Capturing Domain Semantics with Representation Learning: Applications to Health and Function

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Denis R. Newman-Griffis, B.A., M.S.

Graduate Program in Computer Science and Engineering

The Ohio State University

2020

Dissertation Committee:

Prof. Eric Fosler-Lussier, Advisor

Prof. Albert M. Lai

Prof. Huan Sun

Prof. Michael White

Abstract

Natural language processing research is constantly expanding to new domains of text, new types of information, and new applications. A key factor for success in new settings is an ability to capture the characteristics of the language to be analyzed: i.e., the sublanguage of interest. One powerful tool for capturing information about language use is neural representation learning, a family of methods for mathematically representing words, phrases, and other units of language, based on usage patterns in large text corpora. Representation learning for language is predicated on the observation that lexical usage patterns convey important information about meaning, and models this information in terms of geometric relationships between lexical representations. Thus, learned representations provide a lens for analyzing and capturing patterns of language use within restricted domains, as well as for general applications.

This thesis presents two main contributions to the literature. First, we present a method for moving beyond word-level information to learn representations of domain concepts from arbitrary text corpora. We demonstrate that these representations capture domain-relevant information about similarity and relatedness, for both biomedical and encyclopedic concepts, and show that they reveal clinically-significant differences in how medical concepts are discussed among different types of health documentation. We further show how concept-level representations learned using a variety of techniques can be effectively combined for semantic grounding of text.

Second, we present the functional status domain as a new area for NLP analysis and application, with far-reaching impact in both healthcare delivery and social benefits administration. We define how functional status information is realized in practical language, and identify rehabilitation medicine documentation as a distinct sublanguage rich in functional status information. Finally, we show that a combination of neural representation learning from well-chosen data sources and modeling techniques informed by the characteristics of functional status information achieve high-quality extraction of mobility-related information from clinical data, helping to address issues of syntactic complexity and poor coverage in standardized vocabularies. We conclude by identifying future directions leading from our work, including broader application of representation-based analyses of differences in language use, combination of different representation strategies for NLP applications, and further analyses of the structure of functional status information to guide the development of new representation methods for this domain.

For Eric Griffis and Robert Bauman, my first professors

Acknowledgments

It takes a village to raise a PhD student, and I've been lucky enough to have several.

First and foremost, I want to thank the mentors who have made this journey not just possible, but one of enormous learning and growth. Eric Fosler-Lussier, you have been the best advisor I could ask for. You have constantly pushed me: to think more deeply, communicate more clearly, and to pursue scientific inquiry with thoroughness and an insatiable curiosity. I am a much better scientist, communicator, and teacher for your training—and I remain forever grateful that you never did learn to close your office door to keep me from dropping in with the latest crazy idea.

Beth Rasch, your unflagging support throughout this journey has meant the world to me. I have learned an incredible amount from you: about function and health, conducting interdisciplinary research, and keeping a large team running well. Getting the chance to be a part of this team's mission and apply my research to something that can make a real difference has been immensely fulfilling and has set a template I hope to pursue for the rest of my career. I will always be proud to have been a part of the change NIH makes every day.

Albert Lai, working with you has been a real pleasure. You've taught me a great deal about medical informatics, about working across disciplinary boundaries, and

about the academic world. Even though most of our work together has been across several states, your perspective and ideas have always been invaluable.

To my colleagues in the SLaTe lab: it's been real. Chaitanya Shivade and Joo-Kyung Kim, I learned a lot from our conversations and our work together—not to mention the chess games. Adam Stiff, Deblin Bagchi, Peter Plantinga, and Prashant Serai—complaining about ridiculous bugs, laughing at terrible jokes, and getting donuts won't be the same without you guys. Ryan He, Yi Ma, Manirupa Das, Andy Plummer, and all the other SLaTe lab folks—I've learned a lot about how to be a scientist from working with all of you. To everyone in the Clippers group, especially Michael White, Marie-Catherine de Marneffe, William Schuler, and Micha Elsner: your discussions and presentation feedback have been invaluable over the years, and I will dearly miss trekking over to Oxley every Tuesday.

EpiBio, my second scientific family: I have been immensely proud to work as part of this amazing team, and it has been a delight to learn from all of you and to share lunches, SSA meetings, and fabulous baked goods with all of you. Ayah Zirikly, you have been a fantastic co-author, colleague, and friend. Bart Desmet and Guy Divita: it's been great getting to know both of you, and a pleasure to work with you; I look forward to more collaborations (and game nights) in future. Julia Porcino, you have saved my bacon more times than I can count, and talking over ideas with you never fails to bring new insight. Pei-Shu Ho, Jona Camacho Maldonado, and Maryanne Sacco: talking with you is always a joy, and you have taught me so much about annotation and conceptualizing information. Chunxiao Zhou, you have always kept a dozen ships running smoothly, and your curiosity and energy are infectious. To Dr.

Chan and all the incredible folks in RMD and the Clinical Center, and to everyone else in EpiBio (past and present), an enormous thank you.

To my family: I wouldn't be half the person I am today without you. To my father: your joy in and love of learning helped set my on this road, and I'm more proud than I can say to have had you with me on this journey. To my mother: your scholasticism, wit, and endless love are my guiding lights. Matthew: you've always given me something to strive for, and someone to share highs, lows, and games with. To my friends, spread afar from Carleton and here in Columbus: you bring light to my life and a laugh to brighten any cloudy day. I have grown so much in this city, sharing in the Symphony Chorus, the art, the food, the parks, and everything Ohio State has had to offer, and I will take many wonderful memories from my life here.

My PhD studies would not have been possible without research funding from multiple sources. My initial studies were supported by a Graduate Administrative Assistantship from the Engineering Career Services office at OSU, and most of my degree was supported by a Pre-Doctoral Fellowship from the NIH Clinical Center, funded in part by the NIH Intramural Research Program and the U.S. Social Security Administration.

Finally, to Anna: there are no words. This would not be without you.

Vita

July 23, 1991	Born - Hobart, IN, USA
2012	B.A. Computer Science / Russian, Carleton College, Northfield, MN, USA
2017	M.S. Computer Science & Engineering, The Ohio State University, Columbus, OH, USA
2015-present	Pre-Doctoral Fellow, National Institutes of Health Clinical Center, Bethesda, MD, USA

Publications

Journal Articles

Denis Newman-Griffis, Julia Porcino, Ayah Zirikly, Thanh Thieu, Jonathan Camacho Maldonado, Pei-Shu Ho, Min Ding, Leighton Chan, and Elizabeth Rasch. "Broadening horizons: the case for capturing function and the role of health informatics in its use." *BMC Public Health*, (2019) 19:1288. DOI: 10.1186/s12889-019-7630-3

Conference Proceedings

Gordon E. Moon, **Denis Newman-Griffis**, Jinsung Kim, Aravind Sukumaran-Rajam, Eric Fosler-Lussier, and P. Sadayappan. "Parallel Data-Local Training for Optimizing Word2Vec Embeddings for Word and Graph Embeddings." *2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)*, (2019) 1:44-55. DOI: 10.1109/MLHPC49564.2019.00010

Denis Newman-Griffis and Eric Fosler-Lussier. "Writing habits and telltale neighbors: analyzing clinical concept usage patterns with sublanguage embeddings." *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, (2019) 146-156. DOI: 10.18653/v1/D19-6218

Denis Newman-Griffis and Eric Fosler-Lussier. "HARE: a Flexible Highlighting Annotator for Ranking and Exploration." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, (2019) 3:85-90. DOI: 10.18653/v1/D19-3015

Denis Newman-Griffis, Ayah Zirikly, Guy Divita, and Bart Desmet. "Classifying the reported ability in clinical mobility descriptions." *Proceedings of the 18th BioNLP Workshop and Shared Task*, (2019) 1-10. DOI: 10.18653/v1/W19-5001

Brendan Whitaker, **Denis Newman-Griffis**, Aparajita Haldar, Hakan Ferhatosmanoglu, and Eric Fosler-Lussier. "Characterizing the impact of geometric properties of word embeddings on task performance." *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, (2019) 8-17. DOI: 10.18653/v1/W19-2002

Denis Newman-Griffis and Ayah Zirikly "Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility." *Proceedings of the BioNLP 2018 workshop*, (2018) 1-11. DOI: 10.18653/v1/W18-2301

Denis Newman-Griffis, Albert M. Lai, Eric Fosler-Lussier. "Jointly Embedding Entities and Text with Distant Supervision." *Proceedings of The Third Workshop on Representation Learning for NLP*, (2018) 195-206. DOI: 10.18653/v1/W18-3026

Thanh Thieu, Jonathan Camacho, Pei-Shu Ho, Julia Porcino, Min Ding, Lisa Nelson, Elizabeth Rasch, Chunxiao Zhou, Leighton Chan, Diane Brandt, **Denis Newman-Griffis**, Ao Yuan, Albert M. Lai "Inductive identification of functional status information and establishing a gold standard corpus: A case study on the Mobility domain." *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (2017) 2,319-2,321. DOI: 10.1109/BIBM.2017.8218042

Denis Newman-Griffis, Albert M. Lai, Eric Fosler-Lussier. "Insights into Analogy Completion from the Biomedical Domain." *BioNLP 2017*, (2017) 19-28. DOI: 10.18653/v1/W17-2303

Denis R Griffis, Chaitanya Shivade, Eric Fosler-Lussier, Albert M Lai "A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain." *AMIA Joint Summits on Translational Science Proceedings*, (2016) 88-97. PMID: 27570656

Fields of Study

Major Field: Computer Science and Engineering

Studies in Artificial Intelligence: Prof. Eric Fosler-Lussier

Table of Contents

Page

List of Tables

List of Figures

List of Acronyms

ADLs Activities of Daily Living

AI Artificial Intelligence

CMS Centers for Medicare and Medicaid Services

CNN Convolutional Neural Network

CRF Conditional Random Field

CUI Concept Unique Identifier

DALYs Disability Adjusted Life Years

EHR Electronic Health Record

FSI Functional Status Information

ICD International Classification of Diseases

ICF International Classification of Functioning, Disability and Health (World Health Organization, 2001)

ICIDH International Classification of Impairments, Disabilities, and Handicaps (World Health Organization, 1980)

IRB Institutional Review Board

JET Jointly embedding Entities and Text (Newman-Griffis et al., 2018)

k-NN k-Nearest Neighbors

LOINC Logical Observation Identifiers Names and Codes

LSTM Long Short-Term Memory network

MCN Medical Concept Normalization

NER Named Entity Recognition

NLP Natural Language Processing

PHI Protected Health Information

PROSE PROjected Sense Embeddings

RNN Recurrent Neural Network

UMLS Unified Medical Language System (Bodenreider, 2004)

UN United Nations

SNOMED Systematized Nomenclature of Medicine

SNOMED CT SNOMED Clinical Terms

SVM Support Vector Machine

WHO World Health Organization

WSD Word Sense Disambiguation

Chapter 1: Introduction

As the field of Natural Language Processing (NLP) has grown, exploration of new applications for textual analysis, and new information domains to analyze, has been a constant factor of research. From processing of English-language literature and newswire text (Francis et al., 1982; Paul and Baker, 1992), NLP has expanded to include legal documents (Biagioli et al., 2005; Aletras et al., 2019), financial information (Hahn et al., 2018, 2019), scientific literature (Fricke, 2018; Nastase et al., 2019), web content (Buck et al., 2014), and social media (Aramaki et al., 2011; Xu et al., 2019), among many other areas, in myriad languages (Bikel and Zitouni, 2012; Mille et al., 2019; Bojar et al., 2019). This process has been mirrored within the domain of biomedicine, where NLP analysis has extended from discharge summary and radiology report processors (Gabrieli and Speth, 1986; Ranum, 1989) to a vast array of tools and analyses covering medical research (Neumann et al., 2019), social media (Gonzalez-Hernandez et al., 2017), and a broad diversity of clinical specialties (Velupillai et al., 2015).

Processing new domains of language requires two things: an understanding of the characteristics of the target language domain, and reflection of these insights in methodological development. A key technological advancement that has provided invaluable support to both of these goals is the development of *neural representation*

learning (Hinton, 1986; Bengio et al., 2003). In the NLP context, representation learning technologies produce real-valued vectors corresponding to linguistic units such as words, phrases, and sentences. These representations are typically calculated from some combination of statistical co-occurrence patterns in text and expert-curated knowledge resources (Bengio et al., 2013), and thereby reflect the characteristics of the type of language (and associated knowledge) they are trained on. Such vectors are easy to incorporate as mathematical features for predictive models, in principle enabling direct translation of statistical insights into actionable data. While representation technologies have been key contributors to major recent advancements in NLP, their role in modeling the semantics of a particular domain has typically focused at the single-word level, limiting their effectivness in specialized language with multi-word expressions. Further, their effectiveness in capturing meaningful semantics remains opaque, as direct evaluation of the semantic information they encode has proven challenging.

1.1 Overall contributions

This thesis provides two main contributions to the NLP literature:

- We present novel techniques for both learning and adapting neural representations to capture domain-specific semantics at the concept level.
- We provide a case study applying representation learning techniques to *functional status information*, a novel domain for NLP, which presents significant research challenges and high impact in healthcare and government settings.

1.2 Outline and detailed contributions

The remainder of this thesis is organized into three parts.

Part I introduces the domains of interest for this thesis. Chapter 2 introduces conceptual frameworks of human function and their realization in language, and highlights key opportunities for representation learning in this domain. Chapter 3 provides background on NLP in the clinical setting, and highlights key considerations affecting NLP development and application within clinical language.

Part II then introduces the area of representation learning. Chapter 4 describes the linguistic and mathematical theory underlying representation learning, and outlines methodological advances in the field, methods for analyzing the quality and content of representation spaces, and applications of representation learning techniques to sublanguage analysis. Chapter 5 then describes our novel method for learning representations of knowledge base concepts.

Lastly, Part III describes specific applications of representation learning techniques for capturing the semantics of particular domains. Chapter 6 presents experiments on adapting word-level representations to functional status information. Chapter 7 moves from words to concepts, presenting concept representation-based methods for text disambiguation and normalization, in clinical and non-clinical settings. Finally, Chapter 8 describes fine-grained analysis of concept reference patterns in specific sublanguages using learned concept representations.

Chapter 9 concludes the thesis, highlighting key takeaways and directions for future research.

A brief summary of the contributions of each chapter is provided below.

1.2.1 Part I: The clinical domain and the functional status sub-domain

Chapter 2: Functional status information in theory and practice

Functional status information (FSI) is an emerging application domain for NLP, which we highlight as a case study for representation learning in this thesis. Functional status captures the outcome of an individual (in some health state) interacting with society and the world, typically through engaging in specific activities and participating in social roles. Understanding an individual's level of function is key both for providing effective healthcare and for administration of social benefits programs, such as disability insurance. We describe how functional status concepts may be realized in natural language, and identify key gaps in methods and resources for applying NLP to FSI. In order to evaluate the generalizability of existing clinical NLP tools to extract FSI, we analyze a multi-institution collection of clinical documents from rehabilitation medicine, an area of medical practice focused on restoring and optimizing function. We demonstrate that rehabilitation medicine documents can be clearly distinguished from clinical records focused on diagnosis and treatment both by their vocabulary and the medical concepts used, and identify clear cases of failure when applying a benchmark clinical NLP toolkit to extract FSI.

Chapter 3: Characteristics of clinical text affecting representation learning

Text generated in a clinical healthcare setting, such as documents stored in Electronic Health Record (EHR) systems, pose unique challenges for NLP. Clinical language is highly telegraphic, omitting semantically-significant information and utilizing non-standard utterance structures. We highlight the implications of these characteristics for NLP, particularly in the complex area of functional status information. At the

semantic level, we investigate *ambiguity* in clinical text, a key challenge for reliable identification of medical concepts, and describe twelve distinct types of ambiguity deriving from both lexical and medicine-specific factors. While we find that current datasets for medical concept extraction are insufficient to reliably model or evaluate these different types of ambiguity, our ambiguity typology presents clear opportunities to leverage the lexical and semantic patterns encoded in learned representations in future research.

1.2.2 Part II: Representation learning

Chapter 4: Techniques for learning representations of words and concepts

Representation learning has become a fundamental component of modern NLP. Techniques for representation learning rely on specific linguistic and mathematical insights about patterns of language use; we explain these insights, and illustrate what they enable in terms of capturing information about language in restricted domains. We then highlight key shifts in representation learning methodology in recent years, including the development of word-level representation learning methods such as word2vec (Mikolov et al., 2013a) and BERT (Devlin et al., 2019), and discuss different strategies for evaluating the semantic content of these different representation spaces. Finally, we summarize how these techniques have been used to capture and explore distinctive lexical and semantic patterns within different domains.

Chapter 5: Learning representations for domain concepts from text

We then describe JET, a novel method for learning representations of *concepts* from curated terminologies, using an arbitrary text corpus with no human annotations. JET outperforms prior biomedical concept representations, requiring annotated data or specialized preprocessing, on benchmark similarity and relatedness datasets, and shows promise for unsupervised biomedical word sense disambiguation. We further provide a new dataset of similarity and relatedness rankings for entities in Wikipedia, and show that Wikipedia page representations from JET achieve strong performance in this web data evaluation.

1.2.3 Part III: Clinical applications of representation learning

Chapter 6: Adapting word representations to the functional status domain

FSI exhibits two distinctive challenges for NLP that we approach with adaptation of learned representations: a lack of standardized terminologies to reliably capture natural language forms, and long, syntactically-complex reports of activity performance. We first utilize a benchmark Named Entity Recognition (NER) model to extract reports of mobility activity performance, and experiment with utilizing input representation features from rehabilitation medicine-focused data, critical care clinical data, and large-scale non-clinical data. We demonstrate that rehabilitation medicine-focused features match the performance of the non-clinical data, despite a three orders of magnitude difference in corpus size, and find that the best results are achieved by striking a balance between corpus size and language representativeness with the critical care clinical features. We then develop a new method for identifying FSI by estimating the likelihood that each token in a document is part of a

mobility activity report, using learned representation features, and show that this model yields significantly higher recall on identifying long, complex FSI strings than prior experiments. We apply this model to real-world healthcare data collected by two U.S. federal agencies (National Institutes of Health and the Social Security Administration), and demonstrate successful model generalization across different data characteristics.

Chapter 7: Resolving lexical ambiguity with concept representations

Representations of knowledge base concepts and entries in sense inventories offer highly informative features for determining which sense or concept is being referred to in a given utterance. We present a novel supervised method for Word Sense Disambiguation (WSD) that uses a context-sensitive combination of multiple representations of senses. Our method achieves clear gains on benchmark WSD datasets over single or concatenated representations, and outperforms standard baselines on disambiguating lemmas not seen during training. We further apply this model in two clinical settings: normalizing problems, treatments, and tests in clinical records, and identifying mobility activities in physical therapy notes. Our method complements string matching baselines for medical concept normalization, yielding competitive overall performance in a 2019 shared task, and we demonstrate disambiguation across multiple of the ambiguity types described in Chapter 3. On mobility activity reports, our model achieves 90% accuracy over 13 labels in cross-validation experiments, and is outperformed only by SVM classification with in-domain word representations.

Chapter 8: Analyzing concept usage patterns in clinical subdomains

Our representation learning method described in Chapter 5 captures patterns in how domain concepts are referred to within arbitrary text corpora. We learn JET representations from diverse clinical document types, and demonstrate that analysis of nearest neighborhood structure in the resulting representation spaces captures clinically-relevant differences in concept reference patterns across different medical specialties and stages of care.

1.3 Note on notation

Wherever a UMLS CUI is referred to throughout this thesis, it is generally presented with a descriptive string, as <CUI> <*String*>, e.g., C0009443 *Common cold*.

Part I
Clinical Language and the Functional Status Domain

Processing the language of any text genre or information domain requires an understanding of the characteristics of that domain. In this part, we first describe Functional Status Information (Chapter 2), a new domain for NLP that captures the lived experience of health as actualized through interactions with society and the physical world. We identify key information gaps in capturing and analyzing FSI, and describe specific opportunities and challenges for NLP in this domain. We expand this discussion in Chapter 3 key characteristics of the clinical text genre (Chapter 3), including both structural and semantic factors that directly affect the design and evaluation of NLP methods for clinical language. The studies outlined in these two chapters illustrate phenomena of FSI and the clinical genre that are not well addressed by current technologies. In the remainder of this thesis, we describe the development and application of representation learning techniques to close this technology gap.

Chapter 2: Functional Status Information: An Opportunity for Representation Learning

Human activity, and the impact of health conditions on it, is an important component of contemporary conceptualizations of health. However, data on human activity is not captured systematically in current health systems, and methods to analyze these data are under-developed and under-resourced, severely limiting the utility of existing information on activity for decision making. In the first portion of this chapter, we describe key underlying conceptualization of human activity and its role in models of health and disability, and illustrate how these conceptual frameworks can be realized in real data about activity (referred to as Functional Status Information, or FSI). We further identify major information gaps affecting both the availability of FSI and the maturity of analytic models for processing it, and describe concrete steps for the health information management and informatics communities to take towards addressing these gaps.[1] In the second section of this chapter, we provide a significant step towards better understanding of language related to FSI, with an analysis of rehabilitation medicine documentation. Rehabilitation medicine, a discipline focused on optimizing and restoring function, is a rich source for analysis of

[1]Portions of Section 2.1 previously published in D Newman-Griffis, J Porcino, A Zirikly, et al. 2019. "Broadening horizons: the case for capturing function and the role of health informatics in its use." *BMC Public Health*, 19(1):1288. Portions of Section 2.2 have been previously submitted for publication and are currently in revision.

FSI, and a key application area for informatics techniques targeting function and activity. Our analysis demonstrates that rehabilitation medicine documentation forms a distinct sublanguage within the clinical domain, and identifies core characteristics of FSI reports informing development and application of NLP models.

2.1 Defining the functional status domain

2.1.1 Conceptualization and measurement of function

Activity and disability

The way in which we learn about our world as individuals and how we willfully act within it is fundamental to human existence. In sociology, action theory describes human activity, and its purposeful nature, in the context of environments and societies in which activities take place. Although first described in 1937 (Parsons, 1937), the concept of human action has more recently been applied to the fields of medicine and health sciences to characterize the consequences of health conditions as an important and meaningful indicator of health. This concept is reflected in contemporary models of disability, for instance, where disability is conceptualized as the outcome of the interaction between the capabilities of individuals and the demands of environments with which individuals interact. The premise that disability reflects how people function given a particular context was articulated by Saad Nagi in the early 1960s (Nagi, 1965) and formed the basis for every contemporary model of disability that followed. Now codified in the World Health Organization's (WHO) International Classification of Functioning, Disability, and Health (ICF) (World Health Organization, 2001) and adopted internationally, human action is embodied in the domain of activity and participation, where activity represents the execution of an action by an

individual and participation represents actions through involvement in life situations. Actions, which take place at the level of the individual, are distinguished from organ or organ system function (ICF body structures/functions), or cellular/tissue function (ICF health conditions).

What is function?

Human function can be broadly conceptualized as a continuum from body structures and functions to outcomes of interactions between individuals and their environments (World Health Organization, 2013; Beard et al., 2016), and has been argued to reflect "the lived experience of health" (Stucki et al., 2017; Stucki and Bickenbach, 2017a). The ICF defines function as an umbrella term encompassing all aspects of the interaction "between an individual (with a health condition) and that individual's contextual factors (environmental and personal factors)" (World Health Organization, 2013). Within the ICF model, function is broken down into several components, illustrated in Figure 2.1. This model encompasses all aspects of an individual's interaction with the world, including organismal concepts such as individual body functions/structures and pathologies, as well as activity and participation, and all the environmental factors that affect these interactions. Importantly, activity and participation reflect volitional actions that take place at the level of the whole person, such as walking, communicating, applying knowledge, etc., which take place in, and are influenced by, a life situation or social context. For the purposes of this thesis, we operationalize the term "function" at this whole person level, and refer primarily to "activity and participation" in detailed discussion.

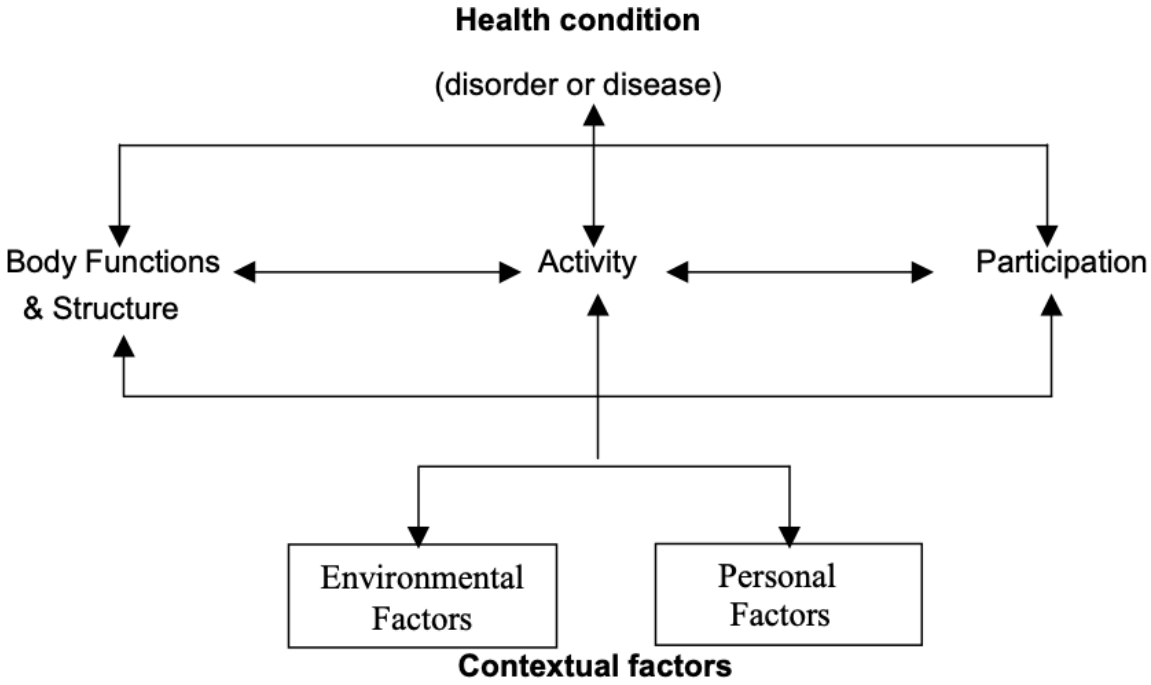


Figure 2.1: Diagram of the International Classification of Functioning, Disability and Health (ICF) model of function. Reproduced by permission of World Health Organization (WHO), from ICF (WHO 2001), p18.

Why are activity and participation important health indicators?

At both the individual and population levels, the ability of people to engage in activities and their participation in social roles shapes the need for resources and the associated response from national agencies, health systems, home and community-based organizations, and other support entities (Hopfe et al., 2016). One timely example of the need for information about activities and participation on a global scale is a consequence of the dramatic shift in the world's demographic profile due to population aging. Among figures that the United Nations (UN) calculates in relationship to population ageing is the support ratio, which is the number of workers per retiree. By 2050, 36 countries, including the U.S., are expected to have support

ratios below 2 (United Nations,, Department of Economic and Social Affairs,, 2017), meaning that there will be fewer than 2 working persons to support each person over the age of 60. Ultimately, an individual's independence and ability to participate in meaningful life activities (i.e., quality of life) will heavily influence resource needs (Taniguchi et al., 2018) and, at the population level, will have an overwhelming impact on national public health, pension, and social programs serving the elderly. As noted in the WHO World Report on Ageing and Health, complex health states resulting from the coexistence of multiple chronic conditions (which can exist at any age) are not adequately represented by identifying or treating one disease at a time. As a result, there is a need for measures that are more meaningful to individuals (Beard et al., 2016).

The need for better information on activity and participation at the individual level has also been widely endorsed (Seals et al., 2016; Stucki et al., 2018). Activity and participation reflect the cumulative outcome of disease burden, i.e. multimorbidity. In the U.S., it has been reported that over half of working age adults experience one or more chronic conditions (Gulley et al., 2011). It is well established that there is a strong and consistent association between a greater number of chronic conditions and the existence and severity of limitations in activities and participation (Verbrugge et al., 1989; Jones and Bell, 2004). Thus, the effect of multiple chronic conditions on the lives of individuals is realized in their overall function (Stucki et al., 2017; Stucki and Bickenbach, 2017a). Since function reflects, among other factors, the cumulative impact of health conditions on the person, and is not disease specific (Cooper et al., 2010), its use as a health indicator helps to address major barriers to holistic,

patient-centered care, such as fragmentation in care resulting from multiple and often competing disease-specific interventions (Hopfe et al., 2018).

In clinical settings, the inclusion of information on activity and participation in case mix calculations has been shown to improve the prediction of patient needs and resource use (Hopfe et al., 2016). Evidence suggests that in cases of multi-morbidity, reducing the complexity of an individual's overall health state to approaches focusing on each disease individually fails to provide adequate care for this growing segment of the global population (Banerjee, 2015). Viewing the outcome of these complexities in the form of whole person function, i.e., activity and participation, is therefore likely to clarify approaches to intervention (Hopfe et al., 2016; Taniguchi et al., 2018). Function reflects a health continuum and thus is more comprehensive in its characterization of health than other endpoints like morbidity or mortality (Hopfe et al., 2018). Indicators of function are strongly predictive of mortality (Keevil et al., 2018) but have the additional advantage of being more proximal health indicators, permitting earlier and potentially more effective interventions (Taniguchi et al., 2018; Cooper et al., 2014). Simple and objective tests of physical performance have been included as biomarkers in studies of ageing, outperforming more traditional impairment measures in models predicting mortality (Cooper et al., 2014). Markers of frailty that include physical function have been associated with employment difficulties in late middle age (Palmer et al., 2016). In addition to predicting mortality, indicators of physical function have been shown to predict other important and more immediate outcomes such as subsequent disability (Perera et al., 2015) and dementia (Beauchet et al., 2016) among older adults. In the context of population ageing, the prevalence of multi-morbidity

within populations and within individuals will have associated consequences in function. Thus, information about function at both the individual and population level is critical for the design of healthcare systems, home and community-based supports, and for resource allocation.

How have activity and participation been measured?

Models of function have historically been developed in the context of discussing disability, which is often described in terms of limitations in function (Nagi, 1965; Institute of Medicine, 1991, 1997). However, these are conceptual models, describing the broad components that contribute to function, and have proven difficult to translate to data models that can capture specific aspects of function in context and how they relate to one another. Even the ICF, the most detailed framework developed for function, does not formally describe the relationships between different structures, activities, and environmental factors. Thus, how best to measure function, and particularly activity and participation, remains an open question despite international efforts (Altman, 2009; Verbrugge, 2016). Many of the existing measurements are at the population level, in the form of national survey questions (see Altman (2009) for a detailed review of many such survey instruments). While these are relatively easy to administer with high coverage, they are necessarily limited in detail, in order to minimize respondent burden, and are unable to capture the individual perspective. Some efforts have been made to systematically capture information on activities of daily living (ADLs) in individual healthcare encounters; however, these have been captured relatively rarely and only present one small piece of the overall picture of activity and participation (Verbrugge, 2016; Bogardus et al., 2004). Notably, information about the environment in which an individual functions is rarely captured

under either approach, despite being central to concepts of function and disability. This continuing debate and development of instruments to measure function means that even where measurements of activity and/or participation are captured, they cannot easily be recognized as such or mapped to standardized vocabularies and data models for analysis.

Definition and examples of terms

One effect of the malleable definitions of function and its measurement is that language used for these concepts varies widely, particularly between different scientific fields. For clarity, we define our key terms here, and provide examples of each.

Function "A dynamic interaction between a person's health condition, environmental factors, and personal factors" (World Health Organization, 2001). This is an umbrella term including cellular and tissue function, organ and body structure function, and whole person function.

Activity and participation the outcome of the interaction between an individual (with some health condition) and their environment, including specific activities and participation, as well as personal contextual factors; also referred to as whole person function. This encompasses basic willful actions, specific tasks, organized activities, and role participation (Altman, 2009; Madans et al., 2004). Examples include walking (including the environment being walked on, anything used to assist in performing the activity, etc), taking public transportation (which combines walking with other activities such as identifying a destination, sitting, etc), or participating in work.

Activity report a recorded observation of activity and/or participation, which identifies relevant components of a specific activity or participation outcome and records them in structured or unstructured data. Examples include, "Patient walked one lap in the hallway," or "Sue reports to work every day at 9 and works with no limitations until 5pm." Prior work has referred to information samples of this type variously as "functioning information" (Stucki and Bickenbach, 2017b), "functional status terms" (Kuang et al., 2015), "functional status information" (Thieu et al. (2017); and used throughout this dissertation), "functional health status" (Skube et al., 2018), and other terms. However, prior studies have not specifically distinguished information about activity and participation from information about other elements of function; thus, we adopt the term "activity report" to clearly distinguish activity and participation information from other types of health information.

2.1.2 The information gap: What's missing?

While information on pathology, and even impairments of individual body functions, has been captured at a high rate for use in many modern health systems (White et al., 2017), information on activity and participation is captured relatively rarely and remains difficult to use effectively (Stucki and Bickenbach, 2017a; Brown et al., 2017). In order to utilize data on activity and participation, i.e., activity reports, the healthcare field has two primary needs: (1) standardized procedures and tools for capturing activity reports routinely and quickly (both in and out of the clinic), and (2) methods for analyzing activity reports to support evidence-based decision making. We suggest approaches towards meeting both of these needs, and provide

four concrete calls to action, with example short term goals for each, to improve both the availability and the utility of activity and participation information for modern health systems.

How can information on activity and participation be captured?

At the population level, most countries collect basic information on function via national censuses and surveys (McPherson et al., 2017), but this information is rarely captured in sufficient detail or frequency to have an impact on healthcare systems (Stucki and Bickenbach, 2017a). Thus, national surveys cannot be responsive to information needs in real time. At the individual level, some self-administered surveys for measuring specific aspects of functional status have been developed (Bowie et al., 2007), and social media technologies have been shown to passively capture some information about individual function (Kuang et al., 2015); wearable devices are also an emerging technology for capturing individuals' activity-related information. However, these tools are, at least currently, difficult to standardize and apply to reliably capture information on activity and participation at scale. Health systems, which many individuals encounter fairly regularly, offer another logical source for capturing information about activity and participation, which can be combined with other sources for a fuller picture of individual function. While some information about activity and participation is already collected during healthcare encounters, there remains significant variability in terms of how often and on whom it is collected, as well as what information is captured (Stucki and Bickenbach, 2017a; Hopfe et al., 2018; Cooper et al., 2014; Brown et al., 2017). In addition to objective observations of activity and participation, expanded documentation of activity reports in health

records can also capture self-reported data, which complements clinical assessments (Bogardus et al., 2004; Burns et al., 1992).

The current scarcity of activity reports at the individual level, recorded via diverse modalities, instruments, and language, presents challenges for their use in decision making. Firstly, to support evidence-based decision making in health systems, health information must be standardized and interoperable to optimize its potential usefulness (Hopfe et al., 2018). Usefulness, in turn, can only be achieved when raw data are translated into knowledge that can change practice, requiring analytics. An extraordinary volume of data generated in health systems (Raghupathi and Raghupathi, 2014), and many of these data may include errors that impact analytics (Ash et al., 2004; Weiskopf and Weng, 2013). Coordination with data from surveys, self-reported tools, and other media can improve accuracy, but increases the volume of data that must be processed. Thus, concerted efforts are needed to tap into the potential of these sources of information on activity and participation. A data-driven approach leveraging current techniques in health informatics to extract information about function, in particular activity and participation, is needed and represents an effort that requires the involvement and coordination of many entities (Beard et al., 2016).

How can information on activity and participation be analyzed?

The field of *health informatics* involves the use of health-related data for scientific inquiry and discovery and for decision making in healthcare and government (Kulikowski et al., 2012). This definition encompasses a wide variety of analytic methods, which can be broadly separated into analyses of *structured data* (i.e., data fields such as vital signs, demographics, lab readings, etc) and *unstructured data* (e.g., free-text health records or medical images). Analysis of structured data has proven

invaluable in advances in medical informatics and public health, such as monitoring cancer incidence and treatment at a population level (White et al., 2017), predicting the need for specific interventions in individual breast cancer treatment (Specht et al., 2005), cohort identification in Nordic countries (Maret-Ouda et al., 2017), and many others (Oellrich et al., 2015; Shortreed et al., 2019). In the area of functional status measures and its correlation to mortality risk, factors such as age, gender, and some ADL information have been used to predict 2-year mortality (Carey et al., 2004). However, a lack of standardized data models means that activity reports are difficult to capture in structured form. Even where some simpler aspects such as ADLs are captured in health records, they are difficult to correlate across samples (Brown et al., 2017); existing structured judgments also often lack the granularity to capture functional limitations informatively (Nicosia et al., 2019). Ongoing development of standards for recording information relevant to activity, such as physical therapy outcomes, offers one way to improve capture of structured data for analysis (Chesbrough et al., 2018). Further, imaging techniques are growing as an area of assessing impairments and associated functional limitations (Steinheimer et al., 2019; Crawford et al., 2019), although such techniques impose high provider burden. Thus, we focus our discussion on unstructured text—particularly in health data—where activity reports have historically been captured (Bogardus et al., 2004; Nicosia et al., 2019), and which offers flexilibity to capture relevant details such as environmental or personal factors. While this flexibility can contribute both to provider burden in writing documentation and analytic burden in extracting useful information from it (Rosenbloom et al., 2011; Payne et al., 2015), technologies such as speech recognition

and natural language processing (NLP) can be used to reduce this burden while enabling automatic extraction, organization, and summarization of relevant information (Payne et al., 2015; Hoyt and Yoshihashi, 2010; Blackley et al., 2019).

How has NLP been used in clinical care and research?

Natural Language Processing (NLP) is a broad field of research that has been used for a variety of purposes in processing health-related text data. The most common application of NLP for health has been automatically extracting and recognizing health-related information in text (Meystre et al., 2008; Kreimeyer et al., 2017; Wang et al., 2018), such as symptoms, procedures, and diseases (Doğan et al., 2014; Soysal et al., 2018; Uzuner et al., 2011), medications (Deléger et al., 2010; Uzuner et al., 2010), health events (Sarker et al., 2015; Haerian et al., 2012), and patient characteristics (Shivade et al., 2013), among other examples. Many advances in NLP for health have been enabled through shared tasks (Huang and Lu, 2016), which engage a wider research community to solve a specific research problem such as detecting smoking status (Uzuner et al., 2008) or heart risk factors (Stubbs et al., 2015). NLP has a long history of research and operational use in clinical informatics (Friedman et al., 1995), and has proven especially helpful for several tasks that are difficult or expensive for humans to complete, such as detecting rates of patient readmission to different facilities (Rastegar-Mojarad et al., 2017). NLP methods have also been incorporated operationally in diverse decision support systems including modeling disease progression, identifying cancer-related information in pathology reports, and risk assessment tools (Gonzalez-Hernandez et al., 2017; Demner-Fushman et al., 2009).

While NLP for healthcare applications has historically focused on diagnostic information such as diseases, symptoms, medications, and procedures, more recent research is expanding both within and outside the clinic to consider contextual factors and other data sources. For example, homelessness is an important social indicator of health that can be extracted from the text of clinical encounters (Bejan et al., 2018; Gundlapalli et al., 2013b). NLP techniques have also been instrumental in leveraging pervasive social media data for diverse applications, from detecting adverse drug reactions to epidemiological surveillance (Demner-Fushman et al., 2009). Social media data have been particularly transformative for monitoring and analyzing mental health, a critical component of function. For instance, NLP techniques have been used to assist moderators of online forums by automatically flagging posts suggesting a mental health crisis—such as suicide risk—for immediate human intervention (Zirikly et al., 2016). Current efforts are also being put into creating datasets that would further application of NLP techniques in this domain (Shing et al., 2018; Zirikly et al., 2019).

How has unstructured activity and participation information been analyzed?

Structured data about activities, participation, and associated limitations are central to disability research, assistive technology development, and many other fields. These data can be gathered from national surveys (Frochen and Mehdizadeh, 2017; Lin and Wu, 2014), obtained via specialized research instruments (Zahuranec et al., 2017), or modeled from available clinical information (Hart et al., 2011), although use of this information in healthcare delivery remains relatively limited (Garçon et al., 2016). Analyzing unstructured text information about activity and participation,

however, along with associated environmental and personal factors, is an emerging area of interest in health informatics research. Recent work has included collecting self-reported function terms by manually reviewing clinical documents and online forums (Kuang et al., 2015), and identifying groups of phrases describing various aspects of function via clinical chart review (Skube et al., 2018); notably, the majority of these terms were not found in established terminological resources like the Unified Medical Language System (UMLS) (Bodenreider, 2004). To address this issue of coverage, some researchers interested in activity and participation have utilized application-specific vocabularies compiled by clinical staff. Such handcrafted approaches have been successful in various applications, including automatically assigning some ICF codes in discharge summaries (Kukafka et al., 2006), using ICF codes for information retrieval (Sundar et al., 2008), and predicting patients' rehospitalization risk (Greenwald et al., 2017). Other work has avoided the coverage issue by using vocabulary-agnostic methods that are targeted to specific types of activity reports (Newman-Griffis and Zirikly, 2018). Additionally, activity and participation information has been used in the extraction and modeling of other functional outcomes, such as frailty or grave illnesses, from clinical text (Shao et al., 2016; Abbott et al., 2017; Davis et al., 2013). These studies represent significant initial efforts in analyzing activity and participation information with NLP, but the lack of systematic alignment with an overall conceptual framework for activity and participation and lack of shared definitions of the analytic tasks pose challenges for synthesizing and building on these efforts.

What is needed to improve analysis of activity and participation information?

While activity reports may not yet be commonplace or a robust part of medical records, important information on activity and participation is currently being recorded, and is most often located in the free text portions of clinical notes. Thus, we focus on NLP as a critical tool for capturing this information for use and analysis. NLP, like other techniques used in health informatics, is a complex field that relies on a multitude of resources to achieve optimal performance. In the following sections, we walk through several factors in effective informatics, what is needed to support them, and the particular challenges of supporting these needs in the context of activity and participation information analysis. These points are also summarized in Table 2.1.

Approach:	**Common datasets for research**	**Shared understanding of analytic tasks**	**Expert knowledge of activity and participation**	**Records of activity and participation**
Analytic Needs:	*Volume:* sufficient data to support modern methods of analysis. *Representation:* data must be widely representative. *Annotation:* gold standard descriptions of activity reports for benchmarking and comparison.	*Problem definitions:* common definitions of analytic tasks and evaluation. *Problem sharing:* information exchange in the community. *Interdisciplinary collaboration:* input from clinical and analytic stakeholders.	*Standardized information structure:* clear standards of information components and their relationships. *Robust sources of information:* capture variation and common usage of language and data.	*Recorded observations:* activity reports explicitly recorded during patient encounters.
Challenges for Activity and Participation:	Records from general encounters often have few activity reports. Activity reports are expressed in diverse language and in varying levels of detail. No common datasets with activity and participation information available for community research.	Prior NLP work on activity and participation information has been highly specific and does not generalize easily. Requires both data science and clinical expertise to effectively adapt existing methods to data that contain activity reports.	Existing resources lack sufficient structure to accurately represent activity and participation information in practice. Current vocabularies have poor coverage of activity and participation concepts and terms.	Multiple competing standards exist for documenting information in rehabilitation medicine. Standards are not widely adopted outside of rehab for standard clinical care.

Table 2.1: Four approaches to addressing the information gap on activity and participation. For further discussion, see Newman-Griffis et al. (2019a).

Approach:	**Common datasets for research**	**Shared understanding of analytic tasks**	**Expert knowledge of activity and participation**	**Records of activity and participation**
Action:	Develop and publish standards for annotating activity reports. Develop resources for research that can be shared through regulatory frameworks.	Identify and define common research problems and applications for processing activity reports.	Develop a clinically-informed ontology for activity and participation information, along with representative terminologies from multiple sources.	Establish common standards for observing and documenting activity reports in patient encounters.
Short-term Goals:	Develop and publish annotation schema for 1-2 specific aspects of activity and participation. Make small sets of annotated data available through existing data sharing mechanisms.	Establish shared tasks for extracting particular activity reports from an annotated dataset.	Develop mappings across existing conceptual frameworks, such as ICF and SNOMED.	Identify minimal interventions that can capture high-impact activity and participation status.

Table 2.1: (continued) Four approaches to addressing the information gap on activity and participation.

What data are needed for successful informatics?

Much of the potential of health informatics is predicated on the availability of data. To develop and evaluate informatics methods for activity and participation, it is necessary to have data that have been annotated, or marked by experts as to what relevant information it contains and where that information can be found. Annotation serves two primary roles in informatics: to tell analysts and machine learning systems what specific information to focus on; and to serve as a gold standard for evaluating proposed automated methods and supporting benchmarking and comparison within a broader research community.

Examples of annotations for activity and participation information might include highlighting descriptions of specific actions (e.g., walking, climbing, shopping, cleaning) or life situations in free text, or even what type of clinical evaluation is being described. Annotating such information requires both identifying and standardizing the components of activity reports in clinical records. Function is defined within the ICF as the outcome of the interaction of individuals with various contextual factors, which means that descriptions of activity and participation tend to be complex and rely on multiple pieces of evidence. For example, a therapist might observe that a patient is able to walk with a rolling walker for 300 feet. While the activity report that needs to be captured is focused on the action ("walk"), this information is contextualized by other factors such as the assistive device ("rolling walker"), and these relationships must be captured in annotation as well.

In addition to annotating data, it is important to devote research and administrative efforts to collecting and sharing large volumes of data that represent activity and participation information. Many recent advances in statistical methods for NLP,

particularly deep learning technologies, have relied on the availability of thousands or millions of documents (Hirschberg and Manning, 2015), but virtually no documents with activity and participation information are available to the broader research community at present. Semantic approaches leveraging expert knowledge have been used to great effect in low-data settings in the past (Jovanović and Bagheri, 2017); however, such methods have typically relied on robust standardized resources that are lacking for activity and participation, emphasizing the value of statistical learning from large datasets.

In medical data, which often contains protected health information (PHI), there are two main strategies for collecting such datasets. First, research groups within a single institution or collaboration may collect private data under an IRB-approved protocol. These data may be re-used or shared after the initial study via mechanisms such as protocol amendments, designing new protocols, and developing business or data use agreements. While these tend to be limited to specific named parties included in the protocol or legal agreements, and may involve lengthy approval processes, such mechanisms have been effectively used for a large variety of data sharing scenarios in health research (Pisani et al., 2016). A second strategy is to curate de-identified datasets that remove PHI and are then made more widely available while taking appropriate precautions for data stewardship. This is not a simple task: though de-identification can be performed without significantly reducing relevant clinical information (Meystre et al., 2014), it is by no means a perfect process (Cimino, 2012; Hripcsak et al., 2016), and defining what qualifies as de-identified requires agreement between all relevant stakeholders, such as IRBs, privacy offices, government entities,

and most certainly patients. De-identified datasets are thus rare, but have an outsize impact in supporting rapid and effective research within a whole community. Under any chosen mechanism, sharable datasets of activity reports will contribute significantly to informatics research and applications using activity and participation information.

How do we make use of these data?

Applying informatic methods to use activity and participation information in clinical and administrative practice requires addressing a wide variety of analytic challenges. One challenge is that many specific analytic tasks do not clearly correspond to existing informatics research problems. For example, activity reports, such as "walks without gait aid 50 feet in hallway", involve the interaction of several concepts. Recognizing and extracting such reports from text requires both identifying the component concepts (e.g., the action "walks", environmental factors "in hallway" and "without gait aid", and the specific distance "50 feet") and linking them together. Walking in an indoor hallway is significantly different from walking across rough terrain outside; connecting these elements is necessary to extract the atomic outcome being recorded. This task is further complicated when multiple outcomes are described in a single report; for example, "ambulate in the hallway and stairs" refers both to walking and to climbing (two distinct activities in the ICF). Thus, modeling the complex semantics of activity reports may involve combining multiple existing research problems, such as named entity recognition, syntactic dependency parsing, and even conceptual inference.

Even well-studied problems such as information retrieval or relation extraction can face new challenges for activity and participation information. For example, some patient records, such as History and Physical Examinations, often contain only a few sentences describing physical and mental function among a much larger concentration of diagnostic history, past procedures, etc. For a healthcare provider or administrator attempting to locate activity and participation information about a patient, such as a physical therapist tracking activity history or an analyst surveying inpatient functional outcomes, it is therefore necessary to pinpoint which sections or paragraphs of a long document include important information to review. Furthermore, such users must be able to quickly access and intuitively organize patient records from a variety of disciplines. These applications encompass diverse NLP tasks, including information extraction and retrieval, for identifying and organizing activity and participation information in the medical record; knowledge representation, for capturing clinically-informed relationships between activity and participation concepts; and determining the relevance of documents with respect to particular criteria, such as potential limitations in function. As with all complex tasks and modern problem solving approaches, addressing these issues for practical care will require interdisciplinary collaboration between clinical or domain experts, knowledge representation specialists, and informaticians at all stages of the analytic process, from defining goals to practical implementation in healthcare systems.

What resources do we need?

Beyond the quantity and quality of available data, many successful clinical applications of NLP have been enabled by robust medical knowledge sources. These sources are referred to by various names, including (but not limited to) taxonomies,

terminologies, and ontologies. These terms are used inconsistently in the literature, so we define each of them for this article as follows. *Terminologies* capture the diverse names used to refer to biomedical concepts, such as diseases, substances, measurements, etc, and are intended to both catalogue distinct concepts and provide a more or less a comprehensive reference for the ways these concepts can be referred to. Biomedical terminologies often include elements of domain-specific *ontology* in their structure, which describe invariant classes of concept, such as diseases, symptoms, biological processes, functions, etc. Ontology also describes relations that hold universally between these classes: for example, that convulsions are a symptom of seizure (Bodenreider et al., 2004). Many terminologies have been developed as formalized coding systems, and can be referred to as *classifications* or *taxonomies*; the International Classification of Diseases (ICD), another WHO reference classification, being a salient example. As a result, the organization of many terminologies distinguishes not only between ontologically different classes (e.g., febrile vs afebrile seizure), but also epistemologically distinct observations (e.g., tuberculosis identified via microscopy or bacterial culture) (Bodenreider et al., 2004). Both types have been critical components of many successes in health informatics (Oellrich et al., 2015; Haendel et al., 2018).

However, comparable knowledge sources are few and far between for non-medical aspects of function. The ICF, originally developed in 1980 as the International Classification of Impairments, Disabilities, and Handicaps (ICIDH) and revised in 2001 to better model environmental aspects of function (Simeonsson et al., 2000), is a conceptual terminology that was designed to provide a common language for a wide variety of administrative and policy needs such as reporting, service coordination,

and policy development (World Health Organization, 2013). Though the ICF has been integrated into the UMLS, and some efforts have been made to map it to other ontological resources (Della Mea and Simoncello, 2012), comprehensive coverage of practical vocabulary has never been its intent, and mappings to other well-developed terminologies such as SNOMED CT or LOINC are minimal. As a result, its coverage and granularity for coding practical information on activity and participation has been shown to lag behind higher-coverage medical terminologies (Tu et al., 2015). Additionally, the distinctions it draws do not necessarily reflect a clinically-based organization of knowledge. As a practical example, the mobility-related action of walking is not linked within the ICF to terms commonly used in practice, such as ambulation. A recent review found several other criticisms of the organization of the ICF, such as its emphasis of the health condition component, the ambiguity of concepts, and its "lack of a clear ontological structure" (Heerkens et al., 2018). Some of these criticisms may be related to the lack of revisions to the ICF over the years. While the WHO publishes updates to the language of the ICF each year, it has never been revised, unlike the ICD, which is currently under its 11th revision. Thus, while the ICF has been hailed as the "best prospect for an internationally recognized, sufficiently complete and powerful information reference for the documentation of functioning information" (Hopfe et al., 2018), and it has the potential to be effectively combined with other vocabularies for coding purposes (Vreeman and Richoz, 2015), a number of practical shortcomings make it difficult to utilize for successful NLP methods relying on dictionary definitions or common patterns in order to extract activity and participation information.

2.1.3 Next steps for FSI research

Function is an important indicator of health from both population and individual perspectives. However, information on function, and particularly on activity and participation, has not been used in a routine and standardized way when evaluating and monitoring the health of individuals from a holistic viewpoint. Informatics can enable identification, extraction, and organization of activity and participation information for applications such as disability assessment and health monitoring (Abbott et al., 2017; Davis et al., 2013), and can also be used in software or devices to assist people with disabilities to engage in daily activities effectively (Sorna et al., 2009; Newell et al., 1998). While existing applications of informatics methodologies to activity and participation information have shown promise, they face several challenges, including reliance on manual collection of non-standardized terminologies in text by domain experts, a lack of a shared systematic framework for activity and participation analysis, and a lack of relevant data.

To drive informatics forward as a tool for capturing and utilizing activity and participation information, we recommend four important steps: (1) make activity and participation annotation standards and datasets available to the broader research community; (2) define common research problems in automatically processing activity and participation information; (3) develop robust, machine-readable ontologies for function that describe the components of activity and participation information and their relationships; and (4) establish standards for how and when to document activity and participation status during clinical encounters. In this thesis, we present initial research on the language of functional status information and the utilization of representation learning methodologies to capture it in diverse data settings.

2.2 Rehabilitation medicine documentation forms a distinct clinical sublanguage

In order to identify data-driven directions for research and development of NLP for FSI, we use *sublanguage analysis* to identify characteristics of FSI-related language. Sublanguage analysis is a corpus linguistic tool that has proven useful for adapting NLP techniques to many different domains (Grishman and Kittredge, 1986; Grishman, 2001). It has been instrumental in the successful development of NLP techniques for biomedical data: for example, Friedman et al. (Friedman et al., 2002) described two distinct sublanguages in the biomedical domain—clinical text and biomedical literature—and each of these has seen significant subsequent development as individual avenues for NLP research (Simpson and Demner-Fushman, 2012; Gonzalez-Hernandez et al., 2017). In addition, as machine learning technologies such as deep learning have continued to advance, large and publicly-sharable linguistic corpora have become more and more critical for enabling modern NLP advances (Hirschberg and Manning, 2015). An understanding of the unique characteristics of a language or sublanguage is a critical element of developing such corpora (Biber, 1993).

To gain an initial understanding of the language of functioning information as something distinct from existing work on health conditions and curative care, we turn to rehabilitation medicine. Rehabilitation is a health strategy focused on functioning outcomes (Stucki et al., 2018), and thus represents a rich source for learning about how functioning concepts are described and evaluated. We analyze three distinct EHR corpora, each including both rehabilitation and non-rehabilitation documents, and demonstrate that rehabilitation documents exhibit markedly different linguistic

characteristics from other healthcare records. We also find significant variation between document types across institutions, reinforcing the importance of moving away from the current single-institution model of clinical NLP to support generalizing to broader and more diverse clinical data. We further demonstrate that a commonly-used system for clinical NLP produces several patterns of error in processing rehabilitation documents, and show that these are related to the structure and terminology of functioning information. We conclude our analysis by identifying several clear directions for improving our understanding of the language of functioning information and developing reliable NLP systems for its extraction.

2.2.1 Materials

We analyzed three corpora of free-text EHR data from different institutions, to control for institutional preferences in linguistic patterns. Each institution used a different EHR system at time of corpus retrieval, and each system assigned documents to different types (e.g., *Discharge Summary*, *History & Physical*) within the EHR. We therefore normalized each set of document types to a shared set of labels, and describe them using the following terms:

- **Document type:** (also referred to as "doctype") the document's type within the originating EHR system.
- **Schema:** three different levels of classification—*Domain*, *Discipline*, and *Functional Area*—described in more detail below. These are non-hierarchical; i.e., a document's Domain label does not restrict its possible Discipline label. Each document may have a label for each of the three schema.

- **Schema class:** (also referred to as a "class label" when applied to an individual document) a specific class within a schema. Classes are mutually exclusive: i.e., a single document cannot belong to multiple Domain classes.

The *Domain* schema includes two classes. *Diagnostic* document types are those primarily concerned with assessing, diagnosing, and treating patient health conditions; e.g., symptoms, procedures, and clinical findings. *Functioning* document types are focused on evaluating and improving patient functioning; these are primarily rehabilitation and therapeutic encounters in our data, though they include other encounter types such as those focused on mental health and social interactions.

The finer-grained *Discipline* schema encompasses four classes: *Medical*, for records from medically-focused (i.e., curative or preventive) encounters; *Therapy*, for therapeutic encounter records; *Ancillary*, for documents regarding psychological or ancillary care services, including mental health assessments and social work evaluations; and *Other*, primarily for administrative documents such as case management records and patient visit reminders.

Finally, documents within the Functioning domain fall into one of eight *Functional Area* classes. These are *PT*, for physical therapy; *OT*, for occupational therapy; *RT*, for recreational therapy; *SLP*, for speech language pathology; *Psych*, for psychological encounters; *SW*, for social worker interactions; *Neuro*, for neurological evaluations focused on functioning; and *General*, for rehabilitation-focused encounters not classified at a more granular level.

Table 2.2 describes the frequency of each Domain, Discipline, and Functional Area class among the three EHR corpora, along with the number of EHR document types within the class. As these broad classes often encompass a wide variety of document

types, we retain the original document types from the source EHR to use in our analysis. High variability in some catchall document types meant that they could not be clearly assigned to an appropriate schema class; we therefore excluded these labels from our analysis. We also excluded any document types with fewer than 20 records as not containing a sufficiently representative sample.

BTRIS

We obtained a dataset of 155,215 free-text documents from the Biomedical Translational Research Information System (BTRIS) of the NIH Clinical Center (Cimino and Ayres, 2010) under an NIH Office of Human Subjects Research determination. These were associated with inpatient and outpatient encounters of 19,008 patients over 2 years (2014-2015) in the NIH Clinical Center. Approximately 40% of these records were sourced from the Rehabilitation Medicine Department of the Clinical Center, with the remainder primarily consisting of consults and consult follow-ups from other departments. All records were collected using the NIHCC Clinical Research Informatics System (CRIS), the Clinical Center's EHR platform. The text records were automatically deidentified before being released to us.

The records in this dataset were assigned fine-grained document types within the Clinical Center EHR, including such information as source department, note type, and stage of care: one representative example is *Urology consult—Initial*. Mapping from these types to schema classes was constructed in consultation with a rehabilitation domain expert, and document types were deemed sufficiently specific that further division was unnecessary.

	Domain							
	Diagnostic				Functioning			
BTRIS	94,976 *(87)*				59,456 *(43)*			
OSUMC	391,853 *(25)*				25,234 *(18)*			
MIMIC	2,065,514 *(14)*				17,624 *(11)*			
	Discipline							
	Medical		Therapy		Ancillary		Other	
BTRIS	98,282 *(86)*		49,011 *(33)*		6,491 *(9)*		648 *(2)*	
OSUMC	368,914 *(23)*		21,902 *(14)*		3,332 *(4)*		22,939 *(2)*	
MIMIC	2,064,375 *(12)*		14,923 *(9)*		2,701 *(2)*		1,139 *(2)*	
	Functional Area							
	PT	OT	RT	SLP	Psych	Neuro	SW	General
BTRIS	9,730 *(5)*	8,435 *(13)*	8,003 *(8)*	2,230 *(3)*	1,637 *(4)*	3,220 *(3)*	2,581 *(2)*	20,613 *(4)*
OSUMC	13,542 *(3)*	6,893 *(3)*	121 *(1)*	658 *(2)*	432 *(2)*	—	2,900 *(2)*	688 *(5)*
MIMIC	2,339 *(3)*	2,571 *(2)*	—	942 *(1)*	31 *(1)*	—	8,932 *(3)*	2,809 *(1)*

Table 2.2: EHR document corpora used for rehabilitation sublanguage study, including the number of EHR documents belonging to each class of the Domain, Discipline, and Functional Area schemas for each of the three corpora used. The number of corresponding EHR document types is given in parentheses. Some EHR document types were not assigned to a schema label, and are not included in the counts. Schema classes with no corresponding document types in a given corpus are marked with —.

OSUMC

We also analyzed a set of documents from the Wexner Medical Center at the Ohio State University, obtained under a protocol approved by OSU Medical Center IRB. This consisted of 418,524 free-text records associated with 4,689 unique patients over 8 years (2005-2012). The patient population included both inpatient and outpatient admissions for patients diagnosed with a variety of chronic diseases, including chronic lymphocytic leukemia, prostate cancer, and heart and kidney failure. All records were collected from the OSUMC EHR system (Epic).

The OSUMC records were distributed to us as CSV files, including note text and document type among other metadata. Documents were originally assigned to one of 36 document types: however, several of these, such as *Progress Note*, included documents from a wide variety of encounter types. We reviewed samples from each document type and designed regular expressions to assign more specific labels to those that showed high variability; for example, *Progress Note* was split into subtypes for PT, OT, SLP, Psych, and Social Work. This yielded a final set of 52 topically-consistent document types.

MIMIC

Finally, we analyze the free-text portion of the publicly-available MIMIC-III dataset (Johnson et al., 2016), including over 2 million deidentified text records associated with over 61,293 hospital admissions of 46,467 distinct patients (38,597 adults and 7,870 neonates) to Beth Israel Deaconess Medical Center, in the period between 2001 and 2012. Records were collected from two EHR systems: Philips CareVue and iMDsoft MetaVision. These data were sourced entirely from critical care units, leading

to a relatively low percentage of documentation focusing on functioning as discussed here.

As with the OSUMC documents, MIMIC records were originally assigned to one of 19 broad document types (e.g., *Nursing/Other*). We followed the same procedure of manual review and developing keyword-based regular expressions for filtering to divide these into 29 topically-consistent types.

2.2.2 Methods

We analyze our three corpora along three primary axes: (1) free text analysis, i.e. the structure and content of the free text documents themselves; (2) document classification, experimentally assigning schema labels to previously unseen documents using automated classifiers; and (3) a qualitative analysis of the predictions of a current clinical NLP system and the structure of information in the documents. Each of these axes is described in detail in the following sections.

Free text analysis

Free-text records were characterized in two different ways: lexical usage (i.e., vocabulary frequency) and keyword identification. Documents were tokenized using the Stanford CoreNLP toolkit (Manning et al., 2014), configured with default settings; these tokens were then analyzed on a per-document basis.

Lexical usage To investigate differences in vocabulary usage between schema classes, we considered two specific questions: (1) How are words used differently in different labels, and (2) What words are most distinctive for each label? Unlike document length, which can be calculated and compared equally across all individual documents, frequencies of individual vocabulary words are sparse and highly variable

when assessed on a per-document basis. We therefore used each document's tokenization to count the frequency of each distinct token (word, suffix, or punctuation), and then aggregated these frequencies within each EHR document type.

To compare vocabulary usage patterns between a pair of schema classes X and Y, we first compared all pairs of a document type belonging to X with a document type belonging to Y. For each of these pairs, we calculated the Kullback-Leibler divergence (KLD) of the normalized frequency distributions from each document type. This yields a distribution of KLD values for the schema class pair X/Y (referred to as the *Cross-class* setting).[2] As neither any schema class nor individual document type is linguistically homogeneous, we also calculate two sets of comparison statistics: the distribution of KLD values for each document type pair within a schema class (the Within-class setting), and the distribution of values calculated by randomly splitting each document type in half and comparing the resulting subsets (Baselines). These distributions were then visualized and compared using Wilcoxon's rank-sum test.

Keyword analysis Lexical statistics describe aggregate trends in vocabulary usage. To complement this with specific tokens most clearly indicative of each schema class, we identified key words as follows. Let A be our target class to characterize, B be the class or classes we wish to distinguish A from, and a *sub-corpus* refer to the collection of documents belonging to A or B. For each word w in the vocabulary of the target class A, we calculate the relative frequency of w in each sub-corpus (denoted F_{wA} and F_{wB}); to control for sub-corpus size, we divide the absolute frequency of w by the total token count of the sub-corpus. We further calculate the coverage D_{wA} of w in A, defined as the fraction of documents in A in which w appears at least once.

[2]Note that KLD is non-symmetric; thus, $\mathrm{KLD}(X/Y) \neq \mathrm{KLD}(Y/X)$. We calculated KLD values for both directions in our analyses.

The keyword score of word w in terms of labels A and B is then calculated as

$$\text{kws}(w|A,B) = D_{wA} * (F_{wA} - F_{wB}) \tag{2.1}$$

Words will be highly scored if they are both (a) distinctly more common in one sub-corpus than the other, and (b) present in a high proportion of documents in the target sub-corpus. We refer to the words that maximize this score as the *keywords* of class A. For each individual schema class, we compare the documents belonging to that class to documents belonging to all other classes within the same schema (e.g., Therapy documents will be compared to all Medical, Ancillary, and Other documents together). We ignored English stopwords (using the default list in the NLTK toolkit[3]) and deidentification placeholders from each corpus.

Document classification

One of our motivating hypotheses is that lexical differences are sufficient to classify documents by the type of information they are most likely to describe, a valuable first step in NLP applications for functioning information. We therefore designed document classification experiments to make use of our extracted lexical statistics. To determine how effectively lexical statistics served to classify previously unseen documents into schema classes, we performed classification experiments for the Domain, Discipline, and Functional Area schemas across all three corpora.

We performed 10-fold cross-validation within each corpus, using stratified sampling to ensure consistent class distributions, in order to evaluate classification of new data from the same institution. Within each training set, we used the procedure described in Section 2.2.2 to identify the top 100 keywords for each schema label;

[3] `http://www.nltk.org/`

these words were then used for feature extraction in both the training and unseen test documents.[4] Our feature set for classification was the frequencies of the combined keyword sets, normalized by document length. Each document was used as a single training sample. We evaluated random forest, k-nearest neighbors (k-NN), and support vector machine (SVM) models for classification, using default parameters as implemented in the Scikit-learn Python toolkit (Pedregosa et al., 2011). Random forest and k-NN are not restricted to binary classification and can accommodate all three schema by default, by choosing the class predicted by the most decision trees (random forest) or neighbors (k-NN); for SVM, we used one-vs-rest training for the non-binary Discipline and Functional Area classification tasks. We calculated precision, recall, and F-1 score for each class, and report overall performance by macro averaging across schema classes. As the outcomes of different models on the same testing data are paired, we compared model predictions using Fisher's exact test over the contingency table of correct/incorrect outcomes.[5]

To evaluate how well differences in vocabulary usage generalized across different institutions, we also performed cross-institution experiments. In this setting, one entire corpus was used for keyword identification and as training data for a classifier; the same keywords were then used (regardless of their presence or absence in the target corpus) to extract features and classify documents from each of the other corpora.

[4]In contrast to popular document-level word features such as TF-IDF, which find words that make one document distinctive from others, we use corpus-level keywords that make one *class* distinctive from others across multiple documents.

[5]Fisher's exact test was chosen over the more common McNemar's test in order to support cases where one or more cells of the contingency table has a value less than 5.

Qualitative analysis

Our free text analyses characterize how lexical and structural properties of rehabilitation documents differ from standard clinical language. Document classification experiments then investigate empirical identification of a document's subject matter (i.e., Domain, Discipline, and Functional Area) from lexical observations, to investigate how clearly these distinctions can be drawn on new documents. Finally, we explore how these distinctions affect existing systems for clinical NLP, and how the semantic and syntactic structure of functioning information differs from the concepts historically explored in clinical NLP research.

We conducted a qualitative review of these questions on a stratified random sample of 75 documents: 60 documents from BTRIS (as the richest source of contrasting documents) and 15 from MIMIC (for reference across corpora). BTRIS documents broke down into 30 Functioning records (20 from the Therapy discipline, 5 from Medical, and 5 from Ancillary) and 30 Diagnostic documents (24 Medical, 3 Ancillary, and 3 Other). MIMIC records were 5 Functioning (3 Therapy records and 2 Ancillary) and 10 Diagnostic (all Medical). We processed each of these documents with the cTAKES clinical text analysis toolkit (version 3.2.2) (Savova et al., 2010), using dictionary lookup derived from the 2016AA release of the Unified Medical Language System (UMLS) (Bodenreider, 2004). Our lookup dictionary included the default SNOMED CT and RxNorm vocabularies extended with the contents of the ICF and the set of MeSH headers.

Documents were reviewed (by DNG in conjunction with two domain experts) with an eye towards information extraction, in the form of clinical concept recognition and normalization. The goal of the review was to describe any repeated patterns of error

in automated outputs, and to judge how effectively automated predictions reflected information deemed important for guiding the patient's continuing course of care.

2.2.3 Results

For all tests of statistical significance, we used a significance value of $p = 0.05$ and employed False Discovery Rate (FDR) correction (Benjamini and Hochberg, 1995; Jones et al., 2008). Where p values are lower than the more conservative Bonferroni–corrected threshold of $p = 3.9 \times 10^{-5}$, we indicate with $p < 3.9 \times 10^{-5}$; all other p values are given explicitly.

Vocabulary usage

Looking at trends in vocabulary frequency, we find firstly that baseline distributions stay between a KLD of 0 to 0.26, indicating that the document types in each corpus are all fairly consistent within themselves. By contrast, Figure 2.2 shows KLD distributions for Within-class and Cross-class settings from each corpus, for the Domain schema (a-f) and Therapy (g-i) and Medical (j-l) classes within the Discipline schema, as compared with the relevant baselines.

At the Domain level, we find that both Diagnostic and Functioning documents in BTRIS have significantly lower Within-class KLD distributions than Cross-class ($p < 3.9 \times 10^{-5}$; Wilcoxon's rank-sum test). In OSUMC, while Diagnostic document types have significantly lower Within-class divergence than Cross-class with Functioning ($p < 3.9 \times 10^{-5}$), Functioning document types are sufficiently varied in vocabulary patterns as to yield Within-class divergences that are indistinguishable from the Cross-class comparison ($p = 0.69$). MIMIC shows the reverse: its Functioning document types are highly self-consistent, but Diagnostic document types are so diverse

in vocabulary usage patterns (see Figure 2.2c) that Cross-class divergence patterns are indistinguishable from Within-class ($p = 0.53$).

At the Discipline level, Therapy document types consistently have significantly lower Within-class variance than Cross-class with Medical document types in all three corpora ($p < 3.9 \times 10^{-5}$); the same holds in reverse for BTRIS and OSUMC, while MIMIC's high diversity in Medical document types leads to no significant difference in KLD distributions. However, Ancillary and Other classes, which have both fewer and highly diverse document types, do not generally have significantly lower Within-class KLD distributions than Cross-class ($p > 0.05$ in most cases). Within Functional Area labels, small sample size makes comparison more difficult, but we find that PT, OT, Psych, and SW typically have lower Within-class divergence than Cross-class, while other classes are harder to distinguish from one another by doctype-level vocabulary frequencies alone.

Vocabulary usage within the same schema class clearly varies across the three corpora, though inter-class trends generally hold; Figure 2.3 shows distributions of Within-class KLD values when comparing BTRIS doctypes to one another and to doctypes from OSUMC and MIMIC. Both BTRIS and OSUMC exhibit significantly lower internal Within-class divergence than when compared with MIMIC at the Domain level ($p < 3.9 \times 10^{-5}$), though the extreme diversity of Diagnostic document types in MIMIC means that the reverse does not hold ($p = 0.62$ comparing with OSUMC, $p = 0.3$ with BTRIS). Interestingly, Within-class divergence within OSUMC at the Domain level is indistinguishable from cross-corpus divergence with BTRIS ($p = 0.76$ for Diagnostic, $p = 0.65$ for Functioning), though BTRIS itself has significantly lower internal divergence than cross-corpus with OSUMC ($p < 3.9 \times 10^{-5}$).

At the Discipline level, the trends are broadly similar to those described above, with Therapy being significantly different between all three corpora, except for comparing internal consistency in OSUMC to cross-corpus consistency with BTRIS ($p = 0.11$). Medical is significantly more internally consistent in BTRIS than when compared with OSUMC and MIMIC, and OSUMC is significantly more internally consistent than when compared with MIMIC ($p < 3.9 \times 10^{-5}$); however, OSUMC internal consistency is indistinguishable from cross-corpus with BTRIS ($p = 0.62$), and MIMIC is indistinguishable from either ($p = 0.46$ with OSUMC, $p = 0.17$ with BTRIS). Ancillary and Other are indistinguishable internally vs between corpora ($p > 0.05$ in all cases). At the Functional Area level, PT and OT in BTRIS can be distinguished from OSUMC and MIMIC ($p < 3.9 \times 10^{-5}$), but other cross-corpus comparisons are largely indistinguishable.

Keyword analysis

Table 2.3 lists the top five keywords identified for each Domain and Discipline class in the three corpora studied. While no class is consistent across corpora, the Domain-level distinction between physiological condition and medication in Diagnostic language (e.g., "blood," "oral") and more holistic assessments in Functioning language (e.g., "therapy," "social") is clear. Keywords from the Domain level largely carry into the Discipline level for Medical and Therapy due to the high overlap of document types across schemas. However, Ancillary keywords reflect the diversity of support services included in this class, and Other keywords clearly reflect the different institution-specific uses of this class (research protocols and consent forms in BTRIS, telephone contacts in OSUMC, and family meetings in MIMIC). Notably, while most identified keywords are highly informative, we observe some spurious words, such as

"signatures" and "authored" in BTRIS Functioning documents. These emerge due to documentation practices specific to originating departments or specialties (here, several areas that include electronic signatures on most EHR documents), which are therefore highly indicative of the overall schema label; this is discussed further in Section 2.2.4.

Document classification

Table 2.4 shows F-1 scores from the ten-fold cross-validation experiments within each corpus. K-nearest neighbors classification significantly outperforms both random forest and SVM models for the Domain and Discipline schemas in each corpus ($p < 3.9 \times 10^{-5}$; Fisher's exact test). In the Functional Area schema, SVM exhibits significantly higher F-1 score than random forest and k-NN ($p < 3.9 \times 10^{-5}$) for BTRIS and MIMIC, due to higher per-class recall. k-nearest neighbors performs most consistently, achieving above 90% F-1 score in all cases except Discipline-level classification in OSUMC.

In the cross-corpus setting, however, classifier performance is much less consistent, and overall scores drop significantly. We analyzed F-1 scores with each model for each schema, over each of our 6 cross-corpus settings. Out of these 18 total comparisons, random forest is best in 4, k-NN in 5, and SVM the remaining 9. Taking SVM as the overall best model in order to examine performance in more detail, Table 2.5 shows that recall falls precipitously when a classifier trained on OSUMC or MIMIC is evaluated on the more diverse BTRIS data, with precision decreasing as the number of possible classes increases. Evaluating on OSUMC and MIMIC, corpus specificity of classification is even clearer, with all three metrics falling below 0.5 in most

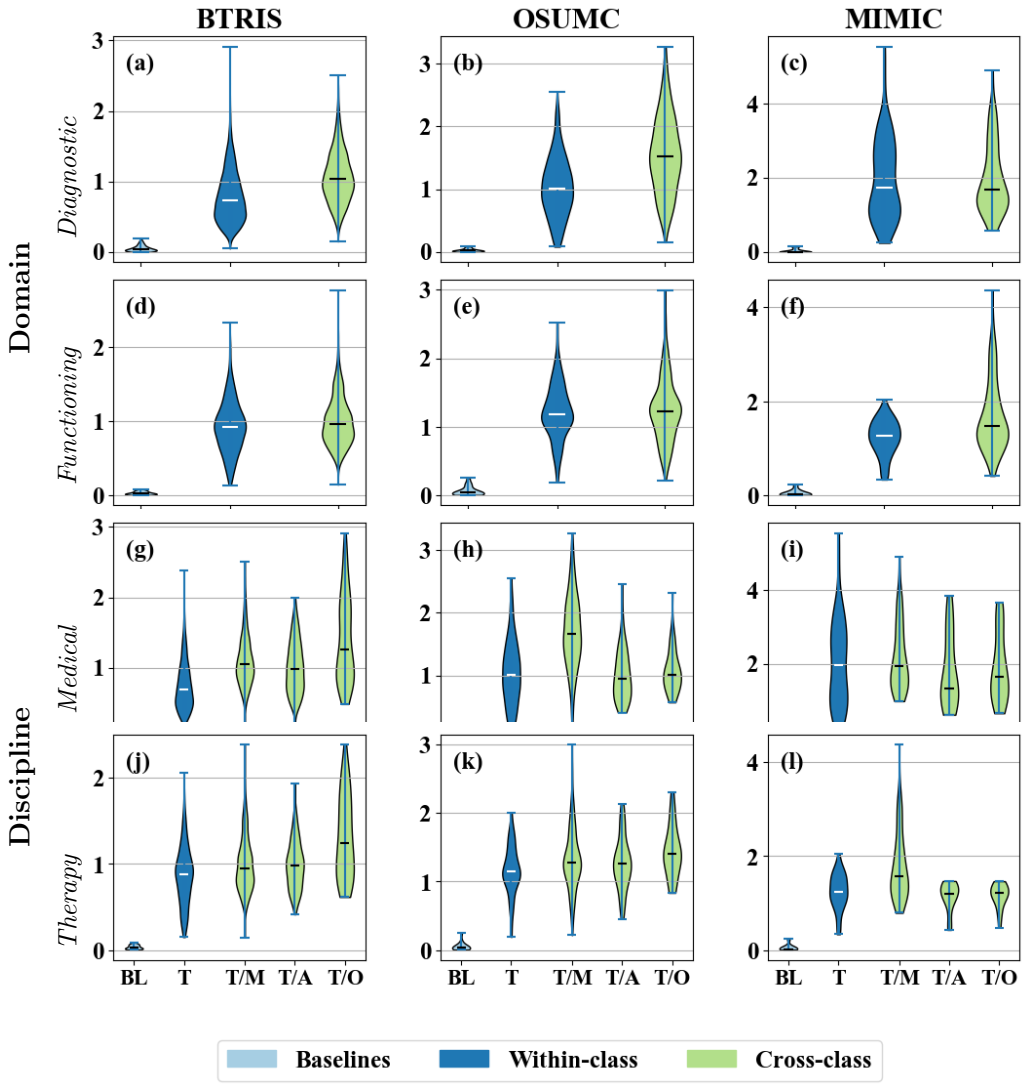


Figure 2.2: K-L divergences of vocabulary distributions between document types, by schema class. Baselines (BL) show KLD distribution for comparing documents within the same document type; Within-class compares document types within the same schema class to one another (e.g. "D" in (a) is comparing Diagnostic types); Cross-class compares document types from one schema class to types from another (e.g. "T/M" is Therapy vs Medical types). (a)-(c) shows distributions for Diagnostic (D) types within the Domain schema, and (d)-(f) shows Functioning (F) distributions (note that KLD is asymmetric). (g)-(i) show distributions for Medical (M) and (j)-(l) for Therapy (T), both in the Discipline schema; Ancillary (A) and Other (O) are omitted for brevity.

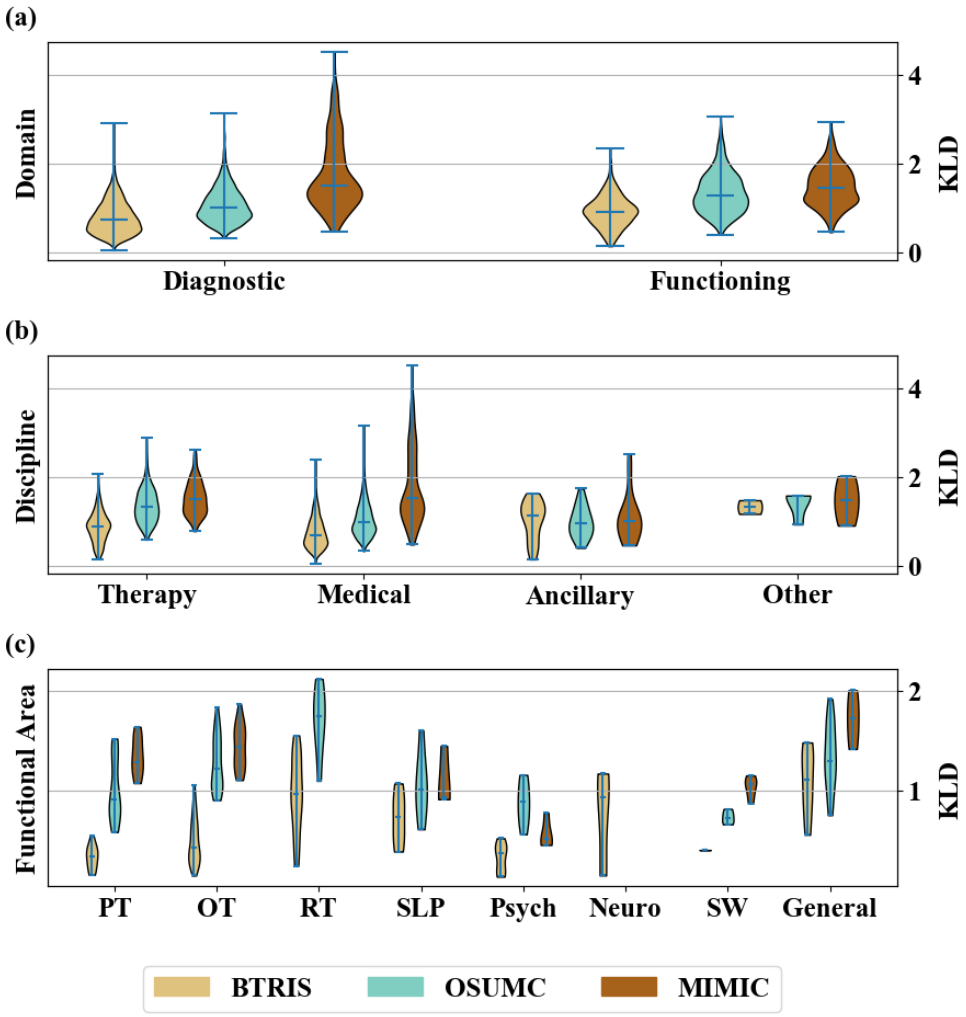


Figure 2.3: Distributions of K-L divergence values for document type pairs within the same schema classes, within BTRIS alone (gold plots on the left of each triad) and comparing doctypes from BTRIS to doctypes from OSUMC (blue middle plots) and MIMIC (brown plots on right). All significance tests were conducted using Wilcoxon's rank-sum statistic. (a) BTRIS has significantly higher Within-class divergence from OSUMC and MIMIC than within itself, for both classes ($p < 3.9 \times 10^{-5}$). (b) BTRIS has significantly higher Within-class divergence from OSUMC and MIMIC for Therapy and Medical classes ($p < 3.9 \times 10^{-5}$), but no significant difference for Ancillary ($p = 0.36$ for OSUMC, $p = 0.64$ for MIMIC) or Other ($p = 0.35$ for OSUMC, $p = 0.64$ for MIMIC) classes. (c) BTRIS has significantly higher Within-class divergence from OSUMC and MIMIC on PT and OT classes ($p < 3.9 \times 10^{-5}$), and from OSUMC on RT ($p = 5.8 \times 10^{-5}$; MIMIC has no RT data). Neither OSUMC nor MIMIC have data for the Neuro Functional Area class.

Corpus	**Domain**		**Discipline**			
	Diagnostic	*Functioning*	*Medical*	*Therapy*	*Ancillary*	*Other*
BTRIS	history	details	history	details	encounter	yes
	normal	therapy	normal	therapy	time	research
	blood	updated	blood	assessment	acupuncture	consent
	daily	signatures	daily	updated	language	protocol
	negative	authored	right	signatures	symptoms	study
OSUMC	daily	assist	daily	assist	social	called
	oral	therapy	oral	therapy	time	please
	normal	physical	normal	physical	states	call
	take	balance	heart	balance	discharge	pharmacy
	heart	goal	rate	goal	would	refill
MIMIC	left	social	left	social	family	discharge
	right	family	right	time	information	care
	chest	follow	chest	activity	support	family
	normal	time	normal	follow	team	case
	reason	work	reason	work	past	insurance

Table 2.3: Top 5 frequency-based keywords for Domain and Discipline classes in each corpus.

Schema	BTRIS			OSUMC			MIMIC		
	RF	*KNN*	*SVM*	*RF*	*KNN*	*SVM*	*RF*	*KNN*	*SVM*
Domain	0.890	**0.977**	0.954	0.811	**0.946**	0.861	0.659	**0.953**	0.790
Discipline	0.427	**0.977**	0.949	0.399	**0.802**	0.773	0.358	**0.887**	0.740
Functional Area	0.632	0.953	**0.965**	0.358	**0.916**	0.867	0.654	0.916	**0.937**

Table 2.4: Macro F-1 scores for cross-validation document classification experiments, by model; includes random forest (RF), k-nearest neighbors (KNN), and support vector machine (SVM) classification models. Results are averaged across all classes for each schema; the best-performing model result in each corpus/schema setting is marked in bold. All differences between models are significant at $p < 3.9 \times 10^{-5}$, using Fisher's exact test.

Source	Domain			Discipline			Functional Area		
	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1
BTRIS									
OSUMC	**0.844**	0.656	0.650	0.640	0.323	0.328	**0.405**	**0.400**	**0.293**
MIMIC	0.823	**0.824**	**0.823**	**0.764**	**0.494**	**0.534**	0.215	0.244	0.137
OSUMC									
BTRIS	**0.615**	**0.860**	**0.634**	**0.419**	**0.500**	0.375	**0.600**	**0.611**	**0.494**
MIMIC	0.601	0.845	0.605	0.377	0.495	0.408	0.464	0.401	0.333
MIMIC									
BTRIS	0.554	**0.840**	0.583	0.287	**0.383**	0.305	0.362	0.425	0.225
OSUMC	**0.577**	0.631	**0.597**	**0.394**	0.315	**0.314**	**0.485**	**0.474**	**0.324**

Table 2.5: Cross-corpus document classification results using SVM classification, reporting macro-averaged precision (Pr), recall (Rec), and F-1 score. Header rows denote target corpus, and Source indicates training data. Results are averaged across all classes for each schema; the best-performing model result in each setting is marked in bold. F-1 is macro-averaged over class-level F-1s, and does not follow F-1 calculation from macro-averaged precision and recall. All differences between source corpora are significant at $p < 3.9 \times 10^{-5}$ (Fisher's exact test), except for comparing BTRIS and OSUMC when testing on MIMIC in the Functional Area schema ($p = 0.494$).

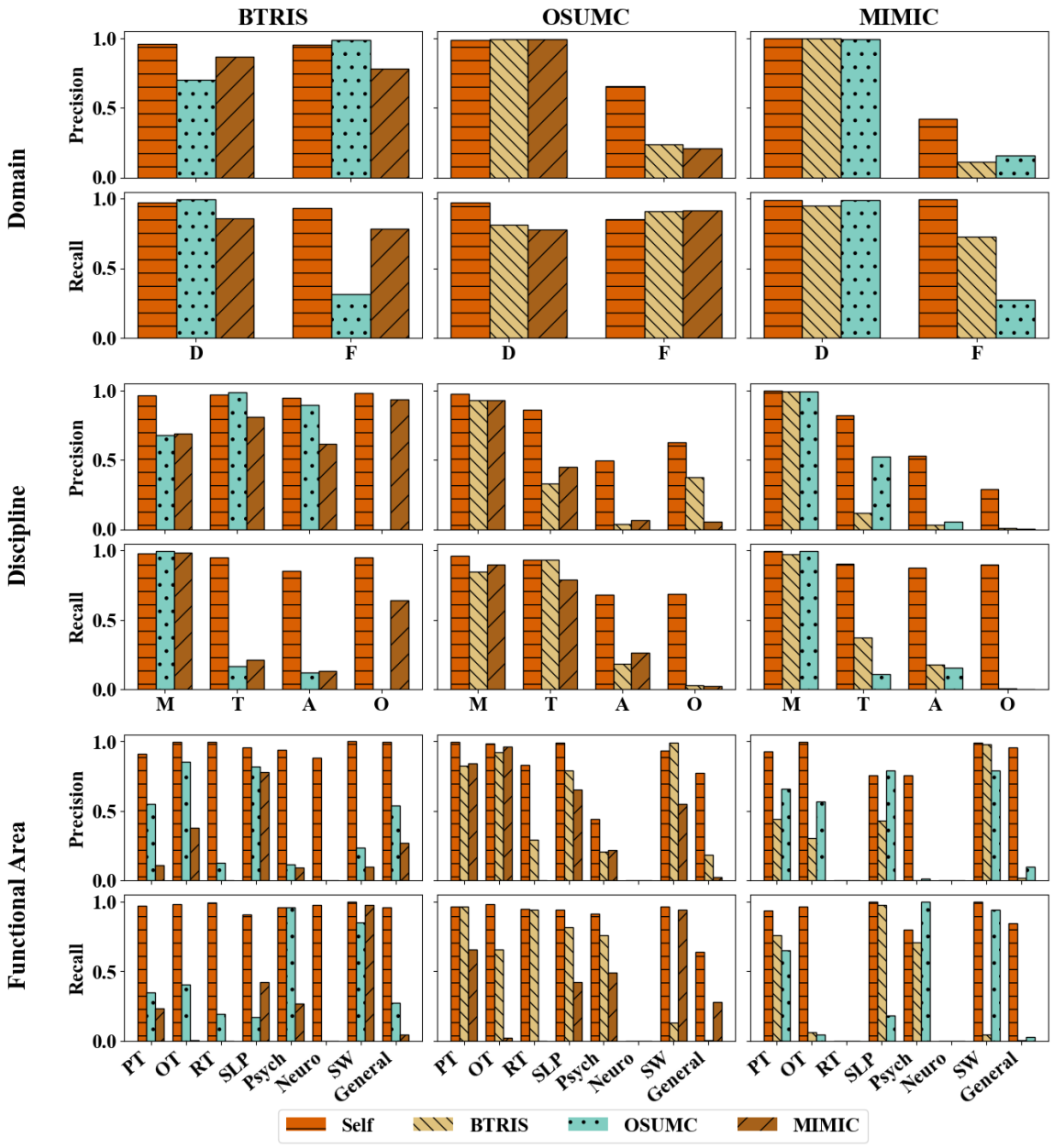


Figure 2.4: Precision and recall for cross-corpus document classification experiments, by schema class. All results reported using SVM classification. Each graph represents testing a specific schema (labeled by row) on a specific corpus (labeled by column), and each colored bar denotes a training corpus. In cases where training and testing are on the same corpus (labeled Self), results are given from ten-fold cross-validation experiments.

Discipline and Functional Area settings. Figure 2.4 clarifies these decreases at the schema class level: precision often decreases slightly for the majority class, while recall falls near zero for rarer classes, reflecting the prevalence of the majority class. Noticeably, while recall does tend to decrease for rarer classes compared to the majority class in the within-corpus cross-validation setting, it remains significantly higher than the cross-corpus performance.

Qualitative analysis

Recognizing key concepts

When it comes to recognizing clinically-relevant pieces of information in rehabilitation text, however, current methods exhibit several distinct points of failure. Most straightforwardly, some of the relevant concepts are simply not encoded at all in the controlled vocabularies available through the UMLS, itself a very high-coverage resource for biomedical vocabulary.[6] For example, "`Patient is an active gentleman`" refers to being *active* in the sense of lifestyle, but available UMLS senses involve medications and biologically active substances. Similarly, the physical therapy evaluation procedure "`balance testing`" was not found in the UMLS, and "`liquid`" as a physical state of food in speech-language pathology evaluation is not directly encoded (though related concepts such as liquid diet are present). Beyond those concepts which are missing entirely, however, a number of relevant concepts are present in controlled vocabularies that were not included in our cTAKES dictionary, such as LOINC and terms from HL7 standards. For example, concepts such as "`hiking`" and "`stationary bike`" were not present in our dictionary subset, but can be found in

[6]UMLS searches in this section were performed using the National Library of Medicine's only search interface at `https://uts.nlm.nih.gov/metathesaurus.html`, using the 2016AA release of UMLS.

various controlled vocabularies indexed in the UMLS. In addition, some terms have multiple senses in different vocabularies: for example, "`depression`" is encoded as a disorder in DSM-IV, but as an anatomical term in SNOMED, leading to "`history of dysthymia and depression`" being tagged with an anatomical site descriptor and no disease/disorder concepts. Thus, for processing rehabilitation documents, it is necessary to expand beyond the default vocabularies to incorporate a broader set of known terms.

It is also necessary to adjust matching and disambiguation procedures to better reflect the distributional characteristics of the rehabilitation sublanguage. Some clinically-relevant concepts in the text were present in the vocabularies included in our cTAKES dictionary, but were not annotated by the full text processing pipeline. For example, "`standing position`" (of a patient) is included in the ICF, but was not included in the final annotations. Other concepts were annotated, but with senses biased towards diagnostic applications: for example, "`plan`" (e.g., of care) was repeatedly marked only as a reference to infantile neuroaxonal dystrophy (PLAN is an acronym for PLA2G6-Associated Neurodegeneration, the overall name for this type of degenerative disorder), and "`bed`" (as a location or device) was annotated as an acronym referring to Bornholm eye disease, despite being included with the correct sense in SNOMED CT and MeSH. As previously observed by Walker and Amsler (Walker and Amsler, 1986) and Pustejovsky et al. (Pustejovsky et al., 1993), multi-sense dictionaries can be effectively combined with knowledge and statistics about a specific sublanguage to tailor disambiguation methods to the domain of interest; such a combination could likely address many of the errors described here.

Structure of functioning information

Functioning is defined in terms of the interaction of various components of a person's health status and life: activities, environment, health condition, etc. This complex nature results in two distinct linguistic challenges for extracting functioning information from text. The first issue is syntactic complexity: as multiple components are involved in describing a single observation, this information cannot be conveyed in single noun phrases or verbs. Prepositional attachment to specify a situation is common, as in "`daily walks with distance less than typical`." It is also often necessary to use a full predicate structure such as an embedded clause or a complete sentence to express a functioning observation, as in "`Patient is unsafe to go home at this time`." Here, multiple components can be identified in the observation: the unsafe status of the patient, the situation with respect to which the patient is unsafe ("`go home`"), and a temporal bound on the information ("`at this time`").

Such complex structures are by no means uncommon in diagnostic and curative observations: for example, "`Crohn's disease and acute back pain which is improved with hydration`" is a representative observation from our documents. However, this latter example can be considered a conjunction of multiple findings, each of which can stand on its own: (1) Crohn's disease, (2) acute back pain, and (3) back pain improving with hydration. By contrast, "`daily walks with distance less than typical`" and "`ambulates across his yard without assistance`" are more atomic: a reference to daily walks conveys little information without an observation of distance, and the patient's ambulation must be contextualized with the

environment (e.g., yard, hallway) and level of independence (e.g., without assistance, with rolling walker) to be informative. As a corollary observation, we also noted several instances of anaphora in functioning observations, often being used to provide more specific context: "`walking with gait aid and doing stairs, two of which are required to enter apartment building`" being an example (anaphora underlined).

The second major challenge is that of implicit information. In several cases, we observed important functioning concepts being described in terms of a specific application or life situation, without explicitly referring to the elements of functioning involved. For example, saying of a patient that he "`crosses his grassy yard to get out into the community`" indicates that (a) the patient can ambulate with some degree of independence, (b) he can do so across a rough surface such as grass, and (c) the action has a further participatory purpose, i.e. getting "out into the community." Similar statements can be made about "`she can manage IV pole independently`," which presupposes manual dexterity and coordination of the IV pole and self ambulation, and "`Wii games utilized from standing position`," which implies some degree of arm mobility and fine motor control. This has some similarities to the investigation of incorporating commonsense knowledge into automated systems (Zellers et al., 2018), though it is more focused on domain-specific knowledge of the associations between activities and their components.

2.2.4 Discussion

We have analyzed rehabilitation medicine documents and contrasted them with records from diagnostic and curative encounters in terms of vocabulary usage, automatically-identified keywords, and document length. Empirical results show that clinical documents can be effectively distinguished from one another by keyword frequency alone at various levels of topical classification, and that these results hold between documents from different institutions. Finally, review of predictions made by an automated clinical language processing system identified several distinct challenges in rehabilitation documents, and suggests new formulations for better understanding the language of rehabilitation medicine and the related language of functioning information. Several intriguing results have emerged from our observations, as have limitations of our current methodology. We discuss four areas of these findings in the following sections.

Assigning schema classes at the document type level is noisy

While we see clear distinctions between schema classes in our analyses, not all doctypes fit cleanly into only one schema class. Some document types, such as *Discharge Summary*, may be used for patients seen for either primarily diagnostic or primarily rehabilitative concerns, and others, such as *History & Physical*, may occasionally include individual pieces of functioning information within a larger weight of diagnostic information. While our heuristic filtering of OSUMC and MIMIC document types accounted for some of this variation, our assignment of schema labels was chosen based on the predominant topics discussed in the majority of documents of a given type, and classification experiments reflect this. Given the diverse nature of functioning information and the diversity of other health-related information it may accompany,

investigating the presence of functioning information independent of document type or overall topic would be a valuable followup to our initial study.

It is also important to note that while rehabilitation medicine is a rich source of functioning information, it is not the only source. Functioning is a broad concept with relevance to curative and preventive health strategies as well. Additionally, though rehabilitation is largely concerned with functioning, diagnostic and other types of medical concepts are by no means excluded. This study draws on documentation of rehabilitation encounters to provide an initial characterization of functioning information language, which can lay the groundwork for a broader investigation of functioning information in other disciplines.

Lexical frequency is an informative piece of a larger picture

Vocabulary frequency provides clear and easily-understood guideposts in analyzing the language used in rehabilitation documents, and empirical classification results demonstrate that it is sufficient information to separate schema classes with high accuracy. We also found an unexpected potential use of our keyword analysis: identifying institutional stopwords within different domains. The keywords shown in Table 2.3 are overall quite informative, but include some examples of non–content-bearing words that are nonetheless indicative of schema label due to differing documentation practices within various departments. For example, as discussed in Section 2.2.3, keywords identified for the Therapy discipline in BTRIS (such as "updated," "authored," and "signatures") were related to electronic signatures, as many of the specialties where these notes originated included this information in every document as standard practice. Thus, expert review of suggested keywords can also

be used to identify department– or note-type–specific stopwords that do not carry content about the care provided to the patient.

However, given the complex nature of functioning information, word-based analysis either at the frequency or topic levels has limitations for discovering representative linguistic forms. Firstly, some words or sets of words may be used equally often, but for different purposes, within different classes and document types. Figure 2.5 shows a t-SNE (Van Der Maaten and Hinton, 2008) visualization[7] of a randomly-selected set of BTRIS documents with Discipline labels, based on the frequencies of identified keywords. Although our classification experiments demonstrate that these document classes are clearly separable using all available keywords, visualization suggests that the classes remain more intermixed than we might expect. This agrees with our observation that some keywords within the Discipline and Functional Area schemas were shared between classes, though remaining distinctive when compared to the corpus as a whole.

Secondly, as discussed in Section 2.2.3, individual observations of functioning generally incorporate multiple components. This structure is difficult to capture from a word-based or even co-occurrence–based approach. However, early work on sublanguage analysis (see (Marsh, 1986; Friedman, 1986), *inter alia*) takes a grammar-based approach, using syntactic structure and/or semantic frames to represent the individual elements of a message and how they relate to one another. Thieu et al. (2017) describe a structured approach in this vein to modeling observations of patient mobility, though they favor semantic type restrictions over syntactic categories. Analysis

[7]We used the t-SNE implementation in scikit-learn for projection.

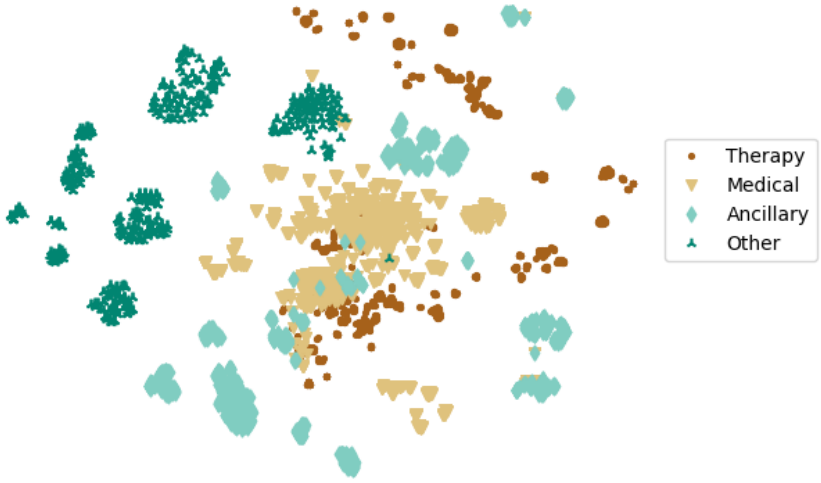


Figure 2.5: t-SNE visualization of keyword features for BTRIS documents by Discipline class (100 from each class). Each document is originally represented as a vector of real-valued relative frequencies for each of keywords identified for every Discipline class using the full BTRIS corpus, and this set of points is projected to 2 dimensions for plotting. Axes in t-SNE visualizations have no specific meaning and are omitted.

of the distributional properties of functioning information thus represents a promising avenue for identifying more general structures for representing descriptions of functioning information in practical use.

A centralized terminology for functioning information is critical

Setting aside the challenges of modeling the structure of functioning information, we observed several issues in finding coverage of its components in controlled terminologies. Many of the relevant terms we identified were present in controlled vocabularies in the UMLS, but were sparsely distributed: some terms could be found in LOINC, others in vocabularies associated with HL7 standards, MEDCIN, the Nursing Outcomes Classification, and more. While a greater coverage could be achieved

by simply including all available vocabularies as potential sources for searching for functioning terms, this also increases the likelihood of finding spurious word senses (such as Bornholm eye disease for "bed"). A more desirable solution to this issue would be to develop a centralized resource of functioning terminology, connecting relevant terms (and relevant *senses* of ambiguous terms) from disparate vocabularies in a single reference.

Although coverage in these various vocabularies is not complete, as observed in Section 2.2.3, many of these terms are to some extent already connected through the UMLS Metathesaurus. However, the underlying ontological structure for functioning concepts to make these connections direct and intuitive is currently lacking. The ICF can provide some basis for this ontological structure, but its applicability to *representing concepts* of functioning (as opposed to its intended purpose of *classifying individual observations*) is limited. Heerkens et al. (2018) describe several related criticisms of the ICF in practice, including its focus on health condition over bio-psycho-social perspectives and the ambiguity of its concepts. We also note that relationships between concepts, such as an activity (e.g., sitting, feeding oneself) and specific elements of the environment involved (e.g., bed, chair, kitchen) are difficult to encode using the ICF alone. An extended ontological structure to support representing all elements of functioning at a more theoretical level would enable more powerful structured analysis of functioning information in practice.

Sublanguage through the lens of institutions and disciplines

Finally, we note two major sources of variability in our sublanguage analysis. The first is the originating institution: we observed significant variance in the language used by the three different institutions we studied; in several cases these differences

were larger than the differences between schema labels within each institution. This aligns with the observations of Carrell et al. (Carrell et al., 2017) regarding idiosyncrasies and variable report structures between institutions. This can be considered an exogenous factor that should be eliminated in a robust characterization of a rehabilitation and/or functioning information sublanguage. Such variation can easily be controlled for by restricting analysis to documents from a single institution, but the significant decrease we observe in document classification results when applying one institution's model to another's data clearly indicates the limitations of this approach. A more lasting, though also more challenging, approach to fix this issue is to more regularly include data from multiple institutions in clinical NLP analysis, a proposal which has recently seen significant attention (Ohno-Machado, 2018).

On the other hand, our analyses at the Functional Area level show clear distinctions between the language used by different specialties, even within rehabilitation medicine. This is to be expected—especially at the vocabulary level—given the different foci of each specialty, and represents an important aspect of our sublanguage analysis. Sublanguages are not mutually exclusive, and can often be analyzed in a hierarchical structure (Marsh, 1986) or even as overlapping items in the same level of a hierarchy (Kittredge and Lehrberger, 1982). The sublanguage of rehabilitation, along with its sibling sublanguage of functioning information, is a subdivision of the broader clinical sublanguage described by Friedman et al. (Friedman et al., 2002), itself a subdivision of the biomedical domain. Thus, it is important when providing further characterization of these sublanguages to ensure that the diverse specialties within a broader domain are represented appropriately, in order to most effectively capture the linguistic trends within the overall domain.

2.2.5 Conclusions

Our work clearly demonstrates that EHR documents from rehabilitation medicine are a valuable text genre for dedicated computational linguistic analysis. The terminology used in these documents is highly distinctive compared to documents that have formed the primary focus of research efforts in clinical NLP. We have illustrated several ways in which existing methods for clinical NLP fail to extract clinically-relevant information from rehabilitation documents. This work provides a corpus-level characterization of rehabilitation notes as a distinct text type across multiple institutions, and describes indicative vocabulary and information structure features of rehabilitation documents.

In addition, we have described initial linguistic characteristics of functioning information, a critical type of healthcare information for global health systems in accommodating the needs of aging populations and people with disabilities. This represents a valuable first step towards using natural language processing to automatically extract and analyze this information within learning health systems. Furthermore, we have identified clear next steps for improving our understanding of functioning information, including studying its prevalence and syntactic and semantic structure within diverse types of healthcare texts, and developing centralized terminological and ontological resources for representing and structuring clinical concepts involved in functioning information.

2.3 Contributions of this thesis towards processing FSI

Whole-person function, as embodied by activity and participation, is a strong predictor of mortality, disability, employment, and resource utilization. Standardized and accessible functional status information will provide valuable knowledge to support holistic and patient-centered care, and to improve the efficiency and effectiveness of health care delivery, management, and planning. We have demonstrated that rehabilitation medicine, a family of healthcare disciplines focusing on optimizing function, exhibits a distinct sublanguage within the clinical domain, utilizing distinct vocabulary and a complex structure of function-related information.

The remainder of this thesis describes how representation learning techniques can help to address both general challenges in clinical NLP, outlined in the next chapter, and the specific challenges of processing FSI. We demonstrate that representation learning methods can capture concept usage patterns within restricted domains, and provide the flexibility to perform high-coverage extraction of functional status information from heterogeneous EHR data.

Chapter 3: Characteristics of Clinical Language to Capture with Representation Learning

Clinical language exhibits distinct features that play an important role when processing any clinical documents, particularly the heterogeneous data in which functional status information is found. This chapter provides a brief overview of two significant issues in clinical text relevant to this thesis: the telegraphic nature of written clinical language, and diverse types of conceptual ambiguity resulting from the complexities of biomedical knowledge.[8]

3.1 Clinical language is telegraphic

One of the key aspects highlighted by Friedman et al. (2002) in their characterization the sublanguages of clinical text and biomedical literature is the the lack of grammatical well-formedness in the clinical sublanguage. Self-contained utterances may consist of noun phrases only, and the rich metadata associated with EHR documents (such as the department where a document was written) mean that semantically significant information may be omitted. In addition, clinical language exhibits a wide variety of semi-structured templates, such as "slot:value" pairs, which often vary based on document section (Divita et al., 2014).

[8]Portions of Section 3.2 have been submitted for publication and are currently under review.

3.1.1 Implications for FSI

For functional status information, which involves the interaction of multiple concepts, this poses two primary challenges. First, the inherent productivity of natural language is compounded by the syntactic flexibility of clinical language to yield a wide variety of ways to present functional interaction information: for example, "`Pt ambulates 300' in clinic with rolling walker`" and "`Ambulation: 4`" (with an implied standardized scale) may express substantively equivalent information. Second, the syntactic complexity of FSI activity reports requires reliable syntactic parsing to fully leverage domain knowledge: for "`Pt ambulates 300' in clinic with rolling walker`", identifying the verb and prepositional phrases is highly informative for identifying the action being performed and the level of assistance required.

One element of the telegraphic nature of clinical text explored in our earlier work is segmentation of a document into "sentences" (i.e., self-contained segments treatable as a complete utterance). In Griffis et al. (2016), we demonstrated that sentence segmentation in clinical document requires very different operational expectations than other domains, including spoken language. We found that off-the-shelf non-clinical sentence segmentation models performed well on newswire and literature data but consistently mis-segmented clinical text, while the reverse was true for a clinical sentence segmentation model. However, the length of activity reports, and the frequent presence of punctuation marks, provides an increased number of over-segmentation opportunities for clinical text segmentation methods, which we previously observed to be somewhat over-eager to segment multiple sentences where no breaks are intended (Griffis et al., 2016). Empirical study of the impact of sentence segmentation algorithms on FSI extraction, as well as the utility of clinical syntactic parsing methods,

will better flesh out the need for further adapting these NLP fundamentals to the FSI domain.

3.2 Clinical language exhibits distinct types of ambiguity in references to medical concepts

At a semantic level, identifying the medical concepts within a document is a key step in analysis of medical records and literature. Mapping natural language to standardized concepts improves interoperability in document analysis (Rosenbloom et al., 2011; Jovanović and Bagheri, 2017) and provides the ability to leverage rich, concept-based knowledge resources such as the Unified Medical Language System (UMLS) (Bodenreider, 2004). This process is a fundamental component of diverse biomedical applications, including clinical trial recruitment (Wu et al., 2018; Weng and Embi, 2019), disease research and precision medicine (Gonzalez et al., 2016; Köhler et al., 2017; Lever et al., 2019), pharmacovigilance and drug repurposing (Ben Abacha et al., 2015; Himmelstein et al., 2017), and clinical decision support (Al-Hablani, 2017). In this work, we identify distinct phenomena leading to ambiguity in Medical Concept Normalization (MCN), and describe key gaps in current approaches and data for normalizing ambiguous clinical language.

Medical concept extraction has two components: (1) Named Entity Recognition (NER), the task of recognizing where concepts are mentioned in the text, and (2) Medical Concept Normalization (MCN), the task of assigning canonical identifiers to concept mentions, in order to unify different ways of referring to the same concept. While MCN has frequently been studied jointly with NER (Savova et al., 2010; Soysal et al., 2018; Elhadad et al., 2015), recent research has begun to investigate challenges specific to the normalization phase of concept extraction.

Three broad challenges emerge in normalization. First, language is productive: practitioners and patients can refer to standardized concepts in diverse ways, requiring recognition of novel phrases beyond those in controlled vocabularies (Elkin et al., 2006; He et al., 2017; Kuang et al., 2015). Second, a single phrase can describe multiple concepts in a way that is more (or different) than the sum of its parts (Osborne et al., 2018; Doğan et al., 2014). Finally, a single natural language form can be used to refer to multiple distinct concepts, yielding *ambiguity*.

Word sense disambiguation (WSD; which often includes phrase disambiguation in the biomedical setting) is thus an integral part of MCN. WSD has been extensively studied in natural language processing methodology (Ide and Veronis, 1998; Navigli, 2009; Raganato et al., 2017a), and ambiguous words and phrases in biomedical literature have been the focus of significant research (Weeber et al., 2001; Savova et al., 2008; Stevenson et al., 2011; Jimeno-Yepes et al., 2011; Jimeno-Yepes, 2017; Charbonnier and Wartena, 2018; Pesaranghader et al., 2019). WSD research in Electronic Health Record (EHR) text, however, has focused almost exclusively on abbreviations and acronyms (Moon et al., 2014; Mowery et al., 2016; Wu et al., 2017; Oleynik et al., 2017; Joopudi et al., 2018). A single dataset of 50 ambiguous strings in EHR data has been developed and studied (Savova et al., 2008; Chasin et al., 2014), but is not freely available for current research. Two large-scale EHR datasets, the ShARe corpus (Elhadad et al., 2015) and MCN (Luo et al., 2019), have been developed for medical concept extraction research, and have been significant drivers in MCN research through multiple shared tasks (Elhadad et al., 2015; Pradhan et al., 2015, 2014; Mowery et al., 2014; Uzuner et al., 2019). However, their role in addressing ambiguity in clinical language has not yet been explored.

Objective

To understand the role of benchmark MCN datasets in designing and evaluating methods to resolve ambiguity in clinical language, we identified ambiguous strings in the ShARe corpus and MCN and analyzed the causes of ambiguity they capture. Using lexical semantic theory and the design of the UMLS as a guide, we developed a typology of ambiguity in clinical language and categorized each string in terms of what type of ambiguity it captures. We found that multiple distinct phenomena cause ambiguity in clinical language, and that the existing datasets are not sufficient to systematically capture these phenomena. Based on our findings, we identified three key gaps in current approaches to MCN, which we hope will spur additional development of tools and resources for resolving medical concept ambiguity.

3.2.1 Background and significance

Linguistic phenomena underpinning clinical ambiguity

Lexical semantics distinguishes between two types of lexical ambiguity: homonymy and polysemy (Cruse, 2004; Murphy, 2010). Homonymy occurs when two lexical items with separate meanings have the same form (e.g., "bank" as reference to financial institution or river bank). Polysemy occurs when one lexical item diverges into distinct, but related meanings (e.g., "coat" for garment or coat of paint). Polysemy can in turn be the result of different phenomena, including default interpretations ("drink" liquid or alcohol), metaphors, and metonymy (usage of a literal association between two concepts in a specified domain, e.g., "The ham sandwich wants his coffee now," uttered in a café setting) (Cruse, 2004; Murphy, 2010). While metaphors are dispreferred in the formal setting of clinical documentation, the telegraphic nature of

medical text (Friedman et al., 2002) lends itself to metonymy by using shorter phrases to refer to more specific concepts, such as procedures (Rindflesch and Aronson, 1994).

Sense relations and ontological distinctions in the UMLS

The UMLS Metathesaurus assigns Concept Unique Identifiers (CUIs) to synonymous lexical items in the medical domain. The semantic relations in the UMLS include taxonomic sense relations that are reflected in lexical phenomena such as hypernymy and hyponymy, as well as meronymy/holonymy in biological and chemical structures (Cruse, 2004). The UMLS has previously been observed to include not only fine-grained ontological distinctions, but also purely epistemological distinctions such as the same disorder resulting from different causes (Bodenreider et al., 2004). This yields high productivity for assignment of different CUIs in cases of ontological distinction, such as "cancer" referring to either general cancer disorders or a specific type in a context such as a prostate exam, as well what Cruse terms propositional synonymy, i.e., different senses which yield the same propositional logic interpretation (Cruse, 2004). Additionally, the difficulty of inter-terminology mapping at scale means that occasional synonyms are assigned different CUIs (Fung et al., 2007).

The role of representative data for clinical ambiguity

Development and evaluation of models for any problem is predicated on the availability of representative data (Borovicka et al., 2012). Prior research has highlighted the frequency of ambiguity in biomedical literature (Weeber et al., 2001; Schuemie et al., 2005) and broken biomedical ambiguity into three broad categories of ambiguous terms, abbreviations, and gene names (Stevenson and Guo, 2010), but an in-depth characterization of the types of ambiguity relevant to clinical data has not

	ShARe Corpus						MCN Corpus
	SemEval-2014 Task 7			CUILESS2016			n2c2 2019 Track 3
	Training	Test	Combined	Training	Dev	Combined	Training
UMLS Version	2011AA			2016AA			2017AB
Source Vocabularies	SNOMED-CT (US)			SNOMED-CT (US)			SNOMED-CT (US), RxNorm
Samples	11,554	8,003	19,557	3,468	1,929	5,397	6,684
Unique Strings	3,654	2,477	5,064	1,519	750	2,011	3,230
Unique CUIs	1,356	1,144	1,871	1,384	639	1,738	2,331

Table 3.1: Details of MCN datasets analyzed for ambiguity, broken down by data subset.

yet been performed. In order to understand what can be learned from the available data for ambiguity and identify areas for future research, it is critical to analyze both the frequency and the types of ambiguity that are captured in clinical datasets.

3.2.2 Materials and methods

Data

The effect of ambiguity in normalizing medical concepts has been researched significantly more in biomedical literature than in clinical data. In order to identify knowledge gaps and key directions for MCN in the clinical setting, where ambiguity may have direct impact on automated tools for clinical decision support, we studied both of the available English-language corpora with concept normalization annotations: the ShARe corpus (Elhadad et al., 2015) and MCN (Luo et al., 2019). Details of these datasets are presented in Table 3.1.

ShARe corpus

The ShARe corpus consists of 531 clinical documents from the MIMIC dataset (Johnson et al., 2016), including discharge summaries, echocardiogram, electrocardiogram and radiology reports. Each document has been annotated for mentions of disorders and normalized to CUIs from SNOMED-CT (Elhadad et al., 2012). The documents were annotated by two professional medical coders, with high Inter-Annotator Agreement (IAA) of 84.6% CUI matches for mentions with identical spans, and all disagreements were adjudicated to produce the final dataset (Pradhan et al., 2015, 2014). Datasets derived from the ShARe corpus have been used as the source for several shared tasks (Pradhan et al., 2013, 2014; Mowery et al., 2014; Elhadad et al., 2015).

SemEval-2014 Task 7: We analyze the subset of 431 documents used for a SemEval-2014 shared task on clinical text analysis[9] (298 training documents, 133 test) (Pradhan et al., 2014), and used as the training set for a following shared task at SemEval-2015, with the remaining 100 corpus documents used for testing (Elhadad et al., 2015). As the SemEval-2015 data is the current version of the corpus, we exclude its 100 test documents from our analysis to preserve their utility as a test set.

CUILESS2016: A significant number of mentions in the ShARe corpus were not assigned a CUI in the original annotations, due either to the appropriate CUI not belonging to the Disorder semantic group in the UMLS or due to compositional mentions requiring multiple CUIs (Elhadad et al., 2015). These mentions were later re-annotated as the CUILESS2016 dataset, with updated guidelines allowing annotation with any CUI in SNOMED-CT (regardless of semantic type) and specified rules

[9]We analyze Track B of the task, focusing on disorder normalization.

for composition (Osborne, 2015; Osborne et al., 2018). These data were split into training and development sets, corresponding to the SemEval-2014 split; the test set from SemEval-2015 was not included in the CUILESS2016 annotations. We follow the same protocol of analyzing the training, development and combined sets separately.

MCN corpus

As the ShARe corpus only provides normalization for mentions of disorder-related concepts, Luo et al. created the MCN corpus to provide mention and normalization data for a wider variety of concepts (Luo et al., 2019). It is derived from the 2010 i2b2/VA shared task on clinical concept extraction, for which documents from multiple healthcare institutions were annotated for all mentions of problems, treatments, and tests (Uzuner et al., 2011). MCN includes 100 of its discharge summaries, with all annotations normalized to CUIs from SNOMED-CT and RxNorm; 2.7% were annotated as "CUI-less". All mentions were dually-annotated with an adjudication phase; pre-adjudication IAA was 67.69% CUI match.

n2c2 2019: The corpus was split into training and test sets, and used for a recent n2c2 shared task on concept normalization (Uzuner et al., 2019). Again, we only analyzed the training set to preserve the utility of the n2c2 test set.

Defining ambiguity

We define ambiguity of a string in two ways: the number of senses a given string can take in general, and the number of senses observed for that string in a finite dataset. Sense inventories such as WordNet (Fellbaum, 1998) or the UMLS Metathesaurus offer a heuristic to measure the former, with the recognition that such inventories may have incomplete coverage of senses (Elkin et al., 2006; Travers and Haas, 2006; ShafieiBavani et al., 2016). The degree to which the sample of ambiguity in

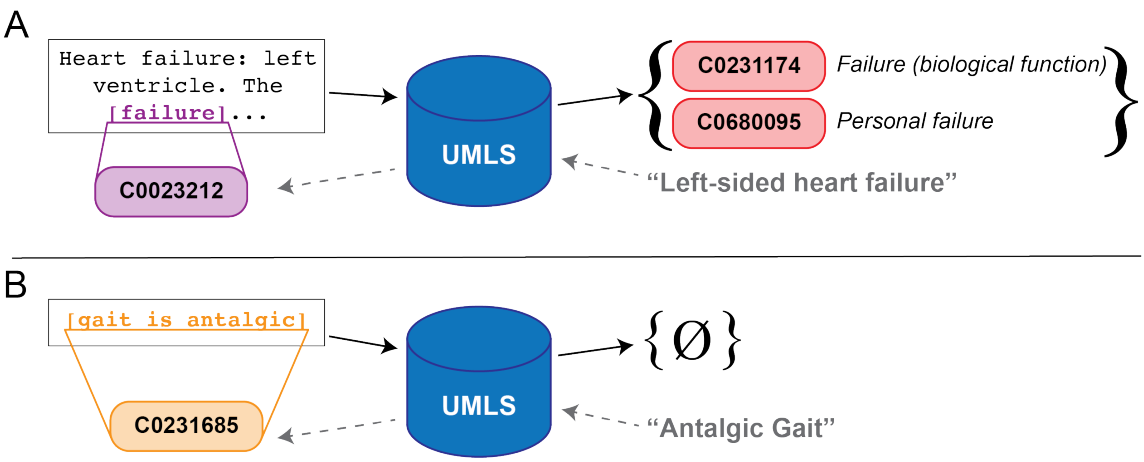


Figure 3.1: Examples of mismatch between medical concept mention string (bold underlined text) and assigned CUI (shown under the mention), due to coreference (A) and predication (B). The right side of each subfigure shows the results of querying the UMLS for the mention string with exact match (top) and the preferred string for the annotated CUI (bottom).

a given dataset is representative of this underlying ambiguity informs the power of statistical models (and their evaluation) to capture the "true" set of senses in the inventory.

While there has been significant work on improving the coverage of synonyms in the UMLS, the breadth and specificity of concepts covered means that useful string-CUI links are often missing (Lang et al., 2017). This is exacerbated by linguistic phenomena such as coreference, allowing seemingly general strings to take very specific meanings, and predication, splitting known strings with a copula (see Figure 3.1 for examples). We therefore measure the ambiguity of medical concept strings in terms of the UMLS from three sources:

- *Dataset ambiguity* – the number of unique CUIs associated with a string in a given dataset.

- *Exact UMLS ambiguity* – the number of unique CUIs canonically associated with the string in the MRCONSO table of the UMLS, filtered to dataset vocabularies.

- *Approximate UMLS ambiguity* – the number of unique CUIs returned when querying the UMLS search API for the string, using word search, filtered to dataset vocabularies.

We also evaluated the coverage of the UMLS search results, in terms of whether they included the CUIs associated with each string in the dataset. As a "CUI-less" result cannot be returned from the UMLS, we excluded all strings with any "CUI-less" annotations; for compositional annotations in CUILESS2016, we treated a label as covered if any of its component CUIs were included in the UMLS results.

Preprocessing

We normalized both the annotated mention strings and the strings in MRCONSO for ambiguity calculation by lowercasing all characters and dropping determiners. We did not apply lemmatization, as initial experiments empirically combined strings and contexts that we deemed too disjoint.

Analysis of ambiguous strings

Inspired by methodological research demonstrating that different modeling strategies are appropriate for phenomena such as metonymy (Markert and Nissim, 2009; Gritta et al., 2017) and hyponymy (Banerjee and Pedersen, 2002; Patwardhan et al., 2003; Navigli and Velardi, 2005; Navigli and Lapata, 2010; Mavroeidis et al., 2005), we analyzed the ambiguous strings in each dataset in terms of the following lexical phenomena: homonymy, polysemy, hyponymy, meronymy, co-taxonomy (sibling relationships), and metonymy (Cruse, 2004; Murphy, 2010). To measure the ambiguity

captured by the available annotations, we performed our analysis only at the level of dataset ambiguity–i.e., only using the CUIs associated with the string in a single dataset. For each string and its CUIs, we answered the following two questions:

What kind of relationship holds between the CUIs associated with this string? This question regarded only the set of annotated CUIs, and was agnostic to specific samples in the dataset. We evaluated two aspects of this relationship: which (if any) of the above lexical phenomena was most representative of the relationship between the CUIs, and if any phenomenon particular to medical language was a contributing factor.

Are the CUI-level differences reflected in the annotations? Given the breadth of concepts in the UMLS, and the subjective nature of annotation, we analyzed whether the CUI assignments in the dataset samples were meaningfully different, and if they reflected the sample-agnostic relationship between the CUIs.

Ambiguity annotations

The answers to these questions determined three variables for each string:

- *Category* – the primary linguistic or conceptual phenomenon underlying the observed ambiguity;
- *Subcategory* – the biomedicine-specific phenomenon contributing to a pattern of ambiguity; and
- *Arbitrary* – the determination of whether the CUIs' use reflected their conceptual difference.

Annotation was conducted by four authors (DNG, GD, BD, AZ) in three phases: (1) initial categorization of the ambiguous strings in n2c2 2019 and SemEval-2014;

(2) validation of the resulting typology through joint annotation and adjudication of 30 random ambiguous strings from n2c2 2019; and (3) re-annotation of all datasets with the finalized typology.

Handling compositional CUIs in CUILESS2016

Compositional annotations in CUILESS2016 presented two variables for ambiguity analysis: single- or multiple-CUI annotations, and ambiguity of annotations across samples. We categorized each string in CUILESS as having (a) unambiguous single-CUI annotation, (b) unambiguous multi-CUI annotation, (c) ambiguous single-CUI annotation, or (d) ambiguous annotations with both single- and multi-CUI labels. The latter two categories were considered ambiguous for our analysis.

Cross-dataset analysis

Finally, we evaluated data representativeness in terms of ambiguity in two ways: between train/test splits in a single dataset (using SemEval-2014 and CUILESS2016), and across datasets. We compared SemEval-2014 and CUILESS2016, both from the same corpus, as well comparing each to n2c2 2019 (cross-corpus). For each string present in a pair of datasets, we compared the annotated CUIs along two axes: (1) differences in ambiguity type and (2) overlap in annotated CUI sets. We further analyzed the coverage of approximate UMLS search for retrieving the combination of CUIs present between the two datasets, to measure the effectiveness of UMLS search for high-coverage CUI retrieval.

		SemEval-2014	CUILESS2016	n2c2 2019
Dataset ambiguity	Unambiguous strings	3,014	1,732	3,066
	Ambiguous strings	130	273	58
	Mean ambiguity	2.15	3.34	2.07
Exact UMLS ambiguity	Unambiguous strings	2,835	1,908	2,705
	No CUIs found	1,328	1,475	1,287
	Correct CUI	1,309	244	1,250
	Wrong CUI	149	117	153
	Ambiguous strings	309	97	420
	Full CUI coverage	231	42	347
	Partial coverage	21	16	9
	No coverage	57	39	64
	Mean ambiguity	2.41	2.70	2.44
Approximate UMLS ambiguity	Unambiguous strings	1,274	1,414	1,006
	No CUIs found	801	1,134	719
	Correct CUI	372	188	228
	Wrong CUI	90	66	58
	Ambiguous strings	1,870	591	2,119
	Full CUI coverage	1,598	378	1,843
	Partial coverage	51	63	21
	No coverage	221	150	255
	Mean ambiguity	21.60	19.51	30.92

Table 3.2: String-level ambiguity analysis results across datasets, by source of ambiguity. Strings are broken down into unambiguous (one CUI only) or ambiguous (multiple potential CUIs). For UMLS ambiguity, coverage relative to the CUIs each string is annotated with in the dataset is provided. The mean number of CUIs associated with each ambiguous string is provided for each source of ambiguity.

3.2.3 Results

String ambiguity

Table 3.2 shows the results of our string-level ambiguity analysis across our three datasets. 86%-98% of non-CUI-less strings were unambiguous at the dataset level. Using exact UMLS search, 86-95% of strings were unambiguous; however, only 46.2% of strings in both SemEval-2014 and n2c2 2019 return the correct CUI, and over 40% of strings have no exact match (as CUILESS2016 strings may combine multiple concepts, exact match is pessimistic for this dataset). Significantly, for 42% of strings, even

approximate UMLS search failed to retrieve any of the correct CUIs (14.9% of strings where at least one CUI was returned). This indicates that synonym coverage in the UMLS remains an active challenge for clinical language. However, when approximate search did return CUIs, it returned multiple of CUIs 68-88% of the time, with average ambiguity of nearly 20 or more CUIs. Thus, choosing between multiple candidates is a significant challenge for high-coverage MCN.

Ambiguity typology

We identified twelve distinct causes of the ambiguity observed in the datasets, organized into five broad categories. Table 3.3 presents our typology, with examples of each ambiguity type; brief descriptions of each overall category are provided below.

Polysemy We combined homonymy (completely disjoint senses) and polysemy (distinct but related senses)(Cruse, 2004; Murphy, 2010) under the category of *Polysemy* for our analysis. While we observed instances of both, we found no actionable reason to differentiate between them, particularly as other phenomena causing polysemy (e.g., metonymy, hyponymy) were covered by other categories. Thus, Polysemy captured cases where more specific phenomena were not observed and the annotated CUIs were clearly distinct from one another. As there is extensive literature on resolving abbreviations and acronyms (Moon et al., 2014; Mowery et al., 2016; Wu et al., 2017; Oleynik et al., 2017; Joopudi et al., 2018), we treated cases involving abbreviations as a dedicated subcategory.

Metonymy Clinical language is telegraphic, meaning that complex concepts are often referred to by simpler associated forms. Normalizing these references requires

Category	Subcategory	Definition	Example ambiguity	
Polysemy	Abbreviation	Abbreviations or acronyms with distinct senses.	Family hx of breast [ca], emphysema	C0006826 *Malignant Neoplasms*
			BP 137/80 na 124 [ca] 8.7	C0201925 *Calcium Measurement*
	Non-abbreviation	Term ambiguity other than abbreviations or acronyms.	BP was [elevated] at last 2 visits	C0205250 *High (qualitative)*
			Her leg was [elevated] after surgery	C0439775 *Elevation procedure*
Metonymy	Procedure vs Concept	Distinguishes between a medical concept and the procedure or action used to analyze/effect that concept.	[Rhythm] revealed sinus tachycardia	C0199556 *Rhythm ECG (Procedure)*
			The [rhythm] became less stable	C0577801 *Heart rhythm (Finding)*
	Measurement vs Substance	Distinguishes between a physical substance and a measurement of that substance.	Pt blood work to check [potassium]	C0032821 *Potassium (Substance)*
			Sodium 139, [potassium] 4.7	C0202194 *Potassium Measurement*
	Symptom vs Diagnosis	Distinguishes between a finding being marked as a symptom or a (possibly diagnosed) disorder.	Current symptoms include [depression]	C0011570 *Mental Depression*
			Hx of chronic [depression]	C0011581 *Depressive disorder*
	Other	All other types of metonymy.	Transfusion of [blood]	C0005767 *Blood (Body Substance)*
			Discovered [blood] at catheter site	C0019080 *Hemorrhage*

Table 3.3: Ambiguity typology derived from ShARe and MCN corpora. Short definitions are provided for each subcategory, along with two samples of an example ambiguous string and their normalizations.

Category	Subcategory	Definition	Example ambiguity	
Specificity	Hierarchical	Combines hyponymy and meronymy; corresponds to taxonomic UMLS relations.	Cardiac: family hx of [failure]	C0018801 *Heart Failure*
			. . . in left ventricle. This [failure]. . .	C0023212 *Left-sided heart failure*
	Recurrence / Number	Distinguishes between singular and plural forms of a finding, or one episode and recurrent episodes.	No [injuries] at admission	C0175677 *Injury*
			Brought to emergency for his [injuries]	C0026771 *Multiple trauma*
Synonymy	Propositional Synonyms	For a general-purpose application, the set of CUIs are not meaningfully distinct from one another.	Negative skin [jaundice]	C0022346 *Icterus*
			Increased girth and [jaundice]	*C0476232 Jaundice*
	Co-taxonyms	The CUIs are (conceptually or in the UMLS) taxonomic siblings; often overspecification.	2mg [percodan]	C0717448 *Percodan*
			5mg [percodan] p.o.	C2684258 *Percodan (reformulated 2009)*
Error	Semantic	Erroneous CUI assignment, due to misinterpretation, confusion with nearby concept, or other cause.	Open to air with no [erythema]	C0041834 *Erythema*
			Edema but no [erythema]	C0013604 *Edema*
	Typos	One CUI is a typographical error when attempting to enter the other (i.e., no real ambiguity).	[Neoplasm] is adjacent	C0024651 *Malt Grain (Food)*
			Infection most likely [neoplasm]	C0027651 *Neoplasms*

Table 3.7: (Continued) Ambiguity typology derived from ShARe and MCN corpora.

inference from their context: for example, a reference to "sodium" within lab readings implies a measurement of sodium levels, a distinct concept in the UMLS. We observed three primary trends in metonymic annotations: reference to a procedure by an associated biological property, mention of a biological substance to refer to its measurement, and the fact that many symptomatic findings can also be formal diagnoses (e.g., "emphysema", "depression").

Specificity The rich semantic distinctions in the UMLS (e.g., phenotypic variants of a disease) lead to frequent ambiguity of *Specificity*. The ambiguity was often taxonomic, captured as Hierarchical; the other pattern observed was ambiguity in grammatical number of a finding, typically due to inflection (e.g., "no injuries" meaning not a single injury) or recurrence.

Synonymy Many strings were annotated with CUIs that were effectively synonymous; we therefore followed Cruse's definition of *propositional synonymy* (Cruse, 2004), in which ontologically distinct senses nonetheless yield the same propositional interpretation of a statement. We also included co-taxonymy in this category, typically involving annotation with either over-specified CUIs or CUIs separated only by negation.

Error A small number of ambiguity cases were due to erroneous annotations stemming from two causes: typological errors in data entry and selection of an inappropriate CUI.

Ambiguity types in each dataset

Figure 3.2 presents the frequency of each ambiguity type across our three datasets. All but 19 strings (3 in SemEval-2014, 16 in CUILESS2016) exhibited a single ambiguity type (i.e., all CUIs were related in the same way). To compare the distribution

of ambiguity categories across datasets, we visualized their relative frequency in Figure 3.3. Polysemy and Metonymy strings were most common in n2c2 2019, while Specificity was the plurality category in SemEval-2014 and Synonymy was most frequent in CUILESS2016. The sample-wise distribution followed the string-wise distribution, except for Polysemy, which included multiple high-frequency strings in SemEval-2014 and CUILESS2016.

Finally, we visualized the proportion of strings within each ambiguity type considered arbitrary (at the sample level) during annotation, shown in Figure 3.4. Arbitrary rates varied across datasets, with the fewest cases in SemEval-2014 and the most in n2c2 2019. Metonymy – Symptom vs Diagnosis, Specificity – Hierarchical, and Synonymy – Co-taxonymy were all arbitrary in more than 50% of cases.

Cross-dataset ambiguity

The majority of strings are unique to the dataset they appear in, even between train/test splits, as shown in Table 3.8. Disagreement between the CUIs annotated for a string in different datasets is frequent, ranging from 18-100%; between SemEval-2014 and CUILESS2016 (both from the ShARe corpus), 7 of the strings with annotated CUIs in both datasets further disagreed in ambiguity type. While most shared strings were annotated with the same sets of CUIs, a large proportion of those that were not were in fact labeled with entirely disjoint CUIs, indicating only partial coverage of the candidate senses for many strings in the existing datasets.

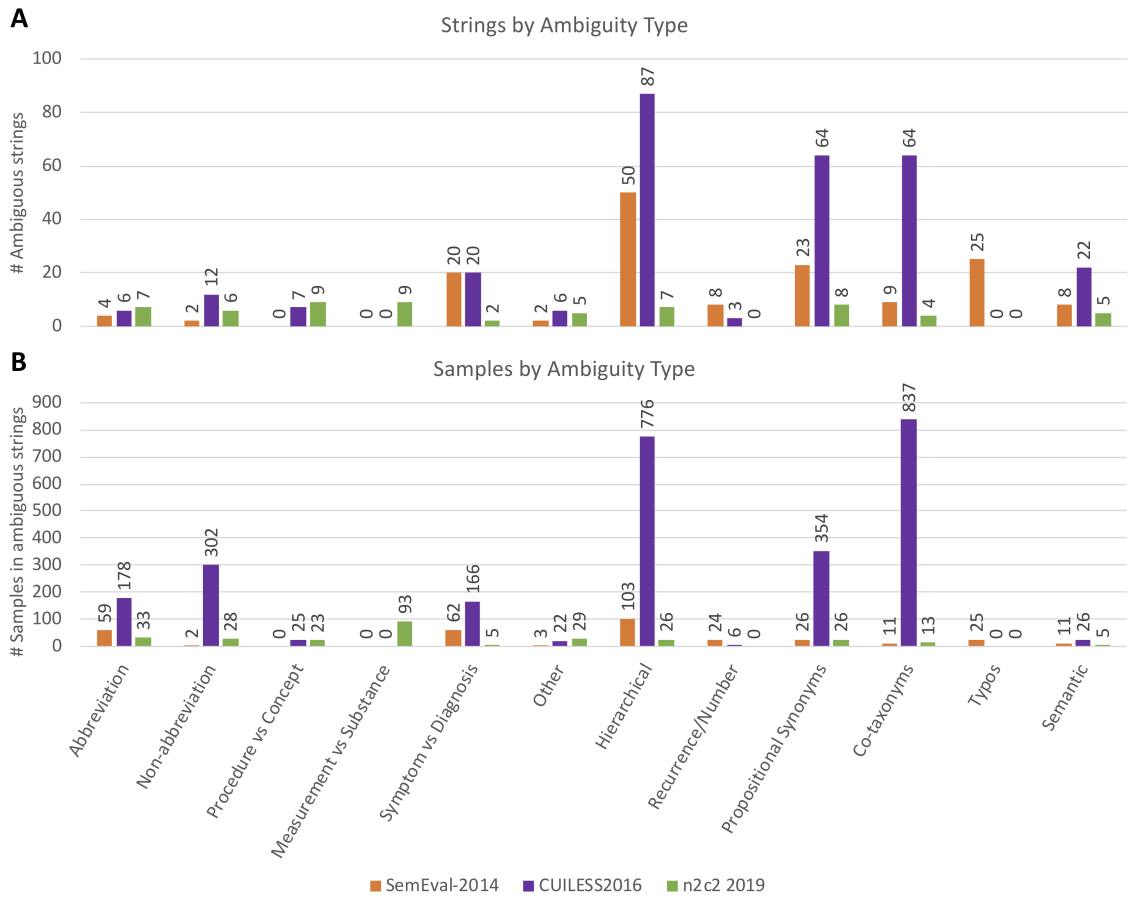


Figure 3.2: Results of MCN ambiguity analysis, showing (A) the number of unique ambiguous strings assigned to each ambiguity type by dataset, along with (B) the total number of dataset samples those strings appear in. (For strings with multiple ambiguity types, the number of affected samples was estimated for each.) The sample counts given for Error subcategories represent the actual count of mis-annotated samples. Total number of ambiguous strings in each dataset – SemEval-2014: 148, CUILESS2016: 274, n2c2 2019: 62. Total number of affected samples in each dataset – SemEval-2014: 326, CUILESS2016: 2,775, n2c2 2019: 295.

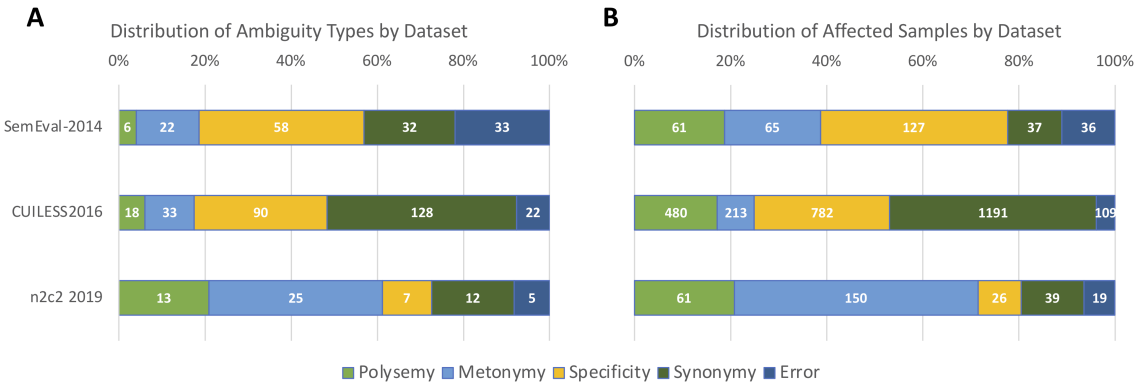


Figure 3.3: Distribution of ambiguity types within each MCN dataset, in terms of (A) the unique strings assigned each ambiguity type and (B) the number of samples in which those strings occur. The number of strings and samples belonging to each typology category is shown within each bar portion.

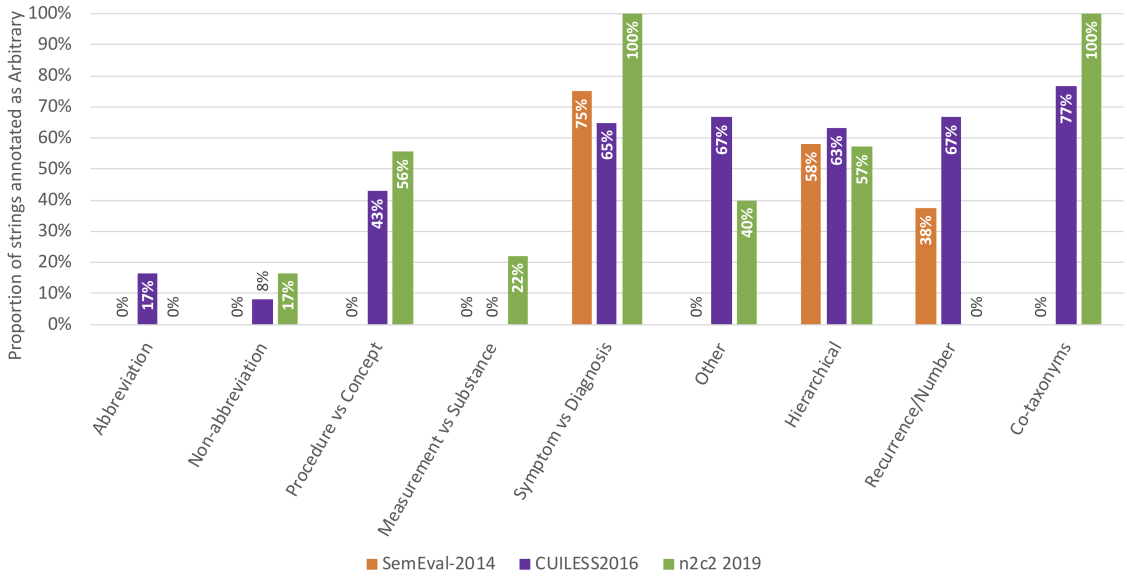


Figure 3.4: Percentage of ambiguous MCN strings in each ambiguity type annotated as "Arbitrary," by dataset. Synonymy – Propositional Synonyms and both Error subcategories are omitted, as they are arbitrary by definition.

	Within dataset		Within corpus		Cross-corpus	
	SemEval-2014 Train/Test	CUILESS 2016 Train/Dev	SemEval-2014 Train / CUILESS2016 Train	SemEval-2014 Test / CUILESS2016 Dev	SemEval-2014 (all) / n2c2 2019 (Train)	CUILESS2016 (all) / n2c2 2019 (Train)
Total strings (left/right)	3,654 / 2,477	1,519 / 750	3,654 / 1,519	2,477 / 750	5,064 / 3,230	2,011 / 3,230
Shared strings	1,067	258	80	52	600	185
Strings with different ambiguity type	*N/A*	*N/A*	7	7	2	0
Strings with different CUI sets	195	166	80	52	271	125
Disjoint annotations	65	65	49	29	200	61
Approximate UMLS search covers some of each (all)	656 (592)	149 (109)	37 (0)	33 (0)	397 (337)	137 (99)

Table 3.8: Cross-dataset comparison of string ambiguity in MCN in three settings: train/test split within the same dataset (SemEval-2014 and CUILESS2016 only), different annotation sets in the same corpus (comparing SemEval-2014 to CUILESS2016, both from the ShARe corpus), and across corpora. All results are the number of strings that fit the given description. Compositional CUILESS2016 annotations were treated as multiple separate CUIs for comparison to approximate UMLS search results.

3.2.4 Discussion

Ambiguity is a key challenge in medical concept normalization. However, relatively little research on ambiguity has focused on clinical language. Our findings demonstrate that clinical language exhibits distinct types of ambiguity, such as clinical patterns in metonymy and specificity, in addition to well-studied problems such as abbreviation expansion. The results highlight three key gaps in the literature for MCN ambiguity, which we discuss as future directions for advancing the state of MCN research.

The next phase of research on clinical ambiguity needs dedicated datasets

The order of magnitude difference between the number of CUIs annotated for each string in our three datasets and the number of CUIs found through approximate UMLS search suggests that our current data resources cover only a small subset of medically-relevant ambiguity. Differences in ambiguity across multiple datasets provide some improvement in addressing this coverage gap, and clearly indicate the value of evaluating new MCN methods on multiple datasets to improve ambiguity coverage. However, the ShARe and MCN corpora were designed to capture an in-depth sample of clinical language, rather than a sample with high coverage of specific challenges like ambiguity. As MCN research continues to advance, more focused datasets capturing specific phenomena are needed to support development and evaluation of methodologies to resolve ambiguity. Savova et al. (2008) followed the protocol used in designing the biomedical NLM WSD corpus (Weeber et al., 2001) to develop a private dataset containing a set of highly-ambiguous clinical strings; adapting and expanding

this protocol with resources such as MIMIC-III (Johnson et al., 2016) offers a proven approach to collect powerful new datasets.

Distinct ambiguity phenomena in MCN call for different evaluation strategies

Evaluation of MCN systems typically uses accuracy (Pradhan et al., 2014, 2015), in which a predicted CUI is either right or wrong. However, as illustrated by the distinct ambiguity types we observed, in many cases a CUI other than the gold label may be highly related (e.g., "Heart failure" and "Left-sided heart failure"), or even propositionally synonymous. As methodologies for MCN improve and expand, alternative evaluation methods leveraging the rich semantics of the UMLS can help to distinguish between a system with a related misprediction from a system with an irrelevant one. A wide variety of similarity and relatedness measures that utilize the UMLS to compare medical concepts have been proposed (McInnes and Pedersen, 2015; McInnes et al., 2009; Andrews et al., 2007; Verspoor et al., 2006), presenting a fruitful avenue for development of new MCN evaluation strategies.

It is important to note, however, that equivalence classes and similarity measures will often be task- or domain-specific. For example, two heart failure phenotypes may be equivalent for presenting summary information in an EHR dashboard, but may be highly distinct for cardiology-specific text mining, or applications with detailed requirements such as clinical trial recruitment. While dedicated evaluation metrics for each task would be impractical, a tradeoff between generalizability and sensitivity to the needs of different applications represents an area for further research.

The UMLS offers powerful semantic tools for high-coverage candidate identification

Our cross-dataset comparison clearly demonstrates the value of utilizing the UMLS to identify a high-coverage set of candidate CUIs for a medical concept, though the lack of 100% coverage reinforces the value of ongoing research on synonym identification (Lang et al., 2017) While retrieving too many candidates presents its own problems, the UMLS provides a variety of semantic tools to filter out uninformative candidates. Contextual features such as identifying document sections can significantly reduce false positive rates for information extraction (Gundlapalli et al., 2013a); for example, a simple regular expression to detect phrase/number alternations would help identify lab readings sections and resolve ambiguity in over 70% of our observed Metonymy – Measurement vs Substance samples. In our analysis, filtering the candidate list from UMLS approximate search to the correct semantic type reduced ambiguity by 37% on average; Figueroa et al. (2009) and Patterson and Hurdle (2011) describe sublanguage-based approaches to prune out unrelated segments of the UMLS in text analysis. Similar methods leveraging UMLS semantics present a significant opportunity for research on MCN methods.

Limitations

The primary limitation of our study was the lack of a broader collection of clinical datasets for MCN. Since our typology was constructed based on the data observed, it is likely that medical language exhibits ambiguity types that were either (a) not present in our data or (b) too infrequent to merit a separate subcategory. This is exacerbated by the limited scope of the datasets analyzed, including only four document types (primarily discharge summaries), with annotations for only a subset of

medical concepts in each case (disorders for ShARe; problems, tests, and treatments for MCN). Thus, our typology should not be taken as capturing all sources of ambiguity in clinical language, nor should our observed distributions of category frequencies be considered universal.

In addition, some ambiguity types were clearer to determine in practice than others. In particular, Specificity – Hierarchical, Synonymy – Co-taxonyms, and Error – Semantic accounted for 39 of the 51 strings noted by the annotators as very difficult to classify in CUILESS2016. The typological structure we proposed is one of multiple that could fit the observed data: for example, "Recurrence/Number" could be recategorized as Polysemy, and Polysemy could itself be split between homonymy and polysemy (Murphy, 2010). Some strings were also so ambiguous as to defy easy categorization: for example, "lesion" appears with 24 different labels in CUILESS2016, and "masses" with 20 across 50 samples.

Finally, preprocessing decisions affect ambiguity significantly. Dropping determiners often assisted our analysis, but also erroneously collapsed distinct strings like "Hepatitis" and "Hepatitis A". Further normalization, such as lemmatization, will increase the representativeness of ambiguity, but may also introduce additional noise.

3.3 Conclusions

The clinical text domain presents general challenges for NLP, requiring domain-specific adjustments of all components of the NLP pipeline, from sentence segmentation and tokenization to concept extraction. We have demonstrated that off-the-shelf tools for segmenting documents fail on clinical text, and that while specialized segmentation methods yield more meaningful segments, they are frequently over-zealous

and introduce additional splits. For functional status information, including activity reports that typically span many tokens and complex syntactic structure, these challenges mean that extraction methods must be robust to noise in document segmentation, supporting extraction from both over- and under-segmented text. Chapter 6 presents work towards robust models for FSI extraction, and identifies clear directions for further research.

At a more semantic level, we have analyzed the role of ambiguity in Medical Concept Normalization (MCN), a key component of extracting medical information from EHR documents. We demonstrated that benchmark datasets for MCN capture very little ambiguity, with much lower coverage of candidate concepts than are present in the UMLS, and with much lower frequency than ambiguity is observed in practical NLP applications. The ambiguous strings that were observed exhibited distinct phenomena from lexical semantics and ontology theory, and were captured in different proportions across datasets. For functional status information, which often uses common words to describe everyday objects, actions, and situations, the challenge of ambiguity is likely to be exacerbated. Chapter 7 describes methods for using learned representations to support disambiguation in diverse settings, and provides preliminary evidence that representation learning can help to address different kinds of medical ambiguity.

Part II
Learning and Analyzing Representations of Language

Modeling the characteristics of language in any given genre or information domain is key to successful development and application of NLP technologies, particularly in specialized areas. As illustrated in Part I, these characteristics include vocabulary (distinctive words and terms, domain-specific ambiguity), context (who produces the language, and where it is recorded), and structure (what pieces of information are relevant, and how they relate to one another), as well as idiosyncrasies of particular groups of speakers. The family of technologies under the umbrella label of *representation learning* provide a well-equipped toolbox for capturing and modeling these characteristics from observed data. In this part, we first outline the intuitions behind representation learning technologies and how they capture important information about language (Chapter 4); we further highlight key advances in the development of representation learning technologies and what they offer for the purposes of capturing and analyzing language use in specific domains. We then describe our novel contribution to the representation learning family in Chapter 5, a method for learning representations of concepts of interest for new domains that lack large-scale expert-curated resources and annotated datasets. These insights, including our contribution, will be concretely applied to specific challenges in FSI and the clinical genre in Part III, in which we demonstrate that thoughtful design and application of representation learning technologies significantly reduces major information gaps in these areas.

Chapter 4: Capturing lexical and semantic patterns with representation learning

Language is most easily conceptualized symbolically, in terms of discrete words, concepts, utterances, etc. Many NLP tools are thus symbolic in nature, such as regular expressions, n-gram language models, early rule-based parsers, and gazetteers. However, for statistical modeling, including the use of machine learning techniques, finding effective mathematical representations of linguistic units is one of the key research and engineering challenges in NLP. As "effectiveness" is application-dependent, this means that not only is the space of mathematical representation strategies infinite, but no one strategy will necessarily suffice for all settings. For example, a set of documents may be easily discriminated from one another in terms of topic by representing them with word counts or TF-IDF vectors, but these representations will not capture whether the documents were written grammatically.[10]

[10]Portions of Section 4.4.2 have previously been published in D Newman-Griffis, AM Lai, and E Fosler-Lussier. 2017. "Insights into analogy completion from the biomedical domain." *BioNLP 2017*, 19-28. Portions of Section 4.5.1 have been published in D Newman-Griffis and E Fosler-Lussier. 2017. "Second-Order Word Embeddings from Nearest Neighbor Topological Features." *arXiv preprint arXiv:1705.08488.* Portions of Section 4.5.2 have been published in in B Whitaker, D Newman-Griffis, A Haldar, et al. 2019. "Characterizing the impact of geometric properties of word embeddings on task performance." *Proceedings of the Third Workshop on Evaluating Vector Space Representations*, 8-17.

Two primary strategies have been used in developing representations. *Feature engineering* approaches involve researchers and engineers identifying distinct features–such as nearby words, part of speech tag, the presence of particular semantic constructs, etc.–with which to represent a given unit. These approaches are often quite powerful, and can yield insight into both the correlations a statistical model captures and potential underlying phenomena in human cognition.

The other approach, motivated by neuronal activation patterns and models from cognitive science, is *representation learning* (Hinton, 1986; Hinton et al., 1986). Under this approach, representations of linguistic units are estimated from observed data capturing a particular phenomenon, whether logical propositions (Hinton, 1986), word sequence generation (Bengio et al., 2003), or more specific applications. The representation learning framework has come to dominate NLP methods and applications over the last two decades, both through learning application-specific representations and through *pretraining* of representations to be used as features in a variety of applications (Bengio et al., 2003; Mikolov et al., 2013a; Peters et al., 2018). This chapter provides an overview of recent advances in representation learning methods for NLP, and discusses ways in which these representations have been used as a tool to investigate linguistic and conceptual characteristics of specific domains.

4.1 Embedding language: a note on terminology

Vector-valued representations are often referred to in the NLP literature as *embeddings*. The use of this particular term reflects some interesting aspects of what these representations are designed to capture about language use in specific domains, and merits some discussion.

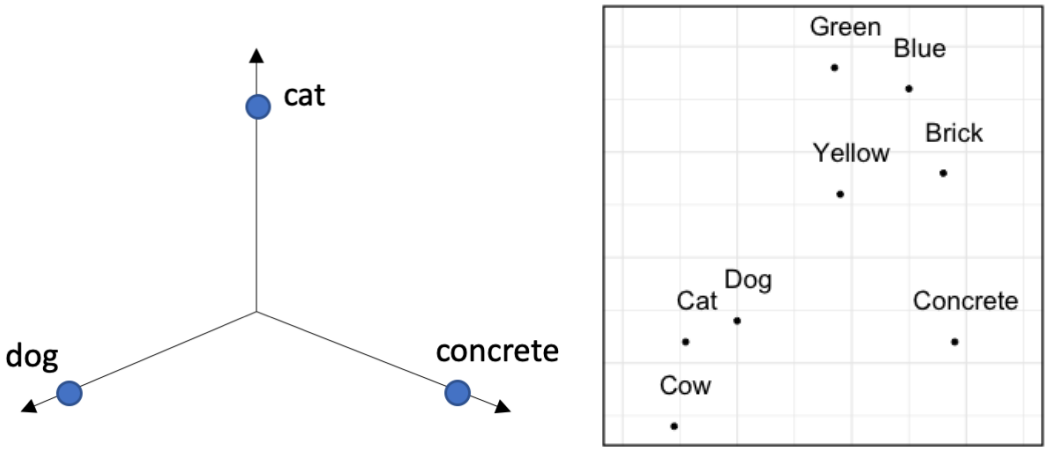
Generally, an *embedding* is a function $f : X \to Y$ that is both injective and structure-preserving (with respect to X). The particular type of structure to be preserved may be defined differently depending on the nature of X and Y; for example, an embedding between topologies (also called a *homeomorphism*) requires that both f and its inverse be bijective and continuous (Munkres, 2013).

The "structure" of natural language as a mathematical object is as yet (and perhaps necessarily) ill-defined. However, three different kinds of structure can be loosely conceptualized to provide insight into what utility learned representations provide: a language's vocabulary; information drawn from a finite source, such as a text corpus or a knowledge base; and information drawn from an infinite manifold representing all potential uses of different units (or knowledge about them) within a language. For ease of discussion, we describe these in terms of words, but the analysis can be easily extended to other linguistic units.

4.1.1 Embedding a vocabulary

The simplest scenario is when we frame X as the vocabulary of a language: an n-dimensional hypercube where each of the n words in the language are represented as a unique one-hot vector (illustrated in Figure 4.1a).[11] In this setting, the only structural characteristics are (1) that every point is unique and (2) that every point is orthogonal to every other, an undesirable aspect (as *cat* and *dog* share significantly more similarity in natural usage than *cat* and *concrete*) that can be ignored in the embedding definition. If a set of n real-valued representations are all unique, then they are then an embedding of X under this perspective. This framing is the easiest

[11]Note that while n is finite in practice, it may in principle be infinite.



(a) One-hot representations (b) Distributional representations

Figure 4.1: Comparison of symbolic and distributional representations of words. (a) shows one-hot symbolic representations, in which all words are orthogonal to one another; (b) shows distributional representations, in which different levels of semantic similarity are expressible.

to model mathematically, but the least powerful for encoding informative aspects of language.

4.1.2 Embedding a finite sample

This scenario, which reflects how most representations are calculated in practice, frames X as some type of information about the units of interest derived from a finite data sample. These data are typically either samples of language use (i.e., a text corpus) or of knowledge in a particular domain (e.g., a knowledge base).

The language modeling perspective for learning representations, discussed in more detail below, is a highly salient example of this approach based the *distributional hypothesis* characterized by Harris (1954). This postulates that a word's meaning can be understood in terms of the contexts in which it appears in natural language

use: thus, because *cat* and *dog* are used in similar contexts, they have similar meaning, while *concrete* is used in nearly disjoint contexts and thus is highly dissimilar to both. (Figure 4.1 illustrates this point, in contrast with a one-hot representation.) On this account, the embedding domain X can be modeled in terms of the cooccurrence statistics of each word within a given sample of text. The structure to be preserved by the embedding function f is then (informally) that words with similar cooccurrence statistics should have similar vectors in the image Y, typically analyzed in terms of cosine or Euclidean distance. By embedding X in a real-valued, low-dimensional range Y, we can improve generalization and computational efficiency over sparse, symbolic cooccurrence statistics while approximating the linguistic structure posed by the distributional hypothesis (Hinton, 1986). A similar argument can be made for the knowledge base case, where the information encoded in X is the propositions in the knowledge base regarding each unit, and f is designed such that the image Y preserves local similarity between nearby points in X.

Importantly, any finite sample of language will exhibit some degree of bias, which will therefore be reflected in the learned representations. In the simplest case, any finite sample of arbitrary size has non-zero probability of not including at least one word in a language, limiting the vocabulary coverage for representation learning (Baayen, 2001). More importantly, the sample will only be able to cover a finite variety of topics and domains–and in practice, most corpora represent only a small number of domains, such as prose literature (Francis, 1964), journalistic text (Marcus et al., 1993), or scientific literature (Kim et al., 2003). Thus, representations learned from such a sample will only reflect the statistics of language use within that particular sample. Though this limits generalizability of any learned representations, it also

presents an opportunity to use this property to study linguistic differences between samples.

4.1.3 Embedding the true distribution

Finally, in the ideal scenario, the embedding domain X represents all information about a particular language: its grammar, lexical semantics, compositional rules, etc. The embedding function f would then map all of this information (within the units of interest) to the real-valued domain Y while so as to preserve all of these inter-unit relationships.

It is questionable whether or not representing this information within a formal system is possible (and Gödel's incompleteness theorem suggests it is not), and it is certainly beyond current technical means; nonetheless, it is a useful concept for what generalizable representations are intended to approximate. Pre-trained language model representations estimated from very large, diverse corpora are used in NLP research and applications precisely because they are assumed to be *sufficiently representative* of general language use to inform models of a particular linguistic task.

4.1.4 Embedded representations are not (typically) formal embeddings

In practice, neither of the criteria for an embedding are strictly met by current representation learning methods. Injectivity is not typically implemented as a strict constraint (and would impose an additional computational burden linear in vocabulary size to include), though the uncountably infinite nature of real space means that it almost always emerges regardless. Preservation of structure would require formally defining what structure in the domain is to be preserved, which is trivial in

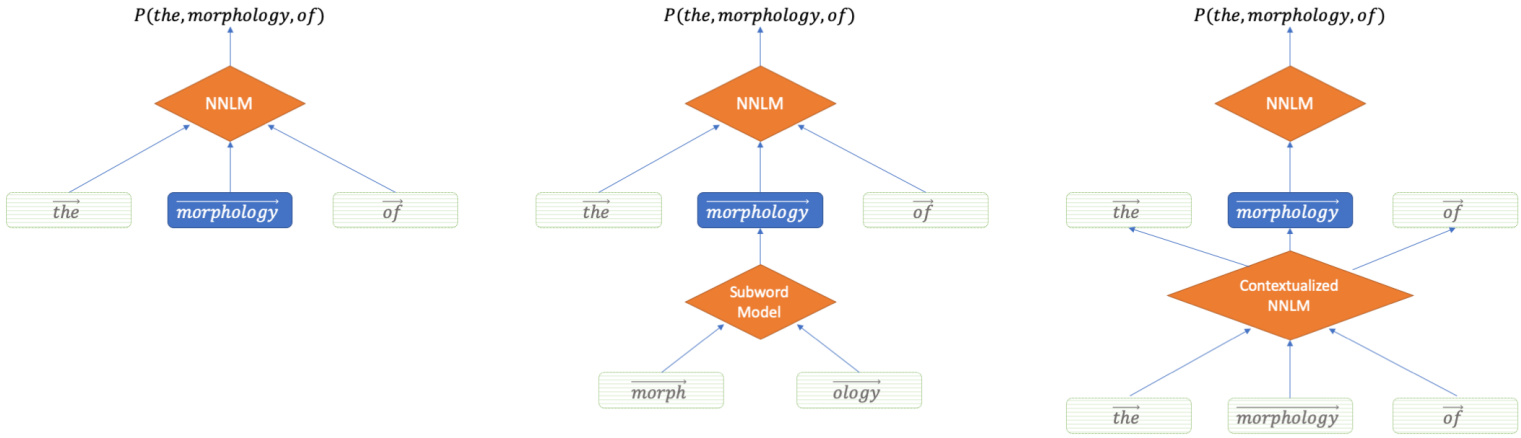
the first scenario and ill-defined in the second and third. Nonetheless, the informal idea of a structure-preserving map is key to conceptualizing representation learning techniques, and an important aspect of their role in characterizing language use.

4.1.5 Usage of the term "embedding" in practice

Mathematically, an *embedding* is a function, as described above. However, as it is often useful to discuss the representations of specific objects (words, phrases, etc. in NLP, or images or other data elswhere in AI research), the term has also come to refer to the *image* of a single point in an embedding function. Thus, learned representations of words are typically referred to as *word embeddings*, representations of sentences as *sentence embeddings*, etc. We follow this convention throughout the rest of this thesis and use "embedding" and "[embedded] representation" interchangeably.

4.2 Neural methods for word representations

Thorough surveys of word representation methods have been provided by Turney and Pantel (2010), for feature-based and early neural representations, and more recently by Camacho-Collados and Pilehvar (2018), who describe and compare a variety of recent methods for learning representations of both words and word senses. In this section, we identify three broad shifts in representation methodologies with direct implications for capturing domain semantics, and refer the interested reader to the earlier surveys for more in-depth discussion. We illustrate the general intuitions behind each of these three families of methods for word representation in Figure 4.2.



(a) Static embedding model (b) Subword embedding model (c) Contextualized embedding model

Figure 4.2: Illustration of the intuitions behind three families of word representation methods, highlighting representation of the word *morphology* in the context "The morphology of". Static representations (Figure 4.2a) learn a lexicalized vector for each word. Subword models (Figure 4.2b) learn lexicalized representations of subword units, which are dynamically combined to represent words. Contextualized models (Figure 4.2c) learn to dynamically combine representations of each of the words in a sequence, yielding context-sensitive representations to be utilized in downstream applications. The illustrations provided are not formal model structures, and do not represent any particular model, but are meant as aids for understanding the design philosophy of specific models.

4.2.1 From engineered features to learned vectors: the development of distributional semantics

Hinton (1986) proposed representing words and concepts with statistically-estimated neuronally-inspired vectors, such that the "activation patterns" (i.e., vector values) correlated with an explicit knowledge source such as a knowledge base of inter-concept relationships. However, estimating these representations was computationally demanding and difficult to model effectively, meaning that vector models of language remained primarily symbolic in nature (Turney and Pantel, 2010). Deerwester et al. (1990) developed a matrix factorization approach to word representation, utilizing word-document occurrences, and showed that the resulting low-dimensional representations improved generalization in information retrieval. Bengio et al. (2003) expanded this idea by utilizing advances in neural network-based modeling techniques and computational efficiency, proposing a Feed-Forward Neural Network (RNN) representation model. This approach moved from from document-level information to word sequences by leveraging a *language modeling* objective, which models the likelihood of each word in a sequence conditioned on the preceding words. Prior language models had calculated this likelihood based on counts of word subsequences (n-gram models); Bengio et al. (2003) modeled the likelihood using a neural network with a fixed number of context words as input.

Morin and Bengio (2005) improved the computational efficiency of the Bengio et al. (2003) model using a hierarchical decomposition of the conditional probability calculations, though at the expense of model accuracy; a later improvement by Mnih and Hinton (2008) yielded both more efficient and more accurate modeling. All of these models used a fixed context window to model the conditional probability of each

word, which negatively impacts some sentence-level NLP tasks such as semantic role labeling; Collobert and Weston (2008); Collobert et al. (2011) therefore proposed a Convolutional Neural Network (CNN) model, which processed a complete sentence at a time and was trained using a variety of objectives, including part of speech tagging, named entity recognition, language modeling, and others. Mikolov et al. (2010) then expanded the idea of processing a larger amount of context, and proposed a Recurrent Neural Network (RNN) language model, in which the probability of each word was conditioned on the entire preceding sequence. This additional information improved the efficacy of the word representations as features in downstream applications.

However, these complex neural models remained expensive to train, requiring significant computational power and time to identify useful parameters. Mikolov et al. (2013a) returned to the idea of a bounded context window and proposed word2vec, a log-bilinear neural representation model. By modeling the likelihood of each word relative to a direct vector similarity calculation between a word and the words co-occurring around it, and approximating normalization of these likelihoods over the full vocabulary using gradient reversal with a small set of negative samples, word2vec approximates factorization of the word cooccurrence matrix (Levy and Goldberg, 2014c). This yields a pseudo language model utilizing learned representations that is easy and fast to train over billions of words, and the utility of these representations in downstream NLP tasks has led to Mikolov et al. (2013a) being one of the highest-cited papers in NLP literature. Subsequent work largely focused on improvements to this modeling strategy: Levy and Goldberg (2014a) incorporated syntactic dependencies into word2vec learning, and Pennington et al. (2014) incorporated insights

from an early neural network-based paper (Huang et al., 2012) to condition the word representations jointly on local and global contexts.

What do they offer for semantic analysis?

The family of neural language model approaches discussed above, which have come to be referred to as *static* word embeddings, yield representations of words that aim to capture the cooccurrence distribution of a word by means of vector similarity (Figure 4.2a). This type of information, drawing on the distributional hypothesis (Harris, 1954), has been termed *distributional semantics* in linguistics (Lenci, 2018; Boleda, 2020). Through a relative organization of word representations that correlates with cooccurrence patterns, distributional semantic models provide in essence a high-level snapshot of language use within the corpus they were trained on. Issues of meaning conflation in polysemous words arise in this setting (where all meanings of *bank*, e.g., are lumped together), and have been addressed in a variety of ways; see Camacho-Collados and Pilehvar (2018) for a review. Nonetheless, as discussed in Section 4.6, these strategies provide a powerful tool for summarizing patterns in word usage within a specific sample of language.

4.2.2 Sub-word modeling for morphology and generalization

Lexicalized representation models of the sort proposed by Bengio et al. (2003) and Mikolov et al. (2013a) learn to represent words as, effectively, independent arbitrary symbols. Thus, the rich linguistic information captured by *morphology* is ignored, meaning that (absent preprocessing) there is no direct commonality between the vectors for *run* and *runs*, or *swimming* and *writing*. While post-processing methods can encourage some of these relationships *post hoc* (Faruqui et al., 2015), capturing some

degree of morphology in word representations offers potential for both linguistic analysis and generalization across words, particularly in multilingual settings. Cotterell and Schütze (2015) augment the log-bilinear model of word2vec with morphological information, in order to encourage morphologically-related words to have similar representations, and demonstrate significant improvement on morphological tasks in German (a morphologically rich language). Kim et al. (2016b) take a more extreme approach and model word representations by using a CNN to combine representations of each *character* in a word; these word representations are then fed into a Long Short Term-Memory (LSTM) recurrent neural language model which trains the full network. While this approach is limited to languages whose written form can be decomposed into sub-word characters, and it does not explicitly model morphological information, it has been shown to capture morphological relationships in multilingual settings (Cotterell and Heigold, 2017).

Character-level representations have frequently been combined with lexicalized word representations to capture both sub-word and distributional information for downstream NLP applications (Lample et al., 2016; Dernoncourt et al., 2017a). Bojanowski et al. (2017) adapted this approach in the log-bilinear setting with the FastText model, in which words are decomposed into character subsequences; lexicalized embeddings are then learned for these subsequences, which are linearly combined to represent the word (a lexicalized word embedding may also be added as a secondary representation in their approach). Data-driven identification of subword units has contributed to recent machine translation models as well: a recent version of Google's machine translation system utilized a combination of "wordpiece" subword units and lexicalized word representations (Wu et al., 2016).

What do they offer for semantic analysis?

The primary utility shown for sub-word modeling (illustrated in Figure 4.2b) thus far has been in multilingual scenarios such as translation and morphology induction. Subword features have been utilized for capturing similarities between agglomerative terminology in the biomedical domain (Zhang et al., 2019), and sub-word features have assisted in biomedical named entity recognition (Gridach, 2017; Galea et al., 2018), but their use for analyzing language in specific domains is largely unexplored. The utility of analyzing sub-word patterns in this context is likely to be limited, as most prior work on analyzing sublanguages has investigated differences in lexical *usage*–rather than semantic content–and domain-specific grammars (Grishman and Kittredge, 1986; Friedman et al., 2002). However, sub-word patterns have potential for analyzing novel linguistic patterns utilizing grapheme substitution or creative morphology (Blashki and Nichol, 2005), and may be useful for identifying relationships between domain concepts in morphologically rich languages (Laippala et al., 2009).

4.2.3 Capturing context with contextualized representations

Recently, a third significant shift has occurred in word representation methodology, with the development of *contextualized* language models, which calculate context-sensitive representations of words for use as input features in downstream applications. Under this approach (illustrated in Figure 4.2c), two instances of the word *bank* in different contexts will be represented using different vectors. This both reduces the impact of meaning conflation (Camacho-Collados and Pilehvar, 2018) and incorporates the modeling effort of word sequence composition into the generation of input features, allowing downstream NLP models to focus model capacity more directly on

the task of interest. Advances in both computational power and efficiency for back-propagation in complex neural models have enabled these developments, although the significant increase in computational demands they pose at inference time presents both a challenge for real-world applications and opportunity for further efficiency improvements.

The first general-purpose contextualized representation model was proposed by Melamud et al. (2016), who used a bidirectional Long Short Term-Memory (LSTM) network, a type of recurrent neural model capable of maintaining information over long sequences, and which processes input sequences both backwards and forwards in a joint model. Replacing static representations with these contextualized embeddings improved performance of state-of-the-art systems for a variety of NLP tasks, indicating a clear gain in broad-coverage information content. Peters et al. (2017) extend this approach for the semi-supervised setting, using contextualized embeddings pre-trained with a language modeling objective as a starting point and further training the representation model for a task of interest. In parallel, McCann et al. (2017) combined contextualized language modeling with encoder-decoder architectures used in machine translation to provide a deep contextualization model with multilingual data. In their approach, static word embeddings are passed into a learned contextualizing encoder to provide context-sensitive representations.

Deep contextualized representation models offer an additional lever for use as downstream features, in that each layer of the representation network can be taken as a different representation of a word in context. Peters et al. (2018) were the first to demonstrate this capability with ELMo, a two-layer bidirectional LSTM language model for general purpose use, in which word representations can be calculated as

a weighted combination of the layers of the representation network corresponding to each element of an input sequence. This capacity, together with pretrained ELMo models from large-scale data, made contextualized representations easy to incorporate into a variety of downstream systems, leading to rapid adoption of contextualized features. Radford et al. (2018) improved on this approach, by replacing the LSTM layers with the Transformer network architecture (Vaswani et al., 2017), which uses self-attention to improve parallelization and representation power across sequences, and incorporating the task-specific language model fine-tuning step proposed by Howard and Ruder (2018).

Both ELMo and the GPT model of Radford et al. (2018) use a language modeling objective, requiring that word sequences be processed in each direction separately (so as not to include the word to be predicted in the representations of the words around it). Devlin et al. (2019) address this issue with the BERT model, which expands GPT by replacing the language modeling objective with a *cloze* task, in which randomly-chosen words are masked in the input sequence, and the training objective is to maximize the likelihood of the original words using bidirectional information. Their approach utilized significantly deeper networks (12- and 24-layer Transformers), and has become the de facto standard for input feature representation. Two notable extensions of the BERT model have been proposed: ALBERT (Lan et al., 2020) is a distilled version of BERT yielding comparable performance with fewer parameters, for application in lower-resourced settings; and RoBERTa (Liu et al., 2019), which provides a more finely-tuned version of BERT for robust application.

The method space for contextualized representations is still an area of active research, with two further notable advances in recent months. Radford et al. (2019)

utilized multitask training to develop a larger-scale transformer model called GPT-2, yielding a generative language model with highly naturalistic output. Concurrently, Yang et al. (2019) identified two issues with BERT-style model training: a discrepancy between training and inference settings due to the presence of masked words at training time, and independence of the predictions of the masked words (effectively decoupling the language model across multiple words within a given sequence). They proposed XLNet, a transformer-based architecture which utilizes a language model objective–thus removing the discrepancies caused by BERT's cloze task–but retains bidirectional context by training with random permutations of input sequence order.

What do they offer for semantic analysis?

Contextualized models are relatively new, and have yet to be broadly deployed for analyzing language in specific domains. A number of studies have investigated different kinds of social biases encoded in contextualized representations: Zhao et al. (2019) demonstrate evidence of gender bias in ELMo representations, and Tan and Celis (2019) expand this analysis to include multiple types of social biases. A growing literature studies linguistic correlates of different components of contextualized models: Liu et al. (2018) demonstrate that LSTM networks using language data retain sequence information across longer sequences than with non-language data, and Clark et al. (2019) show that attention heads in BERT correlate with specific syntactic constructs. At a broader level, Jawahar et al. (2019) and Tenney et al. (2019) demonstrate that layer-wise representations in BERT correlate with different NLP tasks such as syntactic parsing and named entity recognition.

While contextualized models have not yet been used to study the *properties* of language in specific domains, they offer intriguing possibilities for analysis. Bryden

et al. (2013) and Tamburrini et al. (2015) note distinctive patterns in word usage on Twitter among different communities: the context-sensitive representations of words provided by contextualized models provide a tool for investigating similar questions of community-specific patterns of word or phrase usage at scale. Contextualized representations have been shown to separate sentences into a restricted set of classes across multiple settings (Wang et al., 2019); this suggests that they could also be used to study semantic patterns in larger linguistic units within a domain, such as investigating politeness or informativity of different utterances within different communities.

4.3 Representing lexical units other than words

Creating effective representations for larger units of language, such as sentences and documents, is an essential component for a wide variety of natural language tasks. The literature on sentence- and document-level representation learning is concomitantly vast, and even a high-level summary is out of scope of this thesis. However, a few general approaches in learning representations of phrases, sentences, and documents draw on some of the word-level insights described above, and are worth discussing in brief here.

Defining operations to compose word-level representations to create representations of larger units has long been an area of research interest in NLP, including both vector-based operations (Blacoe and Lapata, 2012; Fyshe et al., 2015) and higher-order tensor algebra (Baroni and Zamparelli, 2010; Socher et al., 2012). An extensive literature has investigated recursive and recurrent neural network structures for composition (Socher et al., 2011; Kalchbrenner and Blunsom, 2013; Kiros et al., 2015),

continuing through to the current contextualized representation models (Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019). In many cases, discourse- or document-level representations are then created via hierarchical models that compose words into sentence representations, and sentences into documents (Chang et al., 2019; Zheng et al., 2019).

Learning representations of phrases has also been an area of focused research within specific domains, where standardized phrases and names often capture information of interest. Mikolov et al. (2013b) present a statistical approach for automatically identifying phrases such as named entity mentions, and treating these as unigrams for representation learning; Yin and Schütze (2014) extend this idea to a broader set of bigrams. O'Neill et al. (2017) learn representations of legal phrases using word2vec and employ them for legal document classification. Recently Phan et al. (2019) employ recurrent neural models to learn representations of standardized biomedical terms; further approaches in this vein are discussed in Chapter 5.

4.4 Interpreting learned representations in terms of natural language semantics

In feature engineering approaches to representing language, a word or other linguistic unit is typically represented with a vector of values derived from specific criteria, such as co-occurrence counts, part of speech tags, etc. In representation learning approaches, however, the learned values in representation vectors have no inherent meaning. Hinton (1986) laid out a conceptual framework for this aspect of representation learning that is worth quoting at length here:

> The real criterion for a good set of role-specific representations is that it makes it easy to express the regularities of the domain. [...] Instead of saying that activity in a unit means that the person [being represented]

> is old, we can simply specify a set of people for which the unit is active. Each unit then corresponds to a way of partitioning all the people into two subsets, and good representations are ones for which the partitions are helpful in expressing the regularities. (Hinton, 1986)

Learned representations are thus meant to model a type of associative memory: activation patterns are defined specifically such that items which are associated in the domain to be represented are assigned similar representations. Interpreting these representations therefore requires analyzing how correlations between the mathematical representations correspond to correlations in the items being represented, and vice versa. It is thus an important part of working with representation spaces to note the distinction between *correlation* and *meaning*: even when correlation is observed between the values of some subset of representation features and a linguistic phenomenon, this is not sufficient to say that the actual observed values mean something about language. That is, perturbation of feature values along a correlation trend does not signify change in the associated phenomenon: it merely may (or may not) be correlated with it in the particular space of learned representations. Analysis and interpretation is therefore *restricted* to a matter of finding correlated correlations in the representation space and the domain of objects being represented. On this understanding, the ability to map directly *back* from representation space to the object space becomes a powerful tool for analysis: Appendix A presents some initial arguments that sequence models of language may provide this capability.

This section and Section 4.5 present a brief overview of methods that have been used to analyze the semantic and task-specific correlations captured in the organization of a representation *space*; i.e., the image of an embedding function. It is important to note that, as neural representation methods have become more and

more complex, a growing literature has also begun to investigate methods for analyzing the internals of representation *methods*; e.g., intermediate spaces and operations with a composite embedding function. While discussion of these methods is out of the scope of this thesis, we refer the interested reader to Belinkov and Glass (2019) for a recent survey of methods for analyzing the representation models themselves.

4.4.1 Semantic similarity as an evaluation criterion

When it comes to representing language, the associations between linguistic units are typically modeled via a language modeling objective, as discussed in Section 4.2. In this approach, the distributional hypothesis comes into play, suggesting that the primary *regularity* to be expressed by a "good" representation model is semantic similarity.

Evaluating semantic similarity, however, is by no means straightforward. One frequently-used approach is *word pair similarity*, in which inventories of word pairs are assigned a similarity score based on the corresponding word representations, and these scores are then compared to human-assigned similarity judgments. Rubenstein and Goodenough (1965) developed a set of 65 word pairs, assigned a real-valued synonymy score between 0 and 4; this was later extended to 353 noun pairs by Finkelstein et al. (2001). Later researchers observed that these datasets included aspects of both *similarity* (i.e., sharing various properties) and *relatedness* (i.e., association) (Agirre et al., 2009; Hill et al., 2015), and proposed new datasets separating the two aspects;[12] a variety of other datasets have also been developed, targeting different types of words

[12]Interestingly, though the distributional hypothesis holds that *similarity* (in the sense of common properties) is captured by shared context environments, conflation of similarity and relatedness in learned representation spaces has been observed by multiple researchers (Faruqui et al., 2016; Gladkova and Drozd, 2016). It remains unclear which is more desirable to capture in a representation space, or whether the two ideas should be evaluated using different methods.

(Luong et al., 2013; Bruni et al., 2014; Radinsky et al., 2011), providing a robust set of samples for evaluation. With neural representations, word pairs are typically scored using the cosine similarity between the corresponding representations; these scores are then ranked and compared with the ranking by human similarity judgments (Whitaker et al., 2019).

Similarity and relatedness evaluation measures only one aspect of semantics, and suffers from a variety of issues (Gladkova et al., 2016; Faruqui et al., 2016). Following on observations of semantically-correlated clustering in representation space, a number of studies have presented word categorization via clustering as an additional semantic analysis strategy (Baroni et al., 2008; Baroni and Lenci, 2011; Whitaker et al., 2019). Nonetheless, analyzing representation spaces directly as a way of measuring the semantic information they capture remains an unclear proposition at best. This thesis proposes one alternative reframing of this analytic problem in order to enable a more direct analysis, discussed in Appendix A; however, the utility of this idea is as yet unproven.

4.4.2 Geometric translation and relational regularities: the analogy completion task

Analogical reasoning has long been a staple of computational semantics research, as it allows for evaluating how well implicit semantic relations between pairs of terms are represented in a semantic model. Mikolov et al. (2013c) observed an intriguing artifact of language model-based representations: a select set of semantic and syntactic relationships could be modeled as geometric translations in the representation space, allowing analogical reasoning through vector arithmetic. They described an *analogy completion task*, in which a system is presented with an example term pair

and a query, e.g., *London:England::Paris:*_____ , and the task is to correctly fill in the blank.

Formally, given analogy *a:b::c:d*, the task is to guess d out of the embedded vocabulary, given a, b, c as evidence. The standard methods for this task use the vector difference between embedded representations of the related pairs to rank all terms in the vocabulary by how well they complete the analogy, and choosing the best fit (Mikolov et al., 2013c; Levy and Goldberg, 2014b). The vector difference is most commonly used in one of three ways, where *cos* is cosine similarity:

$$argmax_{d \in V} \left(cos(d, b - a + c) \right) \tag{4.1}$$

$$argmax_{d \in V} \left(cos(d - c, b - a) \right) \tag{4.2}$$

$$argmax_{d \in V} \frac{cos(d, b) cos(d, c)}{cos(d, a) + \epsilon} \tag{4.3}$$

The analogy completion task is intuitively appealing, and continues to be a popular tool for evaluating learned representations (Flamholz et al., 2019; Fathiamini et al., 2019). However, the standard formulation of the task suffers from several significant flaws, discussed by Linzen (2016) and Rogers et al. (2017), among others. In Newman-Griffis et al. (2017), we described several significant assumptions in the analogy task that break down when attempting to model complex domain knowledge such as biomedical relationships. These included:

Single-Target Each analogy has only one correct answer;

Single-Relationship All the information relating a to b also relates c to d;

Informativity The relationship between the exemplar pair must be both representative and informative.

In order to adjust to help relax these assumptions, we described three different evaluation strategies allowing more information on both sides of the analogy, and proposed the use of several metrics from information retrieval to provide a more nuanced picture of analogy completion results. We introduced a biomedical analogy dataset drawn from the UMLS, in order to enable analogical reasoning evaluation for large-scale biomedical knowledge, and demonstrated that these new evaluation strategies and metrics better captured the successes and failures of learned representations at reflecting the complex nature of biomedical relationships.

Is the analogy task informative for analyzing semantic correlates in representation space?

Though the modifications described above help to reduce some of the issues identified in the analogy task, they do not fix them entirely. Further, as shown by Gladkova et al. (2016) and Newman-Griffis et al. (2017), when analogy datasets are systematically collected at scale, the standard methodologies by and large fail to identify the correct answer with any reasonable level of consistency. However, *analogical reasoning* itself is not the culprit in these cases; rather, it is somewhat fallacious to expect that complex semantic relationships can be represented with consistent geometric translations in embedding space, and it is not clear what advantage this property would provide if it were ensured. Rogers et al. (2017) identify some alternative directions to consider for analogical reasoning in representation spaces, including the use of supervised models to identify relationships that can be modeled either as hyperplane projections or non-linear transformations, as well as the development of new evaluation paradigms to capture graded relational similarity.

4.5 Analyzing the effectiveness of representation features

By and large, representation learning methods have not been designed to help study language (though they are increasingly being adopted as a tool for linguistic inquiry, as discussed in Section 4.6), but rather to capture informative features for modeling specific NLP tasks. Several researchers have in fact observed that the "intrinsic" evaluations of natural language semantics discussed in Section 4.4 are not necessarily correlated with utility of the features for downstream applications (Chiu et al., 2016b; Rogers et al., 2018). Thus, "extrinsic" evaluations, in which different learned representations are used as input features for a variety of downstream applications and compared in terms of their impact on task performance, are a more direct evaluation of the *utility* of learned representations for NLP.

Hinton's observation that "good representations are the ones for which the partitions are helpful in expressing the regularities" (Hinton, 1986) raises two questions about "good" representation methods. The first is: which regularities are being expressed through various methods of partitioning the space? This is the *intrinsic semantics* question explored by the methods discussed in Section 4.4.

The second question, then, is this: since a chosen representation space (Euclidean or otherwise) comes with its own geometric and topological properties, how do *these properties* affect the ability to reflect desirable regularities by partitioning the space? We have conducted two studies of this question, which we summarize in the following sections.

4.5.1 Topological properties of representations reflect informative relationships for downstream tasks

Newman-Griffis and Fosler-Lussier (2017) proposed a method for deriving *second-order* word representations, using the nearest neighborhood topology of pretrained representations. Given a set of learned representations for words, a nearest neighborhood graph can be induced, in which each node is a word in the embedding vocabulary, and all words are connected to their top k nearest neighbors (Figure 4.3 illustrates this process). This graph abstracts away not only from absolute feature values in the representation space, but also from global geometric organization within the space, focusing exclusively on local relationships. This graph is then passed as input to an algorithm for learning representations of graph nodes using random walks along edges in the graph (Grover and Leskovec, 2016), and the resulting representations are taken as the second-order embeddings for each word. To help control for the instability of nearest neighborhoods across different representation spaces learned from the same training data (Wendlandt et al., 2018), multiple embedding samples are used for nearest neighbor graph induction, and graph edges are weighted based on the frequency of the corresponding nearest neighbor relationships across samples.

Despite the fact that the second-order embeddings can only approximate the structure of the original representation space, we observed that they retained the majority of task performance when replacing language model-based static word representations in recent models. These results are summarized in Table 4.1. Notably, performance is degraded most significantly for the intrinsic word similarity task, suggesting that this task (which does not include a learned model on top of the representation space) is

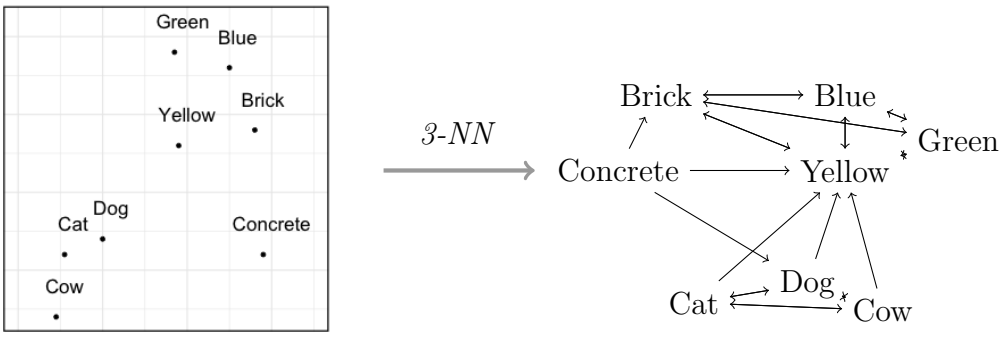


Figure 4.3: Example induction of a 3-nearest neighbor graph over an embedded vocabulary, using Euclidean distance. Note that some components may not be accessible from other components, e.g., *concrete* is inaccessible from any other vertex.

Task	Metric	First-order performance	Second-order performance	Performance delta
Named Entity Recognition (CoNLL-03)	Macro F-1	87.56	86.48	-1.08
Natural Language Inference (SNLI)	Accuracy	82.34	82.7	0.36
Paraphrase Recognition (MSRPC)	F-1	79.8	79.3	-0.5
Word similarity / relatedness (WordSim-353)	Spearman's ρ	52.2	37.9	-14.3

Table 4.1: Summary comparison of first-order and second-order (topological) representation features for NLP applications, adapted from Newman-Griffis and Fosler-Lussier (2017); for each downstream task, results are given for using unmodified, pre-trained word representations (First-order performance) and topologically-derived second-order representations (Second-order performance). The metric used for each task is provided, along with the delta observed between the two settings.

more reliant on the geometric information ablated out by the topological embedding process.

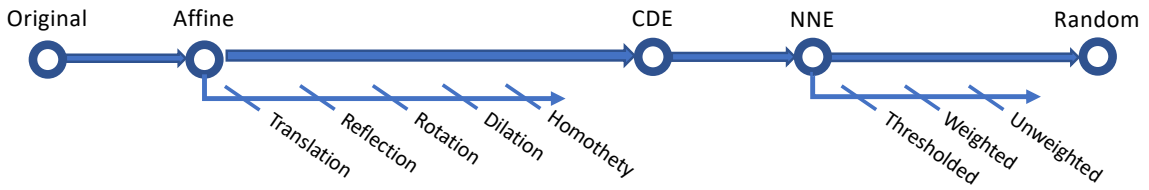


Figure 4.4: Sequence of transformations applied to word representations for geometric analysis, including transformation variants. Each transformation is applied independently to source word embeddings (i.e., transformations are not composed). CDE=Cosine Distance Encoding, NNE=Nearest Neighbor Encoding.

4.5.2 Global geometry of representations is less informative than local geometry

Following on our findings from topological analysis, we investigated the contributions of global geometric features of representation spaces on downstream task performance in Whitaker et al. (2019). In this work, we proposed a set of distinct transformations to be applied to a representation space in order to ablate different aspects of the space's geometry. These transformations, presented in Figure 4.4, included affine transforms such as translation and dilation, a new Cosine Distance Encoding (CDE) function that ablated out absolute feature values while retaining the global organization of points in the representation space, and the Nearest Neighbor Encoding (NNE) function utilized in our topological analysis.

As shown in Figure 4.5, we observed that intrinsic evaluations were somewhat sensitive to affine transformations of the representation space, and intrinsic performance degraded significantly when absolute feature values were ablated (mirroring our observations described in Section 4.5.1). For extrinsic evaluations, affine transformations made no consistent difference in performance (and in some cases improved

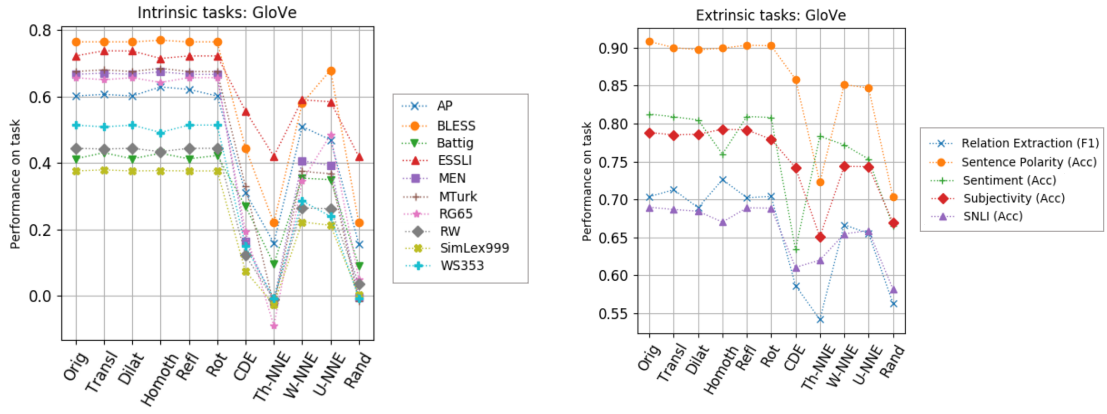


Figure 4.5: Performance metrics of geometric ablations for word representations on intrinsic and extrinsic evaluations, as applied to pre-trained GloVe (Pennington et al., 2014) word embeddings. Transformations are presented in order of decreasing geometric information about the original vectors, and are applied independently of one another to the source embeddings. For full details on tasks and transformations shown, see Whitaker et al. (2019).

it), and while the nearest neighbor encoding retained much of the original performance, the cosine distance encoding degraded it severely. Thus, we suggest that explicitly encoding global organization of learned representations introduces undesirable noise into downstream NLP models, and that local geometry is the primary carrier of information for downstream applications.

4.6 Prior uses of learned representations for sublanguage analysis

Sections 4.4 and 4.5 described methods for analyzing a learned representation space, in order to evaluate if the correlations modeled in the mathematical vector space correspond to linguistically-informative intuitions. However, as highlighted in Sections 4.1 and 4.2, learned representations can provide a powerful tool for asking

the inverse question: what do learned representation spaces tell us about the nature and usage of language?

Sublanguage analysis investigates the characteristics of language use within restricted or specialized domains (Grishman and Kittredge, 1986); one such case was described in Chapter 2, in our study of rehabilitation medicine language. Sublanguages were first defined by Harris (1968), as a subset of a more general natural language, in which a smaller set of sentences are considered valid—later researchers have also identified cases where sentences can be valid in a sublanguage but not so (or at least infelicitous) in the containing language, such as the utterances formed by noun phrases alone discussed by Friedman et al. (2002). Sublanguages have been studied within domains from weather reports, military messaging, and aerospace event reporting (Grishman and Kittredge, 1986) to code comments (Etzkorn et al., 2001), trouble tickets for utility services (Symonenko et al., 2006), and patents (Temnikova et al., 2014), but perhaps nowhere as extensively as in medical language. Identification of sublanguage patterns and characteristics has driven medical NLP from initial analysis of structured questionnaires (Sager et al., 1982) and radiology reports (Ranum, 1989; Friedman et al., 1994) through the description of biomedical literature and clinical language as distinct research targets for NLP (Friedman et al., 2002) up to expansion into social media data (Denecke, 2014; Gonzalez-Hernandez et al., 2017).

However, the majority of sublanguage analysis methods have been symbolic—as the goal is to directly analyze patterns in *language*, the difficulty in mapping back from a neural representation space to corresponding language has made the use of learned representations for sublanguage research difficult. Word representations have been shown to capture diachronic changes in word meaning and usage (Hamilton

et al., 2016b; Kutuzov et al., 2018), as well as changes in biomedical term usage (Vashisth et al., 2019) and the definitions of psychological concepts (Vylomova et al., 2019). Azarbonyad et al. (2017) adapt these ideas to study differences in word usage across political spectra, and Hamilton et al. (2016a) use domain-specific representations to identify lexica for sentiment analysis. In the medical domain, Ye and Fabbri (2018) utilize word representations from different clinical document types to identify keywords for chart review, and our own work, described in Chapter 8, uses concept representations to study differences in how different biomedical concepts are discussed among different document types. These results suggest that, though interpretation may not always be straightforward, learned representations exhibit significant potential as a tool for sublanguage analysis and adaptation.

4.7 Conclusion

Neural representation learning has become a fundamental element of modern NLP. Learned representations of language reflect a variety of different modeling approaches and task formulations, but generally draw on the *distributional hypothesis* to model words, sentences, and other linguistic units in mathematical vector spaces such that vector similarity is correlated with linguistic similarity. Analyzing and interpreting the degree of correlation with natural language semantics in these vector spaces remains an active research area, as does analysis of the geometric characteristics of learned representations. However, the utility of learned representation features for modeling various NLP tasks is increasingly being joined by demonstrated utility of using learned representations as a tool to study properties of language. In the remainder of this thesis, we present a novel approach to learning representations of

domain-specific concepts (Chapter 5), and describe several studies investigating different aspects of using learned representations to capture patterns in language use within specific domains.

Chapter 5: Learning representations of domain concepts with distant supervision

The representation learning methodologies discussed in Chapter 4 have focused largely on words, constituent phrases, and sentences. However, in specialized domains such as medicine, or sub-domains such as functional status, *domain concepts* form a significant portion of the atomic semantic units available. In medicine, these concepts include specific symptoms, diseases, and procedures; for function, concepts include activities, social situations, individual environmental factors, and so on. The important concepts within a domain are learned by humans through specialized training, and the various names for these concepts may be systematically captured in standardized vocabularies (as with many clinical terms) or have yet to be standardized (as for many aspects of functional status). These names can be ambiguous (e.g., "cold"), atomic multi-word expressions ("Lou Gehrig's disease"), or semantically-composed phrases ("diabetes mellitus, insulin-dependent"). Thus, learning representations of these concepts for use within domain-restricted NLP requires the adaptation of standard representation learning techniques to these complexities.[13]

[13]Portions of this chapter previously published in D Newman-Griffis, AM Lai, and E Fosler-Lussier. 2018. "Jointly Embedding Entities and Text with Distant Supervision." *Proceedings of the Third Workshop on Representation Learning for NLP*, 195-206.

Distributed representations of knowledge base entities and concepts are by no means new, having been used as key elements of many recent NLP applications from document ranking (Jimeno-Yepes and Berlanga, 2015) and knowledge base completion (Toutanova et al., 2015) to clinical diagnosis code prediction (Choi et al., 2016a,c). These works have taken two broad tacks for the challenge of learning to represent entities, each of which may have multiple unique surface forms in text. Knowledge-based approaches learn entity representations based on the structure of a large knowledge base, often augmented by annotated text resources (Yamada et al., 2016; Cao et al., 2017). Other methods utilize explicitly annotated data, and have been more popular in the biomedical domain (Choi et al., 2016a; Mencia et al., 2016). Both approaches, however, are often limited by ignoring some or most of the available textual information. Furthermore, such rich structures and annotations are lacking for many specialized domains, and can be prohibitively expensive to obtain.

We propose a fully text-based method for jointly learning representations of words, the surface forms of entities, and the entities themselves, from an unannotated text corpus. We use distant supervision from a *terminology*, which maps entities to known surface forms. We augment the well-known log-linear skip-gram model (Mikolov et al., 2013a) with additional term- and entity-based objectives, and evaluate our learned embeddings in both intrinsic and extrinsic settings.

Our joint embeddings clearly outperform prior entity embedding methods on similarity and relatedness evaluations. Entity and word embeddings capture complementary information, yielding improved performance when they are combined. Analogy completion results further illustrate these differences, demonstrating that entities capture domain knowledge, while word embeddings capture morphological and lexical

information. Finally, we see that an oracle combination of entity and text embeddings nearly matches a state of the art unsupervised method for biomedical word sense disambiguation that uses complex knowledge-based approaches. However, our embeddings show a significant drop in performance compared to prior work in a newswire disambiguation dataset, indicating that knowledge graph structure contains entity information that a purely text-based approach does not capture.

5.1 Related Work

Knowledge-based approaches to entity representation are well-studied in recent literature. Several approaches have learned representations from knowledge graph structure alone (Grover and Leskovec, 2016; Yang et al., 2016; Wang et al., 2017). Wang et al. (2014), Yamada et al. (2016), and Cao et al. (2017) all use a joint embedding method, learning representations of text from a large corpus and entities from a knowledge graph; however, they rely on the disambiguated entity annotations in Wikipedia to align their models. Fang et al. (2016) investigate heuristic methods for joint embedding without annotated entity mentions, but still rely on graph structure for entity training.

The robust terminologies available in the biomedical domain have been instrumental to several recent annotation–based approaches. De Vine et al. (2014) use string matching heuristics to find possible occurrences of known biomedical concepts in literature abstracts, and use the sequence of these noisy concepts (without the document text) as input for skip-gram training. Choi et al. (2016d) and Choi et al. (2016a) use sequences of structured medical observations from patients' hospital stays for context-based learning. Finally, Mencia et al. (2016) take documents tagged with

Medical Subject Heading (MeSH) topics, and use their texts to learn representations of the MeSH headers. These methods are able to draw on rich structured and semi-structured data from medical databases, but discard important textual information, and empirically are limited in the scope of the vocabularies they can embed.

5.2 Methods

In order to jointly learn entity and text representations from an unannotated corpus, we use distant supervision (Mintz et al., 2009) based on known *terms*, strings which can represent one or more entities. The mapping between terms and entities is many-to-many; for example, the same infection can be expressed as "cold" or "acute rhinitis", but "cold" can also describe the temperature or refer to chronic obstructive lung disease.

Mappings between terms and entities are defined by a terminology.[14] We extracted terminologies from two well-known knowledge bases:

The Unified Medical Language System (UMLS; Bodenreider, 2004); we use the mappings between concepts and strings in the MRCONSO table as our terminology. This yields 3.5 million entities, represented by 7.6 million strings in total.

Wikipedia; we use page titles and redirects as our terminology. This yields 9.7 million potential entities (pages), represented by 17.1 million total strings. Table 5.1 gives further statistics about the mapping between entities and surface forms in each of these terminologies.

[14] *Terminology* is overloaded with both biomedical and lexical senses; we use it here strictly to mean a mapping between terms and entities.

	UMLS	Wikipedia
# entities	3,590,353	9,723,785
# terms	7,558,254	17,147,756
Max terms	495	7,077
# entities represented by n terms		
$n = 1$	1,823,569 (51%)	6,828,958 (70%)
$n = 2$	894,932 (25%)	1,565,109 (16%)
$3 \leq n \leq 10$	831,494 (23%)	1,143,452 (12%)
$n > 10$	40,358 (1%)	186,266 (2%)
# terms mapping to n entities		
$n = 1$	7,473,902 (98%)	16,127,138 (94%)
$n = 2$	69,816 (1%)	958,242 (5%)
$3 \leq n \leq 10$	14,366 ($< 1\%$)	62,062 ($< 1\%$)
$n > 10$	170 ($\ll 1\%$)	15 ($\ll 1\%$)

Table 5.1: Terminologies used for JET experiments, listing statistics of the many-to-many mapping between terms and entities in each terminology (including the maximum # of terms per entity).

While iterating through the training corpus, we identify any exact matches of the terms in our terminologies.[15] We allow for overlapping terms: thus, "in New York City" will include an occurrence of both the terms "New York" and "New York City." Each matched term may refer to one or more entities; we do not use a disambiguation model in preprocessing, but rather assign a probability distribution over the possible entities.

Model

We extend the skip-gram model of Mikolov et al. (2013a), to jointly learn vector representations of words, terms, and entities from shared textual contexts. For a given target word, term, or entity v, let $C_v = c_{-k} \dots c_k$ be the observed contexts in a window of k words to the left and right of v, and let $N_v = n_{-k,1} \dots n_{k,d}$ be the d

[15]We lowercase and strip special characters and punctuation from both terms and corpus text, and then find all exact matches for the terms.

random negative samples for each context word. Then, the context-based objective for training v is

$$O(v, C_v, N_v) = \sum_{c \in C_v} \log\sigma(\vec{c} \cdot \vec{v}) + \sum_{n \in N_v} \log\sigma(-\vec{n} \cdot \vec{v}) \tag{5.1}$$

where σ is the logistic function.

We use a sliding context window to iterate through our corpus. At each step, the word w at the center of the window C_w is updated using $O(w, C_w, N_w)$, where N_w are the randomly-selected negative samples.

As terms are of variable token length, we treat each term t as an atomic unit for training, and set C_t to be the context words prior to the first token of the term and following the final token. Negative samples N_t are sampled independently of N_w.

Finally, each term t can represent a set of entities E_t. Vectors for these entities are updated using the same C_t and N_t from t. Since the entities are latent, we weight updates with uniform probability $|E_t|^{-1}$; attempts to learn this probability did not produce qualitatively different results from the uniform distribution. Thus, letting T be the set of terms completed at w, the full objective function to maximize is:

$$\begin{aligned} \hat{O} = {} & O(w, C_w, N_w) + \\ & \sum_{t \in T} \Big[O(t, C_t, N_t) + \sum_{e \in E_t} \frac{1}{|E_t|} O(e, C_t, N_t) \Big] \end{aligned} \tag{5.2}$$

Term and entity updates are only calculated when the final token of one or more terms is reached; word updates are applied at each step. To assign more weight to near contexts, we subsample the window size at each step from $[1, k]$.

	Pubmed	Wikipedia	Gigaword
# tokens	2.6B	1.9B	4.3B
# mentions	1.5B	1.4B	3.2B
Avg CP	2.54	1.01	1.01
% of entities by polysemy impact			
$CP \geq 1$	99.1%	98.6%	98.8%
$CP \geq 2$	9.3%	3.5%	2.2%
$CP \geq 10$	0.3%	0%	$\ll 0.1\%$

Table 5.2: Training corpora used for JET embedding experiments. # mentions is the number of exact matches found for terms in the relevant terminology. CP = corpus polysemy of a given entity. B = billion.

5.2.1 Training corpora

We train embeddings on three corpora. For our biomedical embeddings, we use 2.6 billion tokens of biomedical abstract texts from the 2016 PubMed baseline (1.5 billion noisy annotations). For comparison to previous open-domain work, we use English Wikipedia (5.5 million articles from the 2018-01-20 dump); we also use the Gigaword 5 newswire corpus (Parker et al., 2011), which does not have gold entity annotations.

As our model does not include a disambiguation module for handling ambiguous term mentions, we also calculate the expected effect of polysemous terms on each entity that we embed using a given corpus. We call this the entity's *corpus polysemy*, and denote it with $CP(e)$. For entity e with corresponding terms T_e, $CP(e)$ is given as

$$CP(e) = \sum_{t \in T_e} \frac{f(t)}{Z} \text{polysemy}(t) \tag{5.3}$$

where $f(t)$ is the corpus frequency of term t, Z is the frequency of all terms in T_e, and polysemy(t) is the number of entities that t can refer to.

Table 5.2 breaks down expected polysemy impact for each corpus. The vast majority of entities experience some polysemy effect in training, but very few have an average ambiguity per mention of 50% or greater. Most entities with high corpus polysemy are due to a few highly ambiguous generic strings, such as *combinations* and *unknown*. However, some specific terms are also high ambiguity: for example, *Washington County* refers to 30 different US counties.

5.2.2 Hyperparameters

For all of our embeddings, we used the following hyperparameter settings: a context window size of 2, with 5 negative samples per word; initial learning rate of 0.05 with a linear decay over 10 iterations through the corpus; minimum frequency for both words and terms of 10, and a subsampling coefficient for frequent words of 1e-5.

5.2.3 Baselines

We compare the words, terms,[16] and entities learned in our model against two prior biomedical embedding methods, using pretrained embeddings from each. De Vine et al. (2014) use sequences of automatically identified ambiguous entities for skip-gram training, and Mencia et al. (2016) use texts of documents tagged with MeSH headers to represent the header codes. The most recent comparison method for Wikipedia entities is MPME (Cao et al., 2017), which uses link anchors and graph structure to augment textual contexts. We also include skip-gram vectors as a final baseline; for Pubmed, we use pretrained embeddings with optimized hyperparameters from Chiu et al. (2016a), and we train our own embeddings with word2vec for both Wikipedia and Gigaword.

[16]Unknown terms were handled by backing off to words.

Method	Full		Filtered	
	Sim	Rel	Sim	Rel
Prior work				
word2vec	0.559	0.496		
DeVine'14	0.455	0.422	0.534	0.482
Mencia'16	0.565	0.534	0.573	0.536
Proposed				
Word	0.561	0.490		
Term	0.619	0.557*		
Entity	0.633*	0.563*	0.614*	0.567*
Entity+Word	0.653*	0.586*	0.615*	**0.583***
+Cross	**0.662***	**0.588***	**0.622***	0.573*

Table 5.3: Spearman's ρ results from JET experiments on UMNSRS similarity/relatedness dataset. Filtered results indicate performance on the shared-vocabulary subset. *=significantly better ($p < 0.05$) than word baseline (full), DeVine et al (filtered).

5.3 Evaluations

Following Chiu et al. (2016b), Cao et al. (2017), and others, we evaluate our embeddings on both intrinsic and extrinsic tasks. To evaluate the semantic organization of the space, we use the standard intrinsic evaluations of similarity and relatedness and analogy completion. To explore the applicability of our embeddings to downstream applications, we apply them to named entity disambiguation. Results and analyses for each experiment are discussed in the following subsections.

5.3.1 Similarity and relatedness

We evaluate our biomedical embeddings on the UMNSRS datasets (Pakhomov et al., 2010), consisting of pairs of UMLS concepts with judgments of similarity (566 pairs) and relatedness (587 pairs), as assigned by medical experts. For evaluating our Wikipedia entity embeddings, we created WikiSRS, a novel dataset of similarity

and relatedness judgments of paired Wikipedia entities (people, places, and organizations), as assigned by Amazon Mechanical Turk workers. We followed the design procedure of Pakhomov et al. (2010) and produced 688 pairs each of similarity and relatedness judgments; for further details on this dataset, please see Newman-Griffis et al. (2018).

For each labeled entity pair, we calculated the cosine similarity of their embeddings, and ranked the pairs in order of descending similarity. We report Spearman's ρ on these rankings as compared to the ranked human judgments: Table 5.3 shows results for UMNSRS, and Table 5.4 for WikiSRS.

As the dataset includes both string and disambiguated entity forms for each pair, we evaluate each type of embeddings learned in our model. Additionally, as words and entities are embedded in the same space (and thus directly comparable), we experiment with two methods of combining their information. Entity+Word sums the cosine similarities calculated between the entity embeddings and word embeddings for each pair; the Cross setting further adds comparisons of each entity in the pair to the string form of the other.

Results

Our proposed method clearly outperforms prior work and text-based baselines on both datasets. Further, we see that the words and entities learned by our model include complementary information, as combining them further increases our ranking performance by a large margin. As the results on UMNSRS could have been due to our model's ability to embed many more entities than prior methods, we also filtered the dataset to the 255 similarity pairs and 260 relatedness pairs that all evaluated

Method	Wikipedia Sim	Wikipedia Rel	Gigaword Sim	Gigaword Rel
Prior work				
word2vec	0.630	0.630	0.624	0.623
MPME	0.506	0.567	–	–
Proposed				
Word	0.646	0.655	0.615	0.600
Term	0.607	0.667	0.625	0.673
Entity	0.594	0.648	0.634	0.686
Entity+Word	**0.718***	**0.754***	0.701*	0.722*
+Cross	0.697*	0.753*	0.695*	0.729*

Table 5.4: Spearman's ρ results for JET experiments on WikiSRS similarity/relatedness dataset, training on two corpora. All Proposed results are significantly better than MPME; *=significantly better than strongest word-level baseline ($p < 0.05$).

entity-level methods could represent;[17] Table 5.3 shows similar gains on this even footing. We follow Rastogi et al. (2015) in calculating significance, and use their statistics to estimate the minimum required difference for significant improvements on our datasets.

In UMNSRS, we found that cosine similarity of entities consistently reflected human judgments of similarity better than of relatedness; this reflects previous observations by Agirre et al. (2009) and Muneeb et al. (2015). Interestingly, we see the opposite behavior in WikiSRS, where relatedness is captured better than similarity in all settings. In fact, we see a number of errors of relatedness in WikiSRS predictions, e.g., "Hammurabi I" and "Syria" are marked highly similar, while the composers "A.R. Rahman" and "John Phillip Sousa" are marked dis-similar. MPME embeddings tend towards over-relatedness as well (e.g., ranking "Richard Feynman" and "Paris-Sorbonne University" much more highly than gold labels). Despite better

[17]For WikiSRS, all methods covered all pairs.

Dataset	Words	Entities	Entity+Word+Cross
UMNSRS	Iron/Iron	Iron/Iron	Levaquin/Avelox
	Nausea/Vomiting	Sinemet/Sinemet	Enalapril/Lisinopril
	Lipitor/Zocor	Enalapril/Lisinopril	Carboplatin/Cisplatin
WikiSRS	Minas Tirith/Minas Morgul	Real Madrid/FC Barcelona	Ferrari/Lamborghini
	Moscow/Moscow Kremlin	Minas Tirith/Minas Morgul	Moscow/Moscow Kremlin
	Norway/Denmark	Charlize Theron/Screen Actor's Guild	Toshiro Mifune/Akira Kurosawa

Table 5.5: Top-ranked pairs in UMNSRS and WikiSRS, using different JET features.

similarity performance, this trend of over-relatedness also holds in biomedical embeddings: for example, C0027358 *Narcan* and C0026549 *Morphine* are consistently marked highly similar across embedding methods, even though Narcan blocks the effects of opioids like morphine.

Comparing entities and words

We observe clear differences in the rankings made by entity vs word embeddings. As shown in Table 5.5, highly related entities tend to have high cosine similarity, while word embeddings are more sensitive to lexical overlap and direct cooccurrence. Combining both sources often gives the most inuitive results, balancing lexical effects with relatedness. For example, while the top three pairs by combination in WikiSRS are likely to co-occur, the top three in UMNSRS are pairs of drug choices (antibiotics,

ACE inhibitors, and chemotherapy drugs, respectively), only one of which is likely to be prescribed to any given patient at once.

These differences also play out in erroneous predictions. Entity embeddings often fix the worst misrankings by words: for example, "Tony Blair" and "United Kingdom" (gold rank: 28) are ranked highly unrelated (position 633) by words, but entities move this pair back up the list (position 86). However, errors made by entity embeddings are often also made by words: e.g., C0011175 *Dehydration* and C0017160 *gastroenteritis* are erroneously ranked as highly unrelated by both methods. Interestingly, we find no correlation between the corpus polysemy of entity pairs and ranking performance, indicating that ambiguity of term mentions is not a significant confound for this task.

5.3.2 Analogy completion

We use analogy completion to further explore the properties of our joint embeddings. Given analogy $a : b :: c : d$, the task is to guess d given (a, b, c), typically by choosing the word or entity with highest cosine similarity to $b - a + c$ (Levy and Goldberg, 2014b). We report accuracy using the top guess (ignoring a, b, and c as candidates, per Linzen, 2016).

Biomedical analogies

To compare between word and entity representations, we use the entity-level biomedical dataset BMASS (Newman-Griffis et al., 2017), which includes both entity and string forms for each analogy. In order to test if words and entities are capturing complementary information, we also include an oracle evaluation, in which an analogy is counted as correct if either words or entities produce a correct response.[18] We do

[18]We use the Multi-Answer setting for our evaluation (a single (a, b, c) triple, but a set of correct values for d).

Method	B3	H1	C6	L1	L6
Words	2.9	0.4	**7.9**	**51.5**	**69.3**
Entities	**18.3**	**22.4**	4.5	10.6	10.0
Oracle	20.7	22.9	12.1	55.0	70.9

Table 5.6: Analogy completion results for 5 relations in BMASS with greatest absolute difference in word performance vs entity performance: B3 (*gene-encodes-product*), H1 (*refers-to*), C6 (*associated-with*), L1 (*form-of*), and L6 (*has-free-acid-or-base-form*). Reported numbers are Accuracy (%); the better of word and entity performance is highlighted, and all entity vs word differences are significant (McNemar's test; $p \ll 0.01$).

not compare against prior biomedical entity embedding methods on this dataset, due to their limited vocabulary.

Table 5.6 contrasts the performance of different jointly-trained representations for five relations with the largest performance differences from this dataset. For *gene-encodes-product* and *refers-to*, both of which require structured domain knowledge, entity embeddings significantly outperform word-level representations. Many of the errors made by word embeddings in these relations are due to lexical oversensitivity: for example, in the renaming analogy *spinal epidural hematoma:epidural hemorrhage::canis familiaris:*___, words suggest latinate completions such as *latrans* and *caballus*, while entities capture the correct C1280551 *Dog*. However, on more morphological relations such as *has-free-acid-or-base-form*, words are by far the better option.

The success of the oracle combination method for entity and word predictions clearly indicates that not only are words and entities capturing different knowledge, but that it is complementary. In the majority of the 25 relations in BMASS, oracle results improved on words and entities alone by at least 10% relative. In some cases, as

with *has-free-acid-or-base-form*, one method does most of the heavy lifting. In several others, including the challenging (and open-ended) *associated-with*, entities and words capture nearly orthogonal cases, leading to large jumps in oracle performance.

General-domain analogies

No entity-level encyclopedic analogy dataset is available, so we follow Cao et al. (2017) in evaluating the effect of joint training on words using the Google analogy set (Mikolov et al., 2013a). As shown in Table 5.7, our Wikipedia embeddings roughly match MPME embeddings (which use annotated entity links) on the semantic portion of the dataset, but our ability to train on unannotated Gigaword boosts our results on all relations except *city-in-state*.[19] Overall, we find that jointly-trained word embeddings split performance with word-only skipgram training, but that word-only training tends to get consistently closer to the correct answer. This suggests that terms and entities may conflict with word-level semantic signals.

5.3.3 Entity disambiguation

Finally, to get a picture of the impact of our embedding method on downstream applications, we investigated entity disambiguation.[20] Given a named entity occurrence in context, the task is to assign a canonical identifier to the entity being referred to: e.g., to mark that "New York" refers to the city in the sentence, "The mayor of New York held a press conference." It bears noting that in unambiguous cases, a terminology alone is sufficient to link the correct entity: for example, "Barack Obama" can only refer to a single entity, regardless of context. However, many entity strings

[19]We failed to precisely replicate the analogy numbers reported by Cao et al. (2017); we attribute this primarily to the different training corpus and slightly different preprocessing.

[20]This task is also referred to as entity linking and entity sense disambiguation.

Method	Capital (common)	Capital (all)	Currency	City in State	Family
word2vec (W)	89.1	86.0	15.0	**55.5**	**82.4**
word2vec (G)	90.9	89.7	**18.4**	38.4	81.0
MPME (W)	83.6	80.5	11.9	50.6	78.9
Proposed (W)	90.1	78.7	9.1	42.5	75.5
Proposed (G)	**92.7**	**92.3**	16.4	31.3	81.6

Table 5.7: Analogy completion accuracy with JET features on the semantic relations in the Google analogy dataset. W=Wikipedia, G=Gigaword.

(e.g., "cold", "New York") are ambiguous, necessitating the use of alternate sources of information such as our embeddings to assign the correct entity.

Biomedical abstracts

We evaluate on the MSH WSD dataset (Jimeno-Yepes et al., 2011), a benchmark for biomedical word sense disambiguation. MSH WSD consists of mentions of 203 ambiguous terms in biomedical literature, with over 30,000 total instances. Each sample is annotated with the set of UMLS entities the term could refer to. We adopt the unsupervised method of Sabbir et al. (2017), which combines cosine similarity and projection magnitude of an entity representation e to the averaged word embeddings of its contexts C_{avg} as follows:

$$f(e, C_{avg}) = \cos(C_{avg}, e) \cdot \frac{||P(C_{avg}, e)||}{||e||} \tag{5.4}$$

The entity maximizing this score is predicted.

We compare against concept embeddings learned by Sabbir et al. (2017). They used MetaMap (Aronson and Lang, 2010) with the disambiguation module enabled on a curated corpus of 5 million Pubmed abstracts to create a UMLS concept cooccurrence corpus for word2vec training. As shown in Table 5.8, our method lags behind

theirs, though it clearly beats both random (49.7% accuracy) and majority class (52%) baselines. In addition, we leverage our jointly-embedded entities and words by adding in the definition-based model used by Pakhomov et al. (2016), which calculates an entity's embedding as the average of definitions of its neighbors in the UMLS hierarchy (McInnes et al., 2011). We use this alternate entity embedding in Equation 5.4 to calculate a second score that we add to the direct entity embedding score. This yields a large performance boost of over 6% absolute, indicating that using entities and words together makes up much of the gap between our distantly supervised embeddings and the external resources used by Sabbir et al. (2017). Using the definition-based method alone with our jointly-embedded words, we see a significant increase over Pakhomov et al. (2016), indicating the benefits of joint training. However, the combined entity and definition model still yields a significantly different 2% boost in accuracy over definitions alone. Finally, we evaluate an oracle combination that reports correct if either entity or definition embeddings achieve the correct result; as shown in the last row of Table 5.8, this combination outperforms the entity-only method of Sabbir et al. (2017), and approaches their state-of-the-art result that combines entity embeddings with a knowledge-based approach from the structure of the UMLS.

Specific errors shed more light on these differences. The definition-based method performs better in many cases where the surface form is a common word, such as *coffee* (68% definition accuracy vs 28% entity accuracy) and *iris* (93% definition accuracy vs 35% entity accuracy). Entities outperform on some more technical cases, such as *potassium* (74% entity accuracy vs 49% definition accuracy). Combining both approaches in the joint model recovers performance on several cases of low entity

Method	Accuracy %
Baselines	
Sabbir et al. (2017) (entities; +MetaMap)	89.3
Sabbir et al. (2017) (+MetaMap, UMLS)	**92.2**
Pakhomov et al. (2016) (words)	77.7
Proposed	
Entities	76.4
Definitions (joint words)	80.8
Entities+Definitions	82.7
Oracle (Entities—Definitions)	90.9

Table 5.8: MSH WSD disambiguation accuracy with JET features. Definitions is comparable to Pakhomov et al. (2016), using jointly-embedded words. All differences are significant (McNemar's test, $p \ll 0.01$).

accuracy; for example, joint accuracy on *coffee* is 68%, and on *lupus* (53% entity accuracy), joint performance is 60%.

Newswire entities

AIDA (Hoffart et al., 2011) is a standard dataset for entity linking in newswire, consisting of approximately 30,000 entities linked to Wikipedia page IDs. To reduce the search space, Pershina et al. (2015) provided a set of candidate entities for each mention, which we use for our experiments. The MPME model of Cao et al. (2017) achieves near state-of-the-art performance accuracy on AIDA with this candidate set, using the mention sense distributions and full document context included in the model. As our embeddings are trained without explicit entity annotations, we instead use the same cosine similarity and projection model discussed in Section 5.3.3 for this task. In contrast to our results on the biomedical data, we see performance far below the baseline on these data, as shown in Table 5.9.

Method	Accuracy %
MPME (entities; +graph structure)	**89.0**
Wikipedia	40.9
Wikipedia + mentions	44.6
Gigaword	58.0
Gigaword + mentions	63.9

Table 5.9: Entity linking accuracy on AIDA dataset, using entity embeddings trained on Wikipedia and Gigaword. All differences are significant (McNemar's test, $p \ll 0.01$).

Entity	Words	Terms	Entities	Joint
C0009443 *Common cold*	k(+)-grown	cold	C0041912 *Upper respiratory infections*	C0041912 *Upper respiratory infections*
	legionella-contaminated	short periods	C0234192 *Cold sensation*	C0234192 *Cold sensation*
	hyperinflating	changed	C0719425 *"Cold" pharmaceutical brand*	C0719425 *"Cold" pharmaceutical brand*
C0242797 *Home health aides*	homemaker-home	home health aide	C1553498 *Home health encounter*	home health aide
	voluntary-sector	home health aides	C0019855 *Home care services*	home health aides
	health/social	home health	C1317851 *Home health care specialty*	C1553498 *Home health encounter*

Table 5.10: Top 3 nearest neighbors to two UMLS CUIs using different JET features: words, terms, entities, or all three.

However, we improve this performance slightly by multiplying by the similarity between the entity embedding and the average word embedding of the mention itself; this gives us roughly a further 4% accuracy for both Wikipedia and Gigaword embeddings. Using the surface form recovers several cases where entities alone yield unlikely options, e.g. Roman-era Britain instead of the United Kingdom for *Britain*. However, it also introduces lexical errors: for example, *British* in several cases refers

to the United Kingdom, but the British people are often selected instead. We note that this extra score actually hurts performance on MSH WSD, where the terms are curated to be highly ambiguous, in contrast to the shorter contexts and clearer terms used in AIDA.

Two other issues bear consideration in this evaluation. Prior approaches to the AIDA dataset, including MPME, make use of the global context of entity mentions within a document to improve predictions; by using local context only, we observe some inconsistent predictions, such as selecting the cricket world cup instead of the FIFA competition for *world cup*, in a document discussing football. Additionally, in contrast to the MSH WSD dataset, many instances in AIDA have several highly-related candidates that introduce some confusion in our results. For example, *Ireland* could refer to the United Kingdom of Great Britain and Ireland, the island of Ireland, or the Republic of Ireland. As our embedding training does not include gold entity links, cases like this are often errors in our predictions.

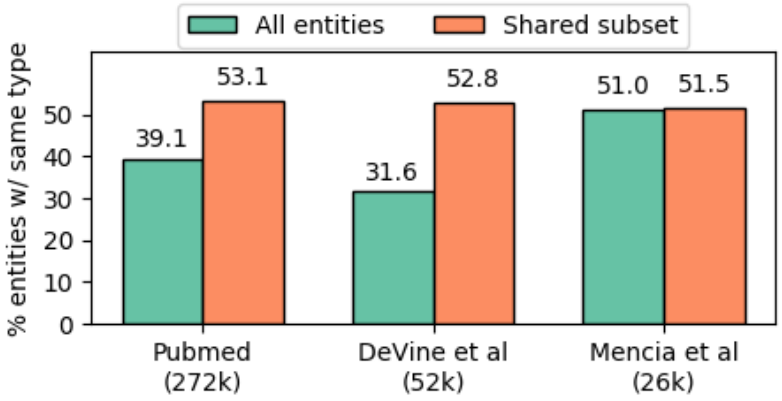


Figure 5.1: Percentage of UMLS entities whose nearest JET neighbor shares a semantic type, with no vocabulary restriction (vocab size in parentheses) and in a shared vocabulary subset.

5.4 Analysis of joint embeddings

To get a more detailed picture of our joint embedding space, we investigate nearest neighbors for each point by cosine similarity. As entities in the UMLS are assigned one or more of over 120 semantic types, we first examine how intermixed these types are in our biomedical embeddings. Figure 5.1 shows how often an entity's nearest neighbor shares at least one semantic type with it, across the three biomedical embedding methods we evaluated. As each set of embeddings has a different vocabulary, we also restrict to the entities that all three can embed (approximately 11,000).

We see that our method puts entities of the same type together nearly 40% of the time, despite embedding over 270 thousand entities. On an even footing, our method puts types together significantly more often Mencia et al. (2016) (McNemar's; $p < 0.05$), and equivalently with De Vine et al. (2014), despite using less entity-level information in training. Within our embeddings, major biological types such as bacteria, eukaryotes, mammals, and viruses all have more than 60% of neighbors with the same type, while less structured clinical types such as Clinical Attribute and Daily or Recreational Activity are in the 10-20% range. Corpus polysemy does not appear to have any effect on this type matching (mean polysemy of 1.5 for both matched and non-matched entities).

Expanding to include the words and terms in the joint embedding space, however, we see definite qualitative effects of corpus polysemy on entity nearest neighbors. Table 5.10 gives nearest word, term, entity, and joint neighbors to two biomedical entities: C0009443 *Common cold* ($CP = 6.71$) and C0242797 *Home health aides* ($CP = 1$). For the more polysemous C0009443, where 95% of its mentions are of the word "cold" (polysemy=7), word-level neighbors are mostly nonsensical, while term

neighbors are more logical, and entity neighbors reflect different senses of "cold". By contrast, the non-polysemous C0242797, which is represented by 14 different unambiguous strings, words, terms, and entities are all very clearly in line with the theme of home health aides. Notably, the common and unambiguous terms for C0242797 are its nearest neighbors out of all points, while only two of the top 10 neighbors to C0009443 are terms.

5.5 Discussion

Faruqui et al. (2016) observe that similarity and relatedness are not clearly distinguished in semantic embedding evaluations, and that it is unclear exactly how vector-space models should capture them. We see more evidence of this, as cosine similarity seems to be capturing a mix of the two properties in our data. This mix is clearly informative, but it empirically favors relatedness judgments, and cosine similarity is insufficient to separate the two properties.

Corpus polysemy plays a qualitative role in our embedding model, but less of a quantitative one. It does not correlate with similarity and relatedness judgments or entity disambiguation decisions, but it clearly affects the organization of the embedding space, by embedding entities with high corpus polysemy in less coherent areas than those with low polysemy. Linzen (2016) points out that for analogy completion, local neighborhood structure can interfere with standard methods; how this neighborhood structure affects predictions in more complex tasks is an open question.

Overall, we find two main advantages to our model over prior work. First, by only using a terminology and an unannotated corpus, we are able to learn entity embeddings from larger and more diverse data; for example, embeddings learned

from Gigaword (which has no entity annotations) outperform embeddings learned on Wikipedia in most of our experiments. Second, by embedding entities and text into a joint space, we are able to leverage complementary information to get higher performance in both intrinsic and extrinsic tasks; an oracle model nearly matches a state-of-the-art ensemble vector and knowledge-based model for biomedical word sense disambiguation. However, our other entity disambiguation results demonstrate that there is additional entity-level information that we are not yet capturing. In particular, it is unclear whether our low performance on disambiguating newswire entities is due to a disambiguation model mismatch, a lack of information in our embeddings, or a combination of both.

5.6 Conclusions

JET learns interoperable representations of language at three levels: (1) general lexical items (i.e., words); (2) lexemes of interest in a specific domain or application (i.e., terms); and (3) concepts of interest (e.g., entities in a knowledge base). This enables direct analysis of the correspondences between these levels of lexical semantics, such as comparison of highly representative terms for a given concept. Further, our method leverages distant supervision to learn context-based representations from any arbitrary text corpus, without the need for expensive manual annotations of concept mentions, based only on a curated list of terms of interest and the concepts they refer to. This makes it a powerful tool for investigating language use in new domains, which may lack for annotated data sets or well-developed knowledge resources, and for dynamically analyzing language within different or rapidly-evolving communities.

In this chapter, we demonstrated that JET embeddings capture similarity and relatedness between both biomedical concepts and encyclopedia entities better than prior methods using additional structured resources, and that they approach the state-of-the-art performance for unsupervised biomedical word sense disambiguation yielded by sophisticated knowledge-based methods. We further showed that the different levels of lexical semantics in our model capture complementary information for semantic analysis. In Chapter 7, we expand our work on word sense disambiguation and describe a method to combine multiple sources of learned concept representations to normalize concept mentions within a specific domain. In Chapter 8, we describe a concrete application of JET to linguistic analysis, by demonstrating that JET representations learned from different clinical document collections capture clinically-relevant distinctions in reference to symptoms, diagnoses, and procedures.

We have released a publicly-availble implementation of our method,[21] along with the source code used for our evaluations and our pretrained entity embeddings. Our novel Wikipedia similarity and relatedness datasets are available at the same source.

[21] `github.com/OSU-slatelab/JET`

Part III
Applications to Domain Semantics

The value of representation learning technologies for specific domains lies in their downstream utility, in engineering applications or scientific inquiry. Part I laid out characteristics of functional status information and the clinical genre that pose challenges for NLP analysis, and Part II described the theory and intuitions underlying the potential of representation learning to address these challenges. In the third part of this thesis, we synthesize these directions to demonstrate the utility of learned representations for capturing domain semantics, illustrated through three broad studies. In Chapter 6, we use the contextual information captured by learned representations as a starting point for automatically extracting FSI reports regarding mobility activities. We demonstrate that in-domain representations learned from small amounts of data are equally informative for this extraction as representations from large-scale out-of-domain data, and we describe our novel HARE system, a simple extraction tool based on learned representations that achieves high recall on extracting mobility reports from multiple datasets In Chapter 7, we present PROSE, a new model of

semantic grounding that leverages the strengths offered by different methods of representing concepts of interest (including our JET method from Chapter 5) via a context-sensitive projection of concept representations into an interoperable space for analysis. We demonstrate that by combining multiple strategies for representing concepts, PROSE achieves strong performance on diverse focused semantic tasks, including lexical word sense disambiguation, normalization of medical concept mentions in clinical language, and classifying activity types described in mobility reports. Finally, in Chapter 8, we use JET to capture differences in usage patterns of medical concepts between different clinical specialties, and demonstrate that differences in nearest neighborhood structure reflect the different topical and conceptual focuses of the respective specialties. Taken together, these studies clearly illustrate the role of representation learning in capturing the semantics of specialized domains, both in engineering systems to automatically analyze domain information and in exploring and describing the characteristics that define new domains.

Chapter 6: The Value of Domain-Sensitive Representations for Extracting Functional Status Information

In order to explore the utility of representation learning techniques for capturing semantic information in restricted domains, we start by investigating their role in *locating* semantically-relevant information for the domain. We take as a case study *mobility activity reports*, functional status information with high relevance to rehabilitation medicine and disability programs. This chapter describes three studies using representation learning to identify mobility activity reports in EHR data.[22] Section 6.1 demonstrates that domain-relevant representations, learned from very small amounts of text data, provide equal or greater utility to representations learned from large-scale out-of-domain data when provided as input features to an off-the-shelf information extraction model. Section 6.2 addresses the challenges of low coverage and syntactic complexity observed in this work, proposing a new model for high-coverage information extraction, with accompanying software for visualization and qualitative

[22]Portions of Section 6.1 have been previously published in D Newman-Griffis and A Zirikly. 2018. "Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility." *Proceedings of the BioNLP 2018 Workshop*, 1-11. Portions of Section 6.2 have been published in D Newman-Griffis and E Fosler-Lussier. 2019. "HARE: A Flexible Highlighting Annotator for Ranking and Exploration." *Proceedings of the 2019 Conference on Empirical Methods for Natural Language Processing: Systems Demonstrations*, 85-90. Portions of Section 6.3 have previously been submitted for publication and are currently in revision.

analysis of model predictions. Finally, Section 6.3 describes application of this high-coverage model to real-world, heterogeneous EHR data collected by the U.S. Social Security Administration for purposes of adjudicating disability benefit claims, and demonstrates that our tool has significant potential to support the efficience of this adjudication process. To our knowledge, the studies described in this chapter are the first investigations into automatically recognizing functional status information in EHR text.

6.1 The tradeoff between representativeness and corpus size in choosing repreesentation features for extracting mobility reports

Thieu et al. (2017) introduced a dataset of EHR documents annotated for descriptions of patient mobility status, one area of activity in the ICF. Automatically recognizing these descriptions faces significant challenges, including their length and syntactic complexity and a lack of terminological resources to draw on. In this study, we view this task through the lens of Named Entity Recognition (NER), as recent work has illustrated the potential of using Recurrent Neural Network (RNN) NER models to address similar issues in biomedical NLP (Xia et al., 2017; Dernoncourt et al., 2017b; Habibi et al., 2017).

An additional strength of RNN models is their ability to leverage pretrained word embeddings, which capture co-occurrence information about words from large text corpora. Prior work has shown that the best improvements come from embeddings trained on a corpus related to the target domain (Pakhomov et al., 2016). However, free text describing patient functioning is hard to come by: for example, even the large MIMIC-III corpus (Johnson et al., 2016) includes only a few hundred documents from

therapy disciplines among its two million notes. While recent work suggests that using a training corpus from the target domain can mitigate a lack of data (Diaz et al., 2016), even a careful corpus selection may not produce sufficient data to train robust word representations.

In this study, we explore the use of an RNN model to recognize descriptions of patient mobility. We analyze the impact of initializing the model with word embeddings trained on a variety of corpora, ranging from large-scale out-of-domain data to small, highly-targeted in-domain documents. We further explore several domain adaptation techniques for combining word-level information from both of these data sources, including a novel nonlinear embedding transformation method using a deep neural network.

We find that embeddings trained on a very small set of therapy encounter notes nearly match the mobility NER performance of representations trained on millions of out-of-domain documents. Domain adaptation of input word embeddings often improves performance on this challenging dataset, in both precision and recall. Finally, we find that simpler adaptation methods such as concatenation and preinitialization achieve highest overall performance, but that nonlinear mapping of embeddings yields the most consistent performance across experiments. We achieve a best performance of 69% exact match and over 83% token-level match F-1 score on the mobility data, and identify several trends in system errors that suggest fruitful directions for further research on recognizing descriptions of patient functioning.

6.1.1 Related work

The extraction of named entities in free text has been one of the most important tasks in NLP and information extraction (IE). As a result, this track of research has matured over the last two decades, especially in the newswire domain for high resource languages such as English. Many of the successful existing NER systems use a combination of engineered features trained using conditional random fields (CRF) model (McCallum and Li, 2003; Finkel et al., 2005). NER systems have also been widely studied in medical NLP, using dictionary lookup methods (Savova et al., 2010), support vector machine (SVM) classifiers (Kazama et al., 2002), and sequential models (Tsai et al., 2006; Settles, 2004). In recent years, deep learning models have been used in NER with successful results in many domains (Collobert et al., 2011). Proposed neural network architectures included hybrid convolutional neural network (CNN) and bi-directional long-short term memory (Bi-LSTM) as introduced by Chiu and Nichols (2016). State-of-the-art NER models use the architecture proposed by Lample et al. (2016), a stacked bi-directional long-short term memory (Bi-LSTM) for both character and word, with a CRF layer on the top of the network. In the biomedical domain, Habibi et al. (2017) used this architecture for chemical and gene name recognition. Liu et al. (2017) and Dernoncourt et al. (2017a) adapted it for state-of-the-art note deidentification. In terms of functioning, Kukafka et al. (2006) and Skube et al. (2018) investigate the presence of functioning terminology in clinical data, but do not evaluate it from an NER perspective.

Evaluation:

[Scoring: 1=totally dependent, 2=requires assistance, 3=requires appliances, 4=totally independent]$_{\text{ScoreDefinition}}$.

[Ambulation: 4]$_{\text{Mobility}}$
Observations:
Pt is weight bearing: [she ambulates independently w/o use of assistive device]$_{\text{Mobility}}$.
Limited to very brief examination.

Figure 6.1: Synthetic document with examples of ScoreDefinition (in blue) and Mobility (in orange).

6.1.2 Data

Thieu et al. (2017) presented a dataset of 250 de-identified EHR documents collected from Physical Therapy (PT) encounters at the Clinical Center of the National Institutes of Health (NIH). These documents, obtained from the NIH Biomedical Translational Research Informatics System (BTRIS; Cimino and Ayres 2010), were annotated for several aspects of patient mobility, a subdomain of functioning-related activities defined by the ICF; we therefore refer to this dataset as BTRIS-Mobility. We focus on two types of contiguous text spans: mobility-related activity reports, here called Mobility entities, and measurement scales related to mobility activity, which we refer to as ScoreDefinition entities.

Two major differences stand out in BTRIS-Mobility as compared with standard NER data. The entities, defined for this task as contiguous text spans completely describing an aspect of mobility, tend to be quite long: while prior NER datasets such as the i2b2/VA 2010 shared task data (Uzuner et al., 2011) include fairly short entities (2.1 tokens on average for i2b2), Mobility entities are an average of 10 tokens long, and ScoreDefinition average 33.7 tokens. Also, both Mobility and ScoreDefinition entities tend to be entire clauses or sentences, in contrast with the constituent noun

Entity	Train	Valid	Test
Mobility	1,533	467	947
ScoreDefinition	82	24	48

Table 6.1: Distribution of Mobility and ScoreDefinition entities in BTRIS-Mobility, broken down by training, validation, and test splits. Due to the rarity of ScoreDefinition entities, we use a 2:1 split of training to test data, and hold out 10% of training data as validation.

phrases that are the meat of most NER. Figure 6.1 shows example Mobility and ScoreDefinition entities in a short synthetic document. Despite these challenges, Thieu et al. (2017) show high (> 0.9) inter-annotator agreement on the text spans, supporting use of the data for training and evaluation.

These characteristics align well with past successful applications of recurrent neural models to challenging NLP problems. For our evaluation on this dataset, we randomly split BTRIS-Mobility at document level into training, validation, and test sets, as described in Table 6.1.

Text corpora

In order to learn input word embeddings for NER, we use a variety of both in-domain and out-of-domain corpora, defined in terms of whether the corpus documents include descriptions of function. For in-domain data, with explicit references to patient functioning, we use a corpus of 154,967 EHR documents shared with us (under an NIH Clinical Center Office of Human Subjects determination) from the NIH BTRIS system.[23] A large proportion of these documents comes from the Rehabilitation Medicine Department of the NIH Clinical Center, including Physical Therapy (PT),

[23]There is no overlap between these documents and the annotated data in BTRIS-Mobility (T. Thieu, personal communication).

Occupational Therapy (OT), and other therapeutic records; the remaining documents are sampled from other departments of the Clinical Center.

Since BTRIS-Mobility is focused on PT documents, we also use a subset of this corpus consisting of 17,952 PT and OT documents. Despite this small size, the topical similarity of these documents makes them a very targeted in-domain corpus. For clarity, we refer to the full corpus as BTRIS, and the smaller subset as PT-OT.

Out-of-domain corpora

As the BTRIS corpus is considered a small training corpus for learning word embeddings, we also use three larger out-of-domain corpora, which represent different degrees of difference from the in-domain data. Our largest data source is pretrained FastText embeddings from Wikipedia 2017, web crawl data, and news documents.[24]

We also make use of two biomedical corpora for comparison with existing work. PubMed abstracts have been an extremely useful source of embedding training in biomedical NLP (Chiu et al., 2016a); we use the text of approximately 14.7 million abstracts taken from the 2016 PubMed baseline as a high-resource biomedical corpus. In addition, we use two million free-text documents released as part of the MIMIC-III critical care database (Johnson et al., 2016). Though smaller than PubMed, the MIMIC corpus is a large sample of clinical text, which is often difficult to obtain and shows significant linguistic differences with biomedical literature (Friedman et al., 2002). As MIMIC is clinical text, it is the closest comparison corpus to the BTRIS data; however, as MIMIC focuses on ICU care, the information in it differs significantly from in-domain BTRIS documents.

[24] `fasttext.cc/docs/en/english-vectors`

6.1.3 Methods

We adopt the architecture of Dernoncourt et al. (2017a), due to its successful NER results on CoNLL and i2b2 datasets. The architecture, as depicted in Figure 6.2, is a stacked LSTM composed of: i) character Bi-LSTM layer that generates character embeddings. We include this in our experimentations due to its performance enhancement; ii) token Bi-LSTM layer using both character and pre-trained word embeddings as input; iii) CRF layer to enhance the performance by taking into account the surrounding tags (Lample et al., 2016). We use the following values for the network hyperparameters, as they yielded the best performance on the validation set: i) hidden state dimension of 25 for both character and token layers. In contrast to more common token layer sizes such as 100 or 200, we found the best validation set performance for our task with 25 dimensions; ii) learning rate = 0.005; iii) patience = 10; iv) optimization with stochastic gradient descent (SGD) which showed superior performance to adaptive moment estimation (Adam) optimization technique (Kingma and Ba, 2015).

Embedding training

We use two popular toolkits for learning word embeddings: word2vec[25] (Mikolov et al., 2013a) and FastText[26] (Bojanowski et al., 2017). We run both toolkits using skip-gram with negative sampling to train 300-dimensional embeddings, and use default settings for all other hyperparameters.[27]

[25]We use word2vec modified to support pre-initialization, from `github.com/drgriffis/word2vec-r`.

[26]`github.com/facebookresearch/fastText`

[27]For PT-OT embeddings, due to the extremely small corpus size, we use an initial learning rate of 0.05, keep all words with minimum frequency 2, and train for 25 iterations.

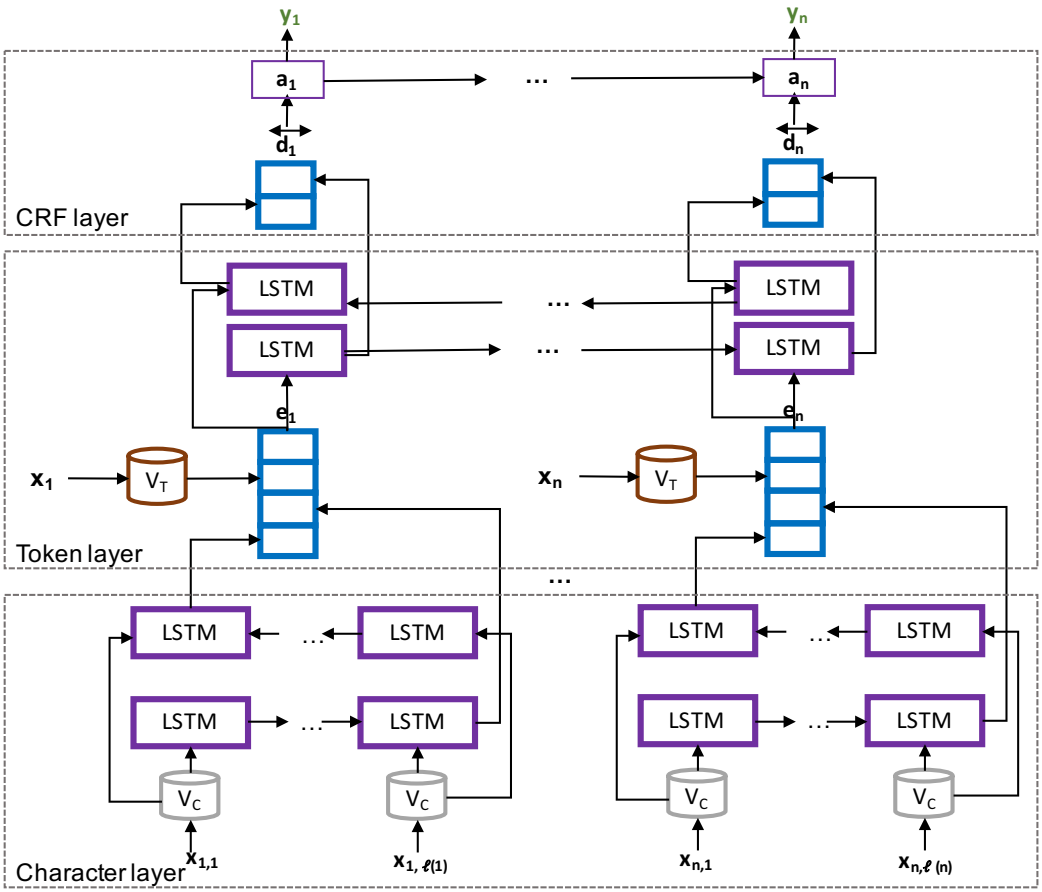


Figure 6.2: Bi-LSTM-CRF network architecture; adapted from Newman-Griffis and Zirikly (2018).

Corpus	Size	Toolkit	Mobility						ScoreDefinition					
			Exact match			Token match			Exact match			Token match		
			Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1
Random initialization			67.7	61.8	64.6	84.0	75.9	79.7	86.5	93.4	90.0	97.7	98.9	98.3
WikiNews	16B	FT	67.0	64.0	65.4	83.0	80.0	81.5	83.3	93.4	88.2	96.8	99.3	98.0
PubMed	2.6B	FT	68.7	**65.9**	67.2	82.0	84.5	83.2	**93.6**	91.7	92.6	**98.1**	97.8	97.9
		w2v	64.9	64.7	64.8	77.4	**87.7**	82.2	90.0	93.8	91.8	97.8	99.6	**98.7**
MIMIC	497M	FT	37.7	10.6	16.5	78.9	21.7	34.0	86.0	90.0	87.8	97.9	97.7	97.8
		w2v	**71.9**	64.9	**68.2**	84.3	83.0	**83.6**	91.7	91.7	91.7	96.5	99.6	98.0
BTRIS	74.6M	FT	66.8	63.8	65.3	80.6	83.4	82.0	90.2	**95.8**	92.9	95.9	99.0	97.4
		w2v	69.7	63.7	66.7	**86.0**	79.2	82.4	88.2	93.8	90.9	96.7	**99.9**	98.3
PT-OT	4.2M	FT	68.8	62.5	65.5	84.5	80.2	82.3	92.0	**95.8**	**93.9**	97.1	97.7	97.4
		w2v	70.8	63.4	67.0	85.8	79.4	82.5	86.3	91.7	88.9	96.3	98.9	97.6

Table 6.2: Comparison of exact- and token-level NER results on BTRIS-Mobility using different embeddings: randomly-initialized embeddings as a baseline and unmodified word2vec (w2v) and FastText (FT) embeddings from different corpora. *Size* is the number of tokens in the training corpus.

Domain adaptation methods

We evaluate several different methods for adapting out-of-domain embeddings to the BTRIS corpus. These methods are summarized below; for full details, we refer to Newman-Griffis and Zirikly (2018).

Concatenation We concatenate out-of-domain and BTRIS/PT-OT embeddings as a baseline, allowing the model to learn a task-specific combination of the two representations.

Preinitialization Pre-trained representations provide a useful starting point for input features, but can be fine-tuned to correlations in a specific domain. We pre-initialize both word2vec and FastText toolkits with each of our reference embeddings and tune on task-relevant BTRIS data.

Linear transform To reduce the impact of limited vocabulary and minimal training data from small, task-specific corpora, we learn a linear transformation to map one (high-resource) set of embeddings to another (low-resource) set, using the Frobenius norm error minimization method of Artetxe et al. (2016).

Non-linear transform Finally, as our embedding domains do not necessarily have a linear relationship, we extend the method of Artetxe et al. (2016) to a non-linear transformation, using a feed-forward neural network to project one set of embeddings onto another.

6.1.4 Results

We report exact match results, calculated using CoNLL 2003 named entity recognition shared task evaluation scoring (Tjong Kim Sang and De Meulder, 2003), which requires that all tokens of an entity are correctly recognized. Additionally, given the

long span of Mobility and ScoreDefinition entities (see Section 6.1.2), we evaluated partial match performance using token-level results. For simplicity, we report only performance on the test set; however, validation set numbers consistently follow the same trends observed in test data. We denote embeddings trained using FastText with the subscript $_{FT}$, and word2vec with $_{w2v}$. Selected findings are presented here; for the full findings of our study, we refer to Newman-Griffis and Zirikly (2018).

Embedding corpora

Exact and token-level match results for both Mobility and ScoreDefinition entities are given for embeddings from each corpus in Table 6.2. In-domain BTRIS and PT-OT embeddings generally yield higher precision than out-of-domain embeddings, though this comes at the expense of recall. Most notably, despite a thousand-fold reduction in training corpus size, we see that PT-OT embeddings match the performance of PubMed embeddings on Mobility mentions and achieve the best overall performance on ScoreDefinition entities. Together with the overall superior performance of PT-OT embeddings even to the larger BTRIS corpus, our findings support the value of using input embeddings that are highly representative of the target domain. Nonetheless, MIMIC embeddings have both the best precision and overall performance on Mobility data, despite the domain mismatch of critical care versus therapeutic encounters.[28] This indicates that there is a limit to the benefits of in-domain data that can be outweighed by sufficient data from a different but related

[28]The poor performance of MIMIC$_{FT}$ embeddings persisted across multiple experiments with two embedding samples, manifesting primarily in making very few predictions (less than 30% as many Mobility entities other embeddings yielded).

Method	Exact match			Token match		
	Pr	Rec	F1	Pr	Rec	F1
WikiNews$_{FT}$	67.0	64.0	65.4	83.0	80.0	81.5
BTRIS$_{w2v}$	70.0	63.7	66.6	**86.0**	79.2	81.5
Concatenated	68.6	**66.7**	**67.6**	84.3	81.8	**83.0**
Preinitialized	66.8	64.5	65.6	78.4	**86.4**	82.2
Linear	**72.5**	58.9	65	79.1	83	81
1-layer ReLU	69.2	63.2	66.0	83.4	76.9	80.0
1-layer tanh	70.6	61.0	65.5	84.9	75.7	80.1
5-layer ReLU	67.3	61.9	64.5	83.5	76.6	79.9
5-layer tanh	67.9	62.1	64.9	82.1	77.0	79.4

Table 6.3: Comparison of domain adaptation methods for Mobility NER using a representative source/target pair: WikiNews$_{FT}$ as source and BTRIS$_{w2v}$ as target. Results are given for exact entity-level match and token-level match for test set Mobility entities.

domain. Token-level results follow the same trends as exact match; as many entity-level errors are only off by a few tokens, token-level scores are generally 15-20 absolute points higher than corresponding entity-level scores.

Mapping methods

Table 6.3 takes a single representative source/target pair and compares the different results obtained on recognizing Mobility entities when the NER model is initialized with embeddings learned using different domain adaptation methods. In this case, as with several other source/target pairs we evaluated, the concatenated embeddings give the best overall performance, stemming largely from an increase in recall over the baselines. However, we see that the nonlinear mapping methods tend to yield high precision: all settings improve over WikiNews embeddings alone, and the 1-layer tanh mapping beats the BTRIS embeddings as well. Reflecting the earlier observed trends of in-domain data, this is offset by a drop in recall, often of several absolute

Source	Target	Method	Pr	Rec	F1
WikiNews_{FT}	PT-OT_{w2v}	Preinit	72.1	66.1	**69.0**
WikiNews_{FT}	BTRIS_{w2v}	Linear	**72.5**	58.9	65
MIMIC_{w2v}	BTRIS_{FT}	Concat	67.4	**67.6**	67.5

Table 6.4: Best exact-match precision, recall, and F-1 for mobility information extraction. Results are for test set Mobility mentions, listed with the source/target pair and domain adaptation method used.

percentage points. As detailed in Newman-Griffis and Zirikly (2018), these differences broadly generalized across source/target pairs, with nonlinear transformations yielding the most consistent results, while concatenation achieved the best overall performance. Notably, domain adaptation experiments overall yielded neither consistent performance improvement nor degradation, though many results achieved notable improvement in precision or recall individually, suggesting that different methods may be useful for different downstream applications.

Source/target pairs

Table 6.4 highlights the source/target pairs that achieved the best exact match precision, recall, and F1 out of all the embeddings we evaluated, both unmapped and mapped. Though each source/target pair produced varying downstream results among the domain adaptation methods, a couple of broad trends emerged from our analysis. The largest performance gains over unmapped baselines were found when adapting high-resource WikiNews and PubMed embeddings to in-domain representations; however, these pairings also had the highest variability in results. The most consistent gains in precision came from using MIMIC embeddings as source, and these were mostly achieved through the nonlinear mapping approach.

Error analysis

Several interesting trends emerge in the NER errors produced in our experiments. Most generally, punctuation is often falsely considered to bound an entity. For example, the following string is part of a continuous Mobility entity:[29]

`supine in bed with elevated leg, and was left sitting in bed`

However, most trained models separated this at the comma into two Mobility entities. Unsurprisingly, given the length of Mobility entities, we find many cases where most of the correct entity is tagged by the model, but the first or last few words are left off, as in

`[he exhibits compensatory gait patterns]`$_{Pred}$ `as a result]`$_{Gold}$

This behavior is illustrated in the large performance difference between entity-level and token-level evaluation discussed in Section 6.1.4.

We also see that descriptions of physical activity without specific evaluative terminology are often missed by the model. For example, `working out in the yard` is a Mobility entity ignored by the vast majority of our experiments, as is `negotiate six steps to enter the apartment`.

Corpus effects

Within correctly predicted entities, we see some indications of source corpus effect in the results. Considering just the original, non-adapted embeddings as presented in Table 6.2, we note two main differences between models trained on out-of-domain vs in-domain embeddings. In-domain embeddings lead to much more conservative models: for example, PT-OT$_{w2v}$ only predicts 850 Mobility entities in test data, and

[29]Several examples in this section have been edited for deidentification purposes and brevity.

BTRIS$_{w2v}$ predicts 863; this is in contrast to 922 predictions from MIMIC$_{w2v}$ and 940 from PubMed$_{w2v}$. This carries through to mapped embeddings as well: adding PT-OT embeddings into the mix decreases the number of predictions across the board.

Several predictions exhibit some degree of domain sensitivity, as well. For example, "fatigue" is present at the end of several Mobility mentions, and both PubMed and MIMIC embeddings typically end these mentions early. PubMed embeddings also append more typical symptomatic language onto otherwise correct Mobility entities, such as `no areas of pressure noted on skin` and `numbness and tingling of arms`. MIMIC and the heterogeneous in-domain BTRIS corpus append similar language, including `and chronic pain`. WikiNews embeddings, by contrast, appear oversensitive to key words in many Mobility mentions, tagging false positives such as `my wife` (spouses are often referred to as a source of physical support) and `stairs are within range`.

Changes from domain adaptation

Domain-adapted embeddings fix some corpus-based issues, but re-introduce others. Out-of-domain corpora tend to chain together Mobility entities separated by only one or two words, as in

> `[He ambulates w/o ad]`$_{Mobility}$`, no walker observed,`
> `[antalgic gait pattern]`$_{Mobility}$

While source PubMed and WikiNews embeddings often collapse these to a single mention, adapting them to the target domain fixes many such cases. However, some of the original corpus noise remains: PT-OT$_{w2v}$ correctly ignored `and chronic pain` after a Mobility mention, but MIMIC$_{w2v}$ mapped to PT-OT$_{w2v}$ re-introduces this error.

The most consistent improvement obtained from domain adaptation was on Mobility entities that are short noun phrases, e.g. `gait instability`, and `unsteady gait`. Non-adapted embeddings typically miss such phrases, but mapped embeddings correctly find many of them, including some that in-domain embeddings miss.

Adaptation method effects

The most striking difference we observe when comparing different domain adaptation methods is that preinitialization universally leads to longer Mobility entity predictions, by both mean and variance of entity length. Though preinitialized embeddings still perform well overall, many predictions include several extra tokens before or after the true entity, as in the following example:

> `(now that her leg is healed [she is independent with wheelchair transfer]`$_{Gold}$ `and using her shower bench)`$_{Pred}$

Preinitialized embeddings also have a strong tendency to collapse sequential Mobility entities. Both of these trends are reflected in the lower token-level precision numbers in Table 6.3.

Comparing nonlinear mapping methods, we find that a 1-layer mapping with tanh activation consistently leads to fewer predicted Mobility entities than with ReLU (for example, 814 vs 859 with WikiNews$_{FT}$ mapped to BTRIS$_{w2v}$, 917 vs 968 with MIMIC$_{w2v}$ mapped to PT-OT$_{w2v}$). However, this difference disappears when a 5-layer mapping is used. Despite their consistent performance, nonlinear transformations seem to re-introduce a number of errors related to more general mobility terminology. For example, `he is very active and runs 15 miles per week` is correctly recognized by concatenated WikiNews$_{FT}$ and BTRIS$_{w2v}$, but missed by several of their nonlinear mappings.

6.1.5 Conclusions

We have shown that a state-of-the-art recurrent neural model is capable of capturing long, complex descriptions of mobility, and of recognizing mobility measurement scales nearly perfectly. Our experiments show that domain adaptation methods for the learned representations used as input features often improve recognition performance over both in- and out-of-domain baselines, though such improvements are difficult to achieve consistently. Most strikingly, we see that embeddings trained on a very small corpus of highly relevant documents nearly match the performance of embeddings trained on extremely large out-of-domain corpora, adding to the recent findings of Diaz et al. (2016).

Viewing this problem through an NER lens provides a robust framework for model design and evaluation, but is accompanied by challenges such as effectively evaluating recognition of long text spans and dealing with complex syntactic structure and punctuation within relevant mentions. The following sections describe reformulations of the extraction problem to mitigate these issues.

6.2 Token-level relevance scoring yields high recall for locating mobility reports

As application of NLP techniques has expanded into an increasing number of new information domains, including FSI, it is not always clear how best to identify information of interest, or to evaluate the output of automatic annotation tools. This can be especially challenging when target data in the form of long strings or narratives of complex structure, e.g., in financial data (Fisher et al., 2016) or clinical data (Rosenbloom et al., 2011).

We introduce HARE, a Highlighting Annotator for Ranking and Exploration. HARE includes two main components: a workflow for supervised training of automated token-wise relevancy taggers, and a web-based interface for visualizing and analyzing automated tagging output. It is intended to serve two main purposes: (1) triage of documents when analyzing new corpora for the presence of relevant information, and (2) interactive analysis, post-processing, and comparison of output from different annotation systems.

In this study, we apply HARE to mobility-related functional status information, and demonstrate that our approach mitigates the issues of sensitivity to long text spans and complex syntactic structure outlined in the previous section. Our model is able to produce a high-quality ranking of documents based on their relevance to mobility, and to capture mobility-likely document segments with high fidelity. We further demonstrate the use of post-processing and qualitative analytic components of our system to compare the impact of different feature sets and tune processing settings to improve relevance tagging quality.

6.2.1 Related work

Corpus annotation tools are plentiful in NLP research: brat (Stenetorp et al., 2012) and Knowtator (Ogren, 2006) being two heavily used examples among many. However, the primary purpose of these tools is to streamline *manual* annotation by experts, and to support review and revision of manual annotations. Some tools, including brat, support automated pre-annotation, but analysis of these annotations and corpus exploration is not commonly included. Other tools, such as SciKnowMine,[30] use automated techniques for triage, but for routing to experts for curation rather

[30] `https://www.isi.edu/projects/sciknowmine/overview`

	SpaCy	WordPiece
Num documents	400	
Avg tokens per doc	537	655
Avg mobility tokens per doc	97	112
Avg mobility segments per doc	9.2	

Table 6.5: Mobility information token-level dataset details, using SpaCy and WordPiece tokenization.

than ranking and model analysis. Document ranking and search engines such as Apache Lucene,[31] by contrast, can be overly fully-featured for early-stage analysis of new datasets, and do not directly offer tools for annotation and post-processing.

Previous efforts towards extracting mobility information have illustrated that it is often syntactically and semantically complex, and difficult to extract reliably (Newman-Griffis and Zirikly, 2018; Newman-Griffis et al., 2019b). Some characterization of mobility-related terms has been performed as part of larger work on functioning (Skube et al., 2018), but a lack of standardized terminologies limits the utility of vocabulary-driven clinical NLP tools such as CLAMP (Soysal et al., 2018) or cTAKES (Savova et al., 2010). Thus, it forms a useful test case for HARE.

6.2.2 System Description

Our system has three stages for analyzing document sets, illustrated in Figure 6.3. First, data annotated by experts for token relevance can be used to train relevance tagging models, and trained models can be applied to produce relevance scores on new documents (Section 6.2.2). Second, we provide configurable post-processing tools for cleaning and smoothing relevance scores (Section 6.2.2). Finally, our system includes

[31] `https://lucene.apache.org/`

interfaces for reviewing detailed relevance output, ranking documents by their relevance to the target criterion, and analyzing qualitative outcomes of relevance scoring output (Sections 6.2.2-6.2.2); all of these interfaces allow interactive re-configuration of post-processing settings and switching between output relevance scores from different models for comparison.

For our experiments on mobility information, we use an extended version of the dataset described by Thieu et al. (2017), which consists of 400 English-language Physical Therapy initial assessment and reassessment notes from the Rehabilitation Medicine Department of the NIH Clinical Center. These text documents have been annotated at the token level for descriptions and assessments of patient mobility status. Further information on this dataset is given in Table 6.5. We use ten-fold cross validation for our experiments, splitting into folds at the document level.

All hyperparameters and design choices discussed in this section were tuned on held-out development data in cross-validation experiments. We report the best settings here, and provide full comparison of hyperparameter settings, along with implementation details, in the online supplemental material to Newman-Griffis and Fosler-Lussier (2019a).[32]

Relevance tagging workflow

Preprocessing

Different domains exhibit different patterns in token and sentence structure that affect preprocessing. In clinical text, tokenization is not a consensus issue, and a variety of different tokenizers are used regularly (Savova et al., 2010; Soysal et al., 2018). As mobility information is relatively unexplored, we relied on general-purpose

[32] `https://arxiv.org/abs/1908.11302`

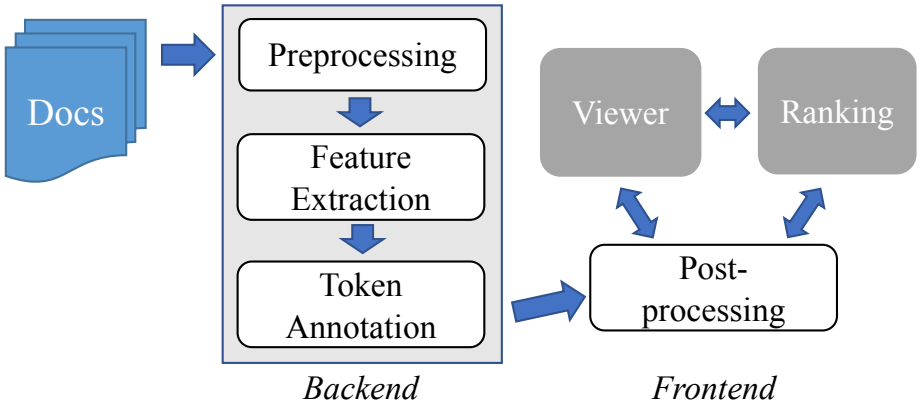


Figure 6.3: HARE workflow for working with a set of documents; outlined boxes indicate automated components, and gray boxes signify user interfaces.

tokenization with spaCy (Honnibal and Montani, 2017) as our default tokenizer, and WordPiece (Wu et al., 2016) for experiments using BERT. We did not apply sentence segmentation, as clinical toolkits often produced short segments that interrupted mobility information in our experiments.

Feature extraction

Our system supports feature extraction for individual tokens in input documents using both static and contextualized word embeddings.

Static embeddings Using static (i.e., non-contextualized) embeddings, we calculate input features for each token as the mean embedding of the token and 10 words on each side (truncated at sentence/line breaks). We used FastText (Bojanowski et al., 2017) embeddings trained on a 10-year collection of physical and occupational therapy records from the NIH Clinical Center.

ELMo (Peters et al., 2018) ELMo features are calculated for each token by taking the hidden states of the two bLSTM layers and the token layer, multiplying

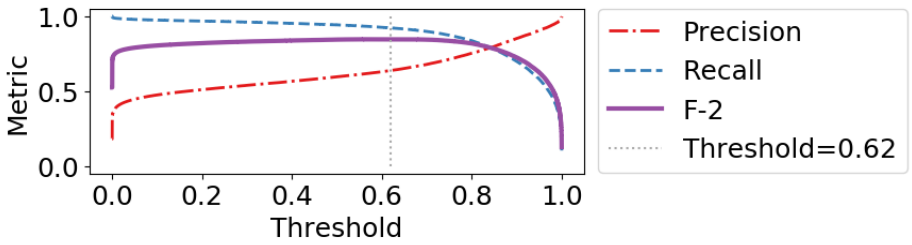


Figure 6.4: Precision, recall, and F-2 when varying HARE binarization threshold from 0 to 1, using ELMo embeddings. The threshold corresponding to the best F-2 is marked with a dotted vertical line.

each vector by learned weights, and summing to produce a final embedding. Combination weights are trained jointly with the token annotation model. We used a 1024-dimensional ELMo model pretrained on PubMed data[33] for our mobility experiments.

BERT (Devlin et al., 2019) For BERT features, we take the hidden states of the final k layers of the model; as with ELMo embeddings, these outputs are then multiplied by a learned weight vector, and the weighted layers are summed to create the final embedding vectors.[34] We used the 768-dimensional clinicalBERT (Alsentzer et al., 2019) model[35] in our experiments, extracting features from the last 3 layers.

Automated token-level annotation

We model the annotation process of assigning a relevance score for each token using a feed-forward deep neural network that takes embedding features as input and produces a binomial softmax distribution as output. For mobility information,

[33] `https://allennlp.org/elmo`

[34] Note that as BERT is constrained to use WordPiece tokenization, it may use slightly longer token sequences than the other methods.

[35] `https://github.com/EmilyAlsentzer/clinicalBERT`

we used a DNN with three 300-dimensional hidden layers, relu activation, and 60% dropout.

As shown in Table 6.5, our mobility dataset is considerably imbalanced between relevant and irrelevant tokens. To adjust for this balance, for each epoch of training, we used all of the relevant tokens in the training documents, and sampled irrelevant tokens at a 75% ratio to produce a more balanced training set; negative points were re-sampled at each epoch. As token predictions are conditionally independent of one another given the embedding features, we did not maintain any sequence in the samples drawn. Relevant samples were weighted at a ratio of 2:1 during training.

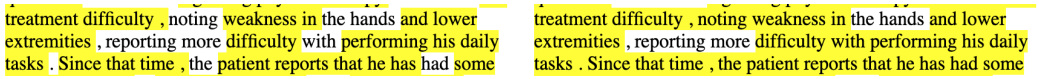
After each epoch, we evaluate the model on all tokens in a held-out 10% of the documents, and calculate F-2 score (preferring recall over precision) using 0.5 as the binarization threshold of model output. We use an early stopping threshold of 1e-05 on this F-2 score, with a patience of 5 epochs and a maximum of 50 epochs of training.

Post-processing methods

Given a set of token-level relevance annotations, HARE provides three post-processing techniques for analyzing and improving annotation results.

Decision thresholding The threshold for binarizing token relevance scores is configurable between 0 and 1, to support more or less conservative interpretation of model output; this is akin to exploring the precision/recall curve. Figure 6.4 shows precision, recall, and F-2 for different thresholding values from our mobility experiments, using scores from ELMo embeddings.

Collapsing adjacent segments We consider any contiguous sequence of tokens with scores at or above the binarization threshold to be a relevant *segment*. As shown in Figure 6.5, multiple segments may be interrupted by irrelevant tokens such as



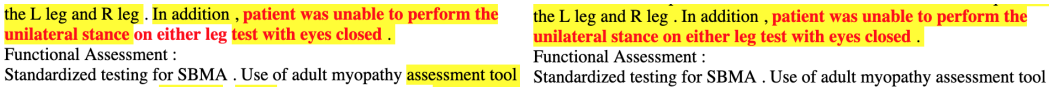
(a) No collapsing (b) Collapse one blank

Figure 6.5: Illustration of collapsing adjacent segments in HARE.

punctuation, or by noisy relevance scores falling below the binarization threshold. As multiple adjacent segments may inflate a document's overall relevance, our system includes a setting to collapse any adjacent segments that are separated by k or fewer tokens into a single segment.

Viterbi smoothing By modeling token-level decisions as conditionally independent of one another given the input features, we avoid assumptions of strict segment bounds, but introduce some noisy output, as shown in Figure 6.6. To reduce some of this noise, we include an optional smoothing component based on the Viterbi algorithm.

We model the "relevant"/"irrelevant" state sequence discriminatively, using annotation model outputs as state probabilities for each timestep, and calculate the binary transition probability matrix by counting transitions in the training data. We use these estimates to decode the most likely relevance state sequence R for a tokenized line T in an input document, along with the corresponding path probability



(a) Without smoothing (b) With smoothing

Figure 6.6: Illustration of Viterbi smoothing in HARE.

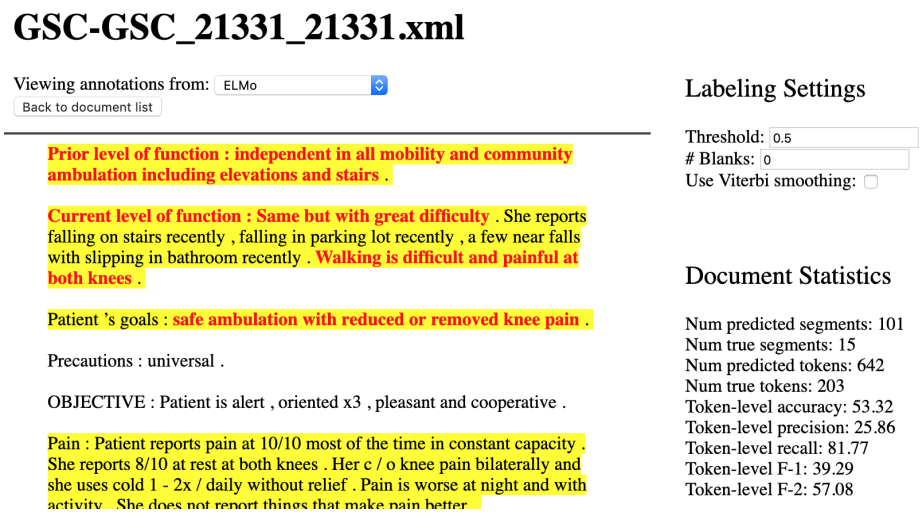


Figure 6.7: HARE annotation viewer interface.

matrix W, where $W_{j,i}$ denotes the likelihood of being in state j at time i given r_{i-1} and t_i. In order to produce continuous scores for each token, we then backtrace through R and assign score s_i to token t_i as the conditional probability that r_i is "relevant", given r_{i-1}. Let $Q_{j,i}$ be the likelihood of transitioning from state R_{i-1} to j, conditioned on T_i, as:

$$Q_{j,i} = \frac{W_{j,i}}{W_{R_{i-1},i-1}} \tag{6.1}$$

The final conditional probability s_i is calculated by normalizing over possible states at time i:

$$s_i = \frac{Q_{1,i}}{Q_{0,i} + Q_{1,i}} \tag{6.2}$$

These smoothed scores can then be binarized using the configurable decision threshold.

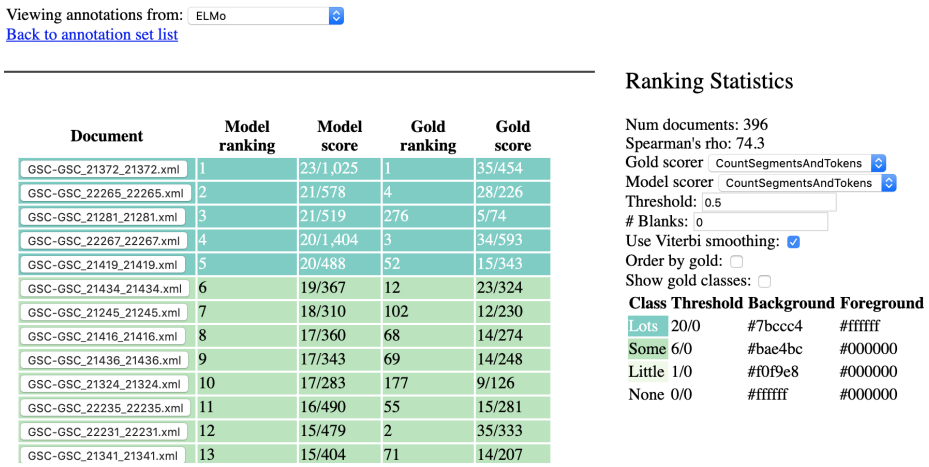


Figure 6.8: HARE document ranking interface.

Annotation viewer

Annotations on any individual document can be viewed using a web-based interface, shown in Figure 6.7. All tokens with scores at or above the decision threshold are highlighted in yellow, with each contiguous segment shown in a single highlight. Configuration settings for post-processing methods are provided, and update the displayed annotations when changed. On click, each token will display the score assigned to it by the annotation model after post-processing. If the document being viewed is labeled with gold annotations, these are shown in bold red text. Additionally, document-level summary statistics and evaluation measures, with current post-processing, are displayed next to the annotations.

Document set ranking

Ranking methods

Relevance scoring methods are highly task-dependent, and may reflect different priorities such as information density or diversity of information returned. In this system, we provide three general-purpose relevance scorers, each of which operates after any post-processing.

Segments+Tokens Documents are scored by multiplying their number of relevant segments by a large constant and adding the number of relevant tokens to break any ties by segment count. As relevant information may be sparse, no normalization by document length is used.

SumScores Documents are scored by summing the continuous relevance scores assigned to all of their tokens. As with the Segments+Tokens scorer, no adjustment is made for document length.

Density Document scores are the ratio of binarized relevant tokens to total number of tokens.

The same scorer can be used to rank gold annotations and model annotations, or different scorers can be chosen. Ranking quality is evaluated using Spearman's ρ, which ranges from -1 (exact opposite ranking) to +1 (same ranking), with 0 indicating no correlation between rankings. We use Segments+Tokens as default; a comparison of ranking methods can be found in Newman-Griffis and Fosler-Lussier (2019a).

Ranking interface

Our system also includes a web-based ranking interface, which displays the scores and corresponding ranking assigned to a set of annotated documents, as shown in Figure 6.8. For ease of visual distinction, we include colorization of rows based on

configurable score thresholds. Ranking methods used for model scores and gold annotations (when present) can be adjusted independently, and our post-processing methods (Section 6.2.2) can also be adjusted to affect ranking.

Qualitative analysis tools

We provide a set of three tools for performing qualitative analysis of annotation outcomes. The first measures lexicalization of each unique token in the dataset with respect to relevance score, by averaging the assigned relevance score (with or without smoothing) for each instance of each token. Tokens with a frequency below a configurable minimum threshold are excluded.

Our other tools analyze the aggregate relevance score patterns in an annotation set. For labeled data, as shown in Figure 6.4, we provide a visualization of precision, recall, and F-2 when varying the binarization threshold, including identifying the optimal threshold with respect to F-2. We also include a label-agnostic analysis of patterns in output relevance scores, illustrated in Figure 6.9, as a way to evaluate the confidence of the annotator. Both of these tools are provided at the level of an annotation set and individual documents.

6.2.3 Results on NIH mobility data

Table 6.6 shows the token-level annotation and document ranking results for our experiments on mobility information. Static and contextualized embedding models performed equivalently well on token-level annotations; BERT embeddings actually underperformed static embeddings and ELMo on both precision and recall. Interestingly, static embeddings yielded the best ranking performance of $\rho = 0.862$, compared

Embeddings	Smoothing	Annotation			Ranking
		Pr	Rec	F-2	ρ
Static	No	59.0	94.7	84.4	0.862
	Yes	60.5	93.7	84.3	0.899
ELMo	No	60.2	94.1	84.4	0.771
	Yes	66.5	91.4	84.8	0.886
BERT	No	55.3	93.8	82.2	0.689
	Yes	62.3	90.8	84.3	0.844

Table 6.6: HARE annotation and ranking evaluation on mobility documents, using three embedding sources. Results are given with and without Viterbi smoothing, using binarization threshold=0.5 and no collapsing of adjacent segments. Pr=precision, Rec=recall, ρ=Spearman's ρ Pr/Rec/F2 are macro-averaged over folds, ρ is over all test predictions.

to 0.771 with ELMo and 0.689 with BERT. Viterbi smoothing makes a minimal difference in token-level tagging, but increases ranking performance considerably, particularly for contextualized models. It also produces a qualitative improvement by trimming out extraneous tokens at the start of several segments, as reflected by the improvements in precision.

The distribution of token scores from each model (Figure 6.9) shows that all three embedding models yielded a roughly bimodal distribution, with most scores in the ranges $[0, 0.2]$ or $[0.7, 1.0]$.

6.2.4 Discussion

Though our system is designed to address different needs from other NLP annotation tools, components such as annotation viewing are also addressed in other established systems. Our implementation decouples backend analysis from the front-end interface; in future work, we plan to add support for integrating our annotation and ranking systems into existing platforms such as brat.

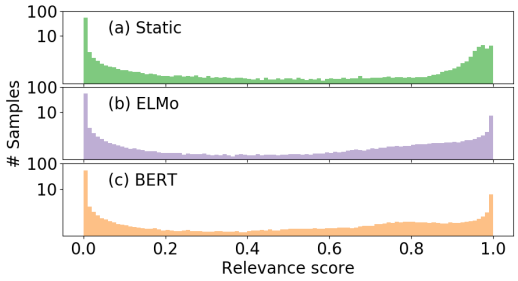


Figure 6.9: Distribution of token-level HARE relevance scores on mobility data: (a) word2vec, (b) ELMo, and (c) BERT.

In terms of document ranking methods, it may be preferred to rank documents jointly instead of independently, in order to account for challenges such as duplication of information (common in clinical data; Taggart et al. (2015)) or subtopics. However, these decisions are highly task-specific, and are an important focus for designing ranking utility within specific domains.

Extending to multi-class/multilabel applications

Our experiments focused on binary relevance with respect to mobility information. However, our system can be fairly straightfowardly extended to both multi-label (i.e., multiple relevance criteria) and multi-class (e.g., NER) settings.

For multi-label settings, such as looking for evidence of limitations in either mobility or interpersonal interactions, the only requirement is having data that are annotated for each relevance criterion. These can be the same data with multiple annotations, or different datasets; in either case, binary relevance annotators can be trained independently for each specific relevance criterion. Our post-processing components such as Viterbi smoothing can then be applied independently to each set of relevance

annotations as desired. The primary extension required would be to the visualization interface, to support display of multiple (potentially overlapping) annotations. Alternatively, our modular handling of relevance annotations could be redirected to another visualization interface with existing support for multiple annotations, such as brat.

Extending to multi-class settings would require fairly minimal updates to both the interface and our relevance annotation model. Our model is trained using two-class cross (relevant and irrelevant) cross-entropy; this could easily be extended to n-ary cross entropy for any desired number of classes, and trained with token-level data annotated with the appropriate classes. In terms of visualization and analysis, the two modifications required would be adding differentiating displays for the different classes annotated (e.g., different colors), and updating the displayed evaluation statistics to micro/macro evaluations over the multiple classes. Qualitative analysis features such as relevance score distribution and lexicalization are already dependent only on the scores assigned to the "relevant" class, and could be presented for each class independently.

Unexpected poor performance of BERT

Using BERT features in the HARE tagger yielded both lower precision and lower recall than using either ELMo or static embedding features, when using raw scores (see Table 6.6)—although Viterbi smoothing erases this gap, BERT features do not improve performance over static embeddings. This result is rather counterintuitive, given the weight of recent literature demonstrating that BERT features are consistently more discriminative than either static embeddings or preceding contextualized methods for text classification tasks, particularly for binary classification.

Two further sets of experiments are needed to explore possible confounding factors. First, while our experiments used averaging of the final three layers of BERT, recent work has demonstrated that semantic content at the lexical level is more directly encoded in lower layers of the network architecture (Tenney et al., 2019); thus, further experimentation using representations from different layers of the BERT network are needed to confirm that this result was not due to choice of layer alone. Second, one of the strengths of the BERT model is the option of fine-tuning the network for a downstream classification task; while our experiments use BERT features as fixed and train a classifier on top of them, BERT fine-tuning further trains the full network parameters in addition to a task-specific output layer. Thus, adapting the BERT fine-tuning approach for our token-level classification will shed further light on the degree to which the observed limitations are due to being unable to tune BERT parameters to the task of interest. If these additional experiments confirm our findings, this suggests that for the task of separating mobility-related tokens from non-mobility tokens, BERT embeddings are less separable than static features when using a feed-forward network, an outcome which would require further detailed analysis to explain.

Token-level modeling improves recall at the expense of precision

Our findings clearly support our initial hypothesis: that modeling the extraction of long, syntactically complex mobility reports as token-level classification improves recall. We observed consistently high recall with all embedding features we experimented with, and strong ranking performance demonstrated the utility of our model for prioritizing information-dense documents. Precision, however, was found to be consistently low, indicating clear room for improvement in de-noising the outputs of

our model. While Viterbi smoothing qualitatively reduces noise in the form of standalone "relevant" tokens, we see improved precision as a key aim for further research on refining our approach. As our understanding of the characteristics of mobility reports improves, a data-driven ensembling approach to combine the recall of token-level modeling with the higher precision of sequence labeling models is likely to be a fruitful direction of development.

6.2.5 Conclusions

We introduced HARE, a supervised system for highlighting relevant information and interactive exploration of model outcomes. We demonstrated its utility in experiments with clinical records annotated for mobility activity reports, and showed that it helps to address the issues of description length and complexity highlighted in Section 6.1. We also provided qualitative analytic tools for understanding the outcomes of different annotation models. In future work, we plan to extend these analytic tools to provide rationales for individual token-level decisions. Additionally, given the clear importance of contextual information in token-level annotations, the static transition probabilities used in our Viterbi smoothing technique are likely to degrade its effect on the output. Adding support for dynamic, contextualized estimations of transition probabilities will provide more fine-grained modeling of relevance, as well as more powerful options for post-processing. Our system is available online at `https://github.com/OSU-slatelab/HARE/`.

6.3 Applications to U.S. Social Security Administration data

The US Social Security Administration (SSA) is responsible for the management of the two largest federal disability programs in the United States, including the review

and adjudication of new applications for disability benefits. The concept of "disability" is operationalized for SSA's purposes in terms of ability to meet the demands of gainful employment (SSA 2008). Some of the most frequent functional limitations leading to reduced ability to meet these demands relate to mobility activities, such as walking, transferring body positions, and using transportation (Courtney-Long et al., 2015). A key part of the disability adjudication process is the review of medical documentation to find evidence to support reported limitations. With current rates of disability applications placing high demand for adjudication (Autor, 2011), it is important to develop automated tools to assist with evidence finding in the adjudication process.

In this study, we investigate the utility of the HARE token-level neural relevance tagger described in Section 6.2 to index mobility-related information in heterogeneous data associated with SSA disability applications. Mobility information is highly sparse in these documents, comprising on average less than 4% of document tokens (see Table 6.7). We evaluate the potential utility of our approach as an AI-assisted support tool for evidence review in disability adjudication, based on three specific use cases.

Use case 1 is document review, evaluated in terms of token-level relevance tagging.

Use case 2 is fine-grained ranking of clinical documents by their expected amount of mobility information, evaluated in terms of ranking correlation.

Use case 3 is coarse-grained document triage to identify a high-impact set of documents for further analysis, evaluated on ranking relevant documents over irrelevant ones.

	CEs	1,200	
		CE	HIT
# documents	304	449	693
Annot type	Spans	Document	Document
Avg tokens/doc	1,795.9	2,471.5	52,299.8
Avg rel seg/doc	8.6	–	–
Avg rel tok/doc	70.7	–	–
# rel docs	245	358	530
# irrel docs	59	91	163

Table 6.7: Two SSA datasets used for mobility information extraction study. Token count is given using SpaCy tokenization; for the span-level annotations, binary document relevance is defined as the presence of 1 or more relevant spans in the document. Span-level relevance statistics are not provided for documents in the 1,200-record corpus, as they are only annotated at the document level. 58 documents were removed from the 1,200-record dataset due to OCR noise.

We demonstrate that the HARE relevance tagging model yields strong performance on all three of these tasks, a first step in developing AI-based tools for reviewing functional status information. Qualitative review of system outputs shows complementary output patterns from static and contextualized embedding features, and identifies trends in output predictions and false negatives that suggest directions for further research.

6.3.1 Materials

We used two document collections for our study, both obtained from the US Social Security Administration through an Inter-Agency Agreement, and annotated by two domain experts; statistics of both document sets are provided in Table 6.7. The first consists of 304 consultative exams (CEs); these are special-purpose documents

recording a detailed evaluation of the individual who filed the claim for disability benefits by an expert provider contracted by SSA for the purpose (SSA 2014). Providers have typically not previously encountered the claimant, and these documents tend to be fairly long, but by and large consist of a set series of sections prescribed by SSA (SSA 2014). These documents were annotated for token-level span boundaries of mobility descriptions, following the protocol used by Thieu et al. (2017).

The second document collection includes 1,200 documents drawn from two types of SSA records: additional CEs (disjoint from the first collection); and Health IT (HIT) documents, sets of records provided directly to SSA from provider EHR systems via regional Health Information Exchanges (HIEs) during the process of developing a disability case. Both of these document types were annotated with a binary label indicating the presence or absence of a substantive evaluation of mobility status anywhere in the document.

Two practical characteristics of the latter dataset impacted the annotation and analysis processes. Many documents were submitted to SSA via fax or scan, stored in image format, and converted into text documents using optical character recognition (OCR). As a result, the digital texts of many of these documents suffered from greater or lesser degrees of OCR noise. Our annotation protocol therefore included a provision that if a document was unreadable, or the status of mobility-related information in it could not be determined due to OCR or other noise, those documents would be removed from the dataset. After this filtering, our final dataset included 888 relevant documents and 254 irrelevant documents; further details are provided in Table 6.7.

Additionally, HIT documents in some cases consisted of a conglomeration of records from multiple encounters. Thus, each HIT document may include several

individual records, sometimes spanning a significant time period. For the purposes of our annotations, we annotated the full document as relevant if any of the records it contained included mobility-relevant information. We did not include record segmentation or sectionization in our experiments, but highlight this as an important consideration for future work consuming HIE data where packaged records may not be automatically separated. This issue did noticeably increase both the size of our HIT documents (as shown in Table 6.7) and the sparsity of relevant information in them.

6.3.2 Methods

The linguistic characteristics of mobility descriptions are as yet poorly understood, and SSA data is unusually heterogeneous in both form and function, particularly compared to the homogeneous physical therapy notes used in the our previous study (Newman-Griffis and Zirikly, 2018). Our focus in this study is on exploring the characteristics of mobility-related information in a setting where it is used in decision-making, and testing whether a simple approach to estimating relevance for information retrieval is is an effective support tool for triage of document sets. Given both the novelty of mobility information and data privacy concerns pertinent to SSA data, we did not have access to well-developed baselines to compare against. However, our experimental goal is to evaluate AI-supported retrieval methods for what is currently a purely manual review-based process, thus our hypotheses are evaluated purely in terms of recovering the target information at an acceptable level for decision support.

Settings for HARE model

We experimented with two approaches to generate word embedding features for input to the HARE tagger: static and contextualized embeddings. For static embeddings, we utilized three 300-dimensional pretrained models: word2vec (Mikolov et al., 2013a) trained on Google News,[36] GloVe (Pennington et al., 2014) trained on 840 billion tokens of Common Crawl web text,[37] and FastText (Bojanowski et al., 2017) trained with subword information on combined Wikipedia and news data.[38] In addition, we trained our own FastText models on a separate corpus of approximately 70,000 medical evidence documents from SSA, using the skip-gram with negative sampling and CBOW training objectives; due to the much smaller corpus size, we trained 100-dimensional embeddings. Using static embeddings, we generated input features by averaging the embeddings for 10 tokens on either side of the target token (ending at linebreaks).

For contextualized embeddings, we used BERT (Devlin et al., 2019), a language model-based model structure using a Transformer network to generate context-sensitive embedding vectors for each token in a sequence. We experimented with three pretrained BERT models: BERT-Base,[39] trained on Wikipedia and book data; BioBERT (Lee et al., 2019), trained on PubMed abstracts;[40] and clinicalBERT (Alsentzer et al.,

[36] Available from `https://code.google.com/archive/p/word2vec/`

[37] `http://nlp.stanford.edu/data/glove.840B.300d.zip`

[38] `https://fasttext.cc/docs/en/english-vectors.html`

[39] `https://github.com/google-research/bert`

[40] `https://github.com/naver/biobert-pretrained`

2019), which is based on BioBERT but fine-tuned on clinical data.[41] All BERT models generate 768-dimensional vectors. We did not fine-tune the BERT models for our task, but rather generated embedding features using the fixed models and trained the HARE tagger on top of those features.

Training and hyperparameter settings

To train the HARE tagger, we used the token-level annotated data from the 304 CEs. These documents were tokenized by spaCy (Honnibal and Montani, 2017) (for static embeddings) or WordPiece (Wu et al., 2016) (required for BERT). We found that the document corpora we used did not lend themselves to clear definitions of sentence boundaries, and that the short segmentation often produced by clinical NLP toolkits (Griffis et al., 2016) frequently interrupted longer narratives; we therefore used linebreaks to separate text segments for embedding feature generation. We trained the model by subsampling a balanced set of relevant and irrelevant tokens from the full training set at each epoch, and training over this set of token samples using binary cross-entropy. After each epoch, we evaluated the model on a held-out 10% of the training data, and calculated F-2 score (which weights recall over precision), using a relevance score of 0.5 as the binarization threshold for discretizing model output. We used an early stopping threshold of 1e-05 on this development data, and trained with a patience of 5 epochs and a maximum training period of 50 epochs.

Our system hyperparameters, identified via F-2 score (a variant of F-1 weighted for recall) on held-out development data, were as follows: HARE DNN configuration of one 768-unit hidden layer with 10% dropout; training with all positive samples

[41] `https://github.com/EmilyAlsentzer/clinicalBERT`

and a random sampling of negative tokens at a 3:1 negative to positive ratio in each epoch; and equal class weights. For BERT features, we used BioBERT; for static embeddings, we utilized both out-of-domain word2vec GoogleNews (GoogleNews_{w2v}) and in-domain SSA_{SGNS} embeddings.

6.3.3 Experiments

We evaluated the utility of neural relevance tagging for three applications, representing different components of a clinical records review workflow. The most direct application (Experiment 1) is document review in order to locate evidence. Experiments 2 and 3 investigate information retrieval applications: Experiment 2 evaluates detailed ranking of different levels of mobility information, and Experiment 3 evaluates a purely triage application of ranking documents with any mobility information over those without any. These sets of experiments are discussed in detail in the following sections.

Experiment 1: token-level relevance tagging

Our first set of experiments was designed to evaluate the accuracy of our relevance tagger at the token level, as a strict measure of our ability to exactly recover the location of mobility-relevant information in SSA documents. Five-fold cross validation was used on the 304 CE corpus for these experiments; held-out development data for halting model training was randomly subsampled from the training set of each fold. At test time, all tokens of each test document were passed as input to the model, and the output relevance probability recorded for each. Evaluation was conducted by binarizing the relevance probabilities at 0.5, and calculating precision, recall, F-1, and F-2 over the full set of test tokens.

Features	P	R	F1	F2
GoogleNews_{w2v}	48.9	82.2	61.2	72.3
SSA_{SGNS}	47.9	82.2	60.5	71.8
BioBERT	46.9	76.9	58.2	68.0

Table 6.8: Token-level HARE relevance tagging results on SSA 304 CE corpus from 5-fold cross validation, using each embedding model. Statistics are averaged across folds, using a relevance score binarization threshold of 0.5. P=Precision, R=Recall.

Table 6.8 shows the results from our three embedding methods. Relevance tagging at the token level achieves high recall in all three cases, although only roughly one in two tokens tagged by the model are "true" relevant tokens. Interestingly, static embeddings outperform contextualized BERT embeddings on both precision (1-2%) and recall (5.3%), suggesting that either the contextualized features are overparameterized for this size of dataset, or that using static embeddings enables leveraging lexicalized triggers in a way that the BERT model has not been tuned to do. For static embeddings, the higher-data GoogleNews_{w2v} embeddings slightly outperform in-domain SSA_{SGNS}features (0.5% F-2). From an application perspective, however, all three embedding methods are effectively equivalent, providing high-recall indexing with a signal to noise ratio of about 1:1.

Experiment 2: document ranking

Our token-level tagging experiments measured our system's ability to strictly recover the information of interest. As highlighted in the Introduction, another application of use to SSA in processing large collections of medical evidence is priority ranking documents by the amount of mobility-related information they are likely to contain. We therefore conducted document ranking experiments, again using the 304

token-level annotated CEs. A gold standard ranking was calculated by counting the number of relevant segments (i.e., contiguous sequences of relevant tokens) in the gold annotations for each document; in the case of a tie in number of segments, the document with the greater overall number of relevant tokens was assigned the higher ranking. The same ranking procedure was applied to binarized token-level relevance predictions to produce a model ranking.

Evaluation of our relevance tagging system was conducted using five-fold cross validation to obtain relevance scores for every token in the 304 CEs dataset in a test scenario, as in our first set of experiments. However, measuring rankings of five different 60-document sets is less informative for a high-volume scenario than ranking all 304 documents; we therefore combined the test set predictions from all five folds and ranked the full document set based on these. In our view, the practical evaluation at a larger scale outweighed potential cross-contamination effects of using test set outputs trained on overlapping training sets; nonetheless, this evaluation is necessarily somewhat optimistic. Ranking performance was measured using Spearman's rank correlation coefficient ρ, which ranges from -1 (indicating perfect anti-correlation) to +1 (indicating perfect correlation), where 0 indicates no correlation (i.e., random re-ranking). We included the Viterbi smoothing technique provided in HARE in our experiments, to reduce noise in output relevance scores.

Results

As shown in Table 6.9, raw token-level relevance scores rank the 304 documents with very strong correlation to the gold ranking ($\rho = .819$ in the worst case). Viterbi smoothing increases ranking quality considerably, to $\rho = .892$ in the best case, without noticeably degrading the token-level annotation quality. The effect of smoothing on

Features	Raw		Smoothed	
	ρ	F-2	ρ	F-2
GoogleNews_{w2v}	.832	72.3	.887	72.1
SSA_{SGNS}	.826	71.9	.892	71.4
BioBERT	.819	68.1	.873	69.3

Table 6.9: Annotation and ranking results for HARE experiments on SSA CEs, reporting Spearman's ρ and token-level F-2. Results evaluated on combined test set predictions from all folds of cross validation. F-2 is micro-averaged, slightly increasing over the macro-averaged F-2 in Table 6.8. Raw uses token-level relevance scores without post-processing; Smoothed includes Viterbi smoothing.

ranking correlation is about the same for all embedding models; however, its effect on token-level annotations is noticeably stronger when using BERT embeddings, where precision is increased by nearly 10% (to 56.2%; compared to a 3% gain for each of the static models), with a 4% drop in recall (3% for static models). Overall, all three models yield extremely strong correlation between model ranking and gold ranking when smoothing is applied, indicating that while token-level annotation may be noisy, it nonetheless captures the relevant information for successful retrieval.

Experiment 3: binary document ranking

At sufficient scale, determining whether a document merits further detailed analysis is an important first step in document triage and prioritization. Additionally, minor re-rankings of documents with similar amounts of mobility information may affect Spearman's ρ while having minor practical impact on system utility. We therefore conducted a third set of experiments evaluating document ranking based on a binary assessment of whether they were likely to have any mobility information in them or not. In this scenario, documents were ranked using the same procedure as

Features	CE	HIT	All
GoogleNews$_{w2v}$	99.1	97.6	97.1
SSA$_{SGNS}$	98.8	98.1	97.9

Table 6.10: Results from binary relevance ranking experiments on SSA data, reporting average precision over documents in our 1,200-document corpus, evaluated on binary document-level relevance annotations. Results are also broken out for CEs (449 documents) and HIT (693). BERT features were not used due to document length. Note that as CE and HIT documents are interleaved in the All setting, overall results can be lower than on individual subsets.

in Experiment 2, and this ranking was compared to the gold document-level binary labels to report average precision (AP). AP measures, for each relevant document, the proportion of the documents ranked higher which are truly relevant, and averages these ratios to report overall ranking quality.

For these experiments, we trained our relevance tagger using the full set of all 304 token-annotated CEs, and generated relevance scores for all tokens in each of the 1,200 binary-annotated documents. Due to the length of the HIT documents in this collection (an average of 52,000 tokens in each document), feature generation using BERT proved logistically infeasible: feature extraction on a subset of 150 documents took several days and produced hundreds of gigabytes of output. We therefore constrained our experiments to static embedding features only; results from our first two sets of experiments suggest that BERT would achieve comparable performance absent logistical difficulties.

Results

Table 6.10 shows the average precision achieved for the 1,200 document dataset, overall and by document type. Both sets of static embeddings overwhelmingly rank

relevant documents higher than irrelevant documents, achieving 97.1% overall AP in the worse case. As can be expected from the considerably larger size of HIT documents, they are slightly more difficult to rank correctly than CE documents are, though both feature sets yield above 97.5% AP. Thus, token-level relevance tagging is clearly effective for prioritizing relevant documents in a triage setting.

6.3.4 Qualitative analyses

Our quantitative system evaluations measured the utility of neural relevance tagging for different application scenarios. We also conducted qualitative analysis of system outputs to gain an understanding of what kinds of data are being tagged as relevant (correctly or erroneously) by different systems, and what implications these trends have for practical evidence retrieval of mobility-related information. We investigated three primary questions:

1. What differences do we observe in system outputs when using static vs contextualized word embedding features?

2. What patterns of error do we observe for false negatives, i.e. true mobility descriptions missed by our relevance tagger? (This analysis is constrained to the 304 CEs, as it requires token-level gold relevance annotations).

3. What patterns do we observe in text segments tagged as relevant? This includes both true and false positives in the token-level 304 CEs dataset, but also review of relevance annotations produced for the 1,200 document dataset.

Static vs contextualized features

While BERT and static embeddings yield comparable results in our experimental evaluations, they exhibit distinct patterns in the relevance annotations they produce. As shown in Table 6.11, BERT features lead to a striking increase in number of relevant segments tagged compared to static embedding features, and a concomitant reduction in the length of each segment. Many of these short segments are in fact close to one another, and are often parts of a longer segment tagged by static features; for example, BERT highlights the underlined phrases in the true segment "her husband estimated that the maximum weight she could lift would be equivalent to a gal ##lon of milk"; static features tag the entire segment contiguously. Many BERT segments are one or two-word phrases that appear somewhat random: for example, the underlined phrases in "her hair was brown and neck length". Interestingly, changing the binarization threshold does not noticeably decrease this noise without removing a considerable degree of useful signal as well. However, as illustrated in Table 6.11, Viterbi smoothing does close some of the gaps between segments and remove noisy segments, considerably decreasing the number of segments and increasing mean segment length. False positives remaining after smoothing are typically reasonable, if not necessarily directly relevant to mobility: for example, "her posture was within normal limits", and "she did not use correct ##ive lenses".

Static embeddings produce many fewer short segments, though some individual words and phrases are still tagged: e.g., "The claimant reported he has a problem with agitation". Static output segments, by contrast, often start before a true relevant segment and extend after it, suggesting lexicalization effects within the 10-token context window; see "enabled him to *take a job as a school bus*

Features	# Segments		Tokens/Segment	
	Mean	Max	Mean	Max
GoogleNews_{w2v}	10.7	57	11.2	114
+Smoothing	6.7	42	16.3	130
SSA_{SGNS}	10.9	64	11.2	99
+Smoothing	6.4	39	17.3	103
BioBERT	46.5	319	2.8	71
+Smoothing	15.8	87	6.6	93

Table 6.11: Statistics of HARE outputs on SSA 304 CE corpus, including number of segments per document and number of tokens per segment, using relevance scores. Results are given using raw scores, binarized at 0.5, and with Viterbi smoothing. GoogleNews_{w2v} and SSA_{SGNS} use SpaCy tokenization, while BioBERT uses WordPiece.

driver" (italics indicate the true relevant segment). Some static segments are also offset from true segments, e.g. "*will get up once during the night* to use the bathroom". This produces an error in token-level evaluation, but is still helpful from a retrieval standpoint.

False negatives

In terms of false negative segments (i.e., true relevant segments in which none of the tokens were tagged as relevant), the noisiness of BERT output proves useful: virtually no relevant segments in the 304 CEs were entirely missed when using BERT features. Static embeddings, while qualitatively appealing in producing long, contiguous relevant segments, are also more susceptible to false negatives. The main trend we observed in these cases was syntactic: examples with mobility-relevant action verbs, such as "walking" or "transfer" are retrieved reliably, but many segments with mobility-relevant nouns were missed when using static features. For example, "right

`and left lateral bending approximately 10 degrees`" was missed by static features, while BERT tagged "and" and "bending approximately." Some examples combining long-distance dependencies with less direct assertions, such as "`claimant has symptomatic limitations in his ability to squat`", were also missed by both sets of static features.

Relevance prediction patterns

Apart from the distinctions between outputs from static and BERT features, we observed several general patterns in relevance tagging outputs. The first is lexicalization: action verbs such as "stoop," "crouch," "climb," and "balance" (along with morphological variants) were tagged as relevant more than 90% of the time by all three embedding models, as were mobility-relevant objects such as "ladders," "ramps," and "stairs." These lexicalizations mostly reflected the true mobility annotations, though some false positives leaked through: for example, lexicalizing "table" from uses as a physical location led to 64 HIT documents being tagged with a single relevant segment, "Table of Contents." More practically, lexicalization effects yielded some acceptable false positives, such as "`she left because she had surgery and had to stand`", as well as more neutral statements such as "`call bell and possessions in reach`" (where "reach" is a common verb in mobility annotations).

The more challenging problem to handle in terms of false positives is a pragmatic one. Some phrases tagged as relevant describe neutral positional information: e.g., "`Patient in chair`", with no information on how the patient reached that position. More significantly, many tagged references to mobility status were hypothetical, describing goals for the patient's therapy or conjectures; several others were field headers in clinical templates, referring to limitations or actions that may or may not have been

Features	# Seg.	Tok./Seg.	% Tok.
CE (2,471.5 avg tokens)			
GoogleNews_{w2v}	14.7	8.9	5.3
SSA_{SGNS}	14.9	9.0	5.4
HIT (52,299.8 avg tokens)			
GoogleNews_{w2v}	140.3	6.5	1.7
SSA_{SGNS}	118.6	6.9	1.6

Table 6.12: Statistics of HARE outputs on SSA 1,200-document corpus, including mean number of segments (Seg.) per document and number of tokens (Tok.) per segment, with the mean proportion of the tokens in each document tagged as relevant, broken down into CE and HIT subsets. Results are given using raw scores, binarized at 0.5, without Viterbi smoothing.

observed. Local context is insufficient to capture these pragmatic implications in the absence of prior knowledge about template format or current section; finding systematic ways of incorporating this knowledge into relevance estimation systems represents a significant area for further work on improving AI-assisted tools for evidence review.

Finally, as our document sets varied considerably in length, we investigated how relevance annotation scales to larger documents. Table 6.12 shows statistics for relevance annotations of the 1,200 document dataset. While the number of segments annotated as relevant increases on average for the longer HIT documents, the fraction of document text annotated as relevant decreases considerably, demonstrating that the relevance tagging model successfully ignores much of the irrelevant data introduced in the longer documents.

6.3.5 Discussion and limitations

Our experiments, while yielding compelling results, are simulations of real document review workflows. In order to evaluate potential utility for operationalizing

AI-assisted tools such as relevance tagging within real-world disability adjudication at SSA, two clear next steps are needed. First, this work can be extended to other types of functional status information, either by developing multiple expert relevance taggers (e.g., one for mobility, another for domestic life activities, etc) and combining their output, or developing multi-stage models to gradually zero in on specific information of interest for an individual's claim. Second, a usability study and/or randomized controlled study should be conducted to evaluate whether disability adjudicators find integrating AI-assisted tools into the adjudication process helpful, and whether this integration results in meaningful process improvements.

Outside of the SSA setting, the potential utility of research on retrieving functional status information like mobility is limited by a lack of appropriate data. The SSA records used in this study were heterogeneous in length, contents, source provider, and document type, but are subject to stringent data privacy protections. Similar protections apply to US Department of Veteran's Affairs data, another data source supporting valuable research in clinical outcomes (Shao et al., 2016). More accessible data sources, such as MIMIC (Johnson et al., 2016), are either from a single institution (and often a single specialty) or lack data relevant to functional status. Efforts to develop more diverse and accessible data sets about functional status will significantly contribute to research such as ours by facilitating easier comparison of systems and enabling a broader body of researchers to be involved.

A few limitations in our experimental evaluations should also be discussed. While our study was intended as a proof of concept for supporting evidence review with NLP, and data privacy and a lack of appropriate models made identification and use of relevant baseline methods difficult, it is quite possible that other methods

would yield superior results for any of our three experiments. We therefore cannot make claims beyond *potential* utility, but by using a published method with publicly-available embedding features, our results can at least serve as a baseline for further research on evaluation and system comparison in similar settings.

Finally, two specific characteristics of the SSA documents we used affected our experimental results. Our filtering process of the OCRed portions of our documents only removed documents that were unreadable, leaving many documents with small amounts of remaining OCR noise; in all cases, line breaks were also introduced by the OCR process. OCR errors impacted model performance slightly, leading to "relevant" tags such as "`1-6-4;e2aw`" that were included in relevance evaluation. Additionally, we did not distinguish in our document set between documents associated with different disability claimants; in a practical setting, document ranking would be applied to triage the records for a specific individual, which might affect the ranking quality.

6.4 Conclusions

Well-chosen learned representations exhibit significant utility as input features for neural models of functional status information extraction. We have demonstrated that domain-representative embedding features improve the performance of off-the-shelf models for information extraction, and that tailored models focusing on word-level representations help to address the challenges of long, complex strings and variable vocabulary in FSI. Our experiments on Social Security Administration data demonstrate significant potential for deploying AI-assisted support tools in a benefits adjudication setting focused on functional status. In the next chapter, we present work addressing

the next step in the information extraction pipeline: normalization of identified text spans to domain-relevant concepts.

Chapter 7: Using Concept Representations for Semantic Grounding

After locating information for extraction, the next step in document processing is to identify the *kinds* of information extracted. Within a restricted domain, this typically translates into *concept normalization*, i.e., assigning a standardized unique identifier to each mention of a given concept, in order to connect information agnostic of surface form. This chapter presents a method for concept normalization utilizing learned representations of domain concepts, including our method described in Chapter 5. Section 7.1 introduces the normalization model and provides proof-of-concept experiments on word sense disambiguation. Section 7.2 describes our application of this model to clinical concept normalization in Track 3 of the the 2019 n2c2 Challenges, and Section 7.3 presents results on identifying the type of functional activity described in mobility activity reports.[42]

7.1 PROSE: Word sense disambiguation with projected sense representations

Identifying the correct sense of words in context is a fundamental and long-standing challenge in natural language processing (NLP). Recent technologies for

[42]Portions of Section 7.1 have previously been submitted for publication and are currently in revision.

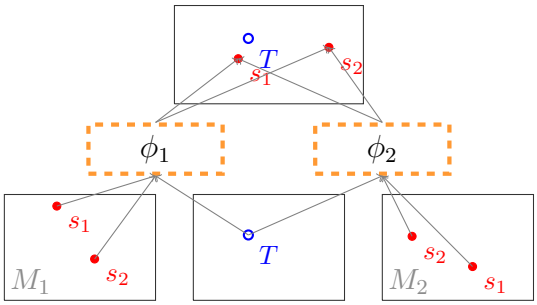


Figure 7.1: Illustration of PROSE intuition: using projectors ϕ_1 and ϕ_2 to map sense embeddings from two spaces (M_1 and M_2) into the same space as context T for disambiguation.

generating context-sensitive vector representations for words have shown significant promise for word sense disambiguation (WSD) (Peters et al., 2018). However, recent state-of-the-art results in WSD have been obtained by task-specific models that do not utilize contextualized embedding features (Uslu et al., 2018; Kumar et al., 2019). This begs the question of what dedicated WSD models contribute over and above what contextualized embeddings encode, and whether these approaches capture contradicatory or complementary information.

In this study, we present PROSE, a straightforward model based on Projecting Sense Embeddings, that effectively combines recent advances in sense representation (Pakhomov et al., 2016; Camacho-Collados et al., 2016; Newman-Griffis et al., 2018) with contextualized embedding features. We demonstrate that PROSE improves results over contextualized embeddings alone on both benchmark WSD datasets and a low-resource biomedical application, and we achieve benchmark WSD results comparable to dedicated state-of-the-art models. Additionally, our approach builds on recent success in zero-shot WSD (Kumar et al., 2019) by combining sense embeddings

from multiple sources (as illustrated in Figure 7.1) to capture complementary information; by using general-purpose sense embeddings instead of a task-specific model, we are able to achieve similar gains on zero-shot senses with a more parsimonious approach. Our results show that PROSE successfully leverages the diversity of different sense embedding approaches to improve disambiguation, and that its main limiting factors are a lack of embeddings for some correct senses and a preference for generic senses.

7.1.1 Related Work

Contextualized word embedding models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have proven useful across many NLP tasks. However, while word embedding features have been used extensively in WSD (Pedersen, 2010; Iacobacci et al., 2016), contextualized embeddings have not yet been systematically incorporated. Many significant WSD advances have leveraged a variety of expert knowledge sources in addition to lexical features, such as dictionary definitions (Lesk, 1986) and WordNet features (Navigli et al., 2011); McInnes and Pedersen (2013) summarize relevant approaches in the biomedical domain.

Sense embeddings offer an approach to capture some of this expert knowledge while supporting dense vector space operations. Embeddings of senses and knowledge base entities have been derived from graph structure (Grover and Leskovec, 2016), dictionary definitions (Pakhomov et al., 2016), lexical statistics (Camacho-Collados et al., 2016), and task-specific encoding models (Kumar et al., 2019), among others. Different embedding methods have been shown to capture complementary information for WSD (Newman-Griffis et al., 2018). Our approach was designed to combine the

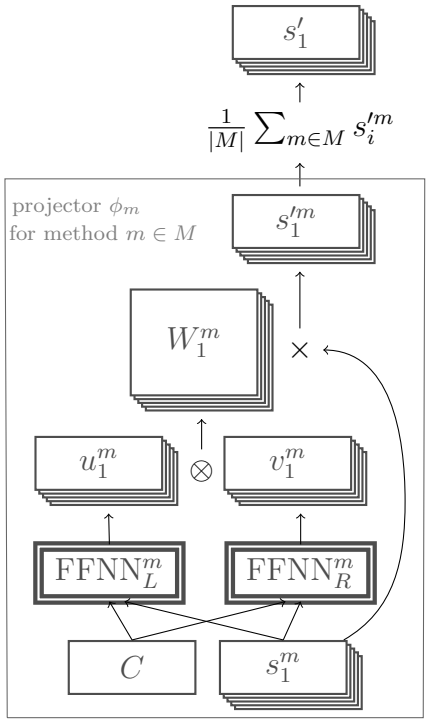


Figure 7.2: Diagram of MatrixMult PROSE projector; one projector ϕ_m is trained for each embedding method $m \in M$ used, and projector outputs are averaged for final sense embeddings.

strengths of different sense embedding sources in order to maximally leverage the information encoded in contextualized embedding features.

7.1.2 Disambiguation Model

Given an ambiguous word w in some context (either a fixed window of words on either side or a complete sentence), let C denote the vector embedding of the context, and $S = s_1 \dots s_n$ be the embeddings of the n candidate senses for w. In this work, we assume that senses are given *a priori*, using a knowledge source such as WordNet.

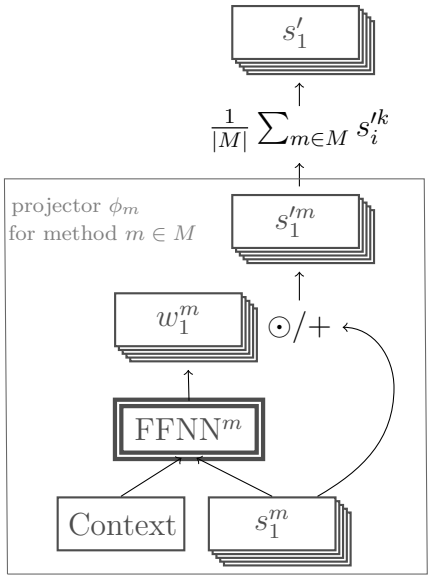


Figure 7.3: Diagram of Re-weighting (using $\odot$) and Residual (using $+$) PROSE projector configurations.

Projection model

We define a projection function ϕ that takes a t-dimensional context vector C and d-dimensional sense vector s and produces a projected t-dimensional sense vector $\hat{s}$:

$$\phi : \mathbb{R}^t \times \mathbb{R}^d \longrightarrow \mathbb{R}^t \tag{7.1}$$

The same ϕ is applied to each sense representation to create $\hat{s_1} \ldots \hat{s_n}$. We consider four configuration of ϕ, detailed in the following paragraphs.

MatrixMult We define two feed-forward neural networks (FFNN) to produce vectors $u \in \mathcal{R}^d$ and $v \in \mathcal{R}^t$:[43]

$$u = \text{FFNN}_L(C, s) \tag{7.2}$$

$$v = \text{FFNN}_R(C, s) \tag{7.3}$$

The outer product of these two vectors forms $d \times t$ matrix W, which is used to transform s:

$$W = u \otimes v \tag{7.4}$$

$$\hat{s} = Ws \tag{7.5}$$

This approach, shown in Figure 7.2, captures cross-correlation between features in sense and context embeddings, and supports differing vector dimensionalities.

Re-weighting Our second configuration (Figure 7.3) uses one FFNN to produce a re-weighting vector w for s:

$$w = \text{FFNN}(C, s) \tag{7.6}$$

$$\hat{s} = s * w \tag{7.7}$$

This transforms each feature of the sense embeddings independently, in order to focus on context-specific linear subspaces. While the MatrixMult approach can support any dimensionalities d and t, this method requires that $t = d$.

Residual Our third configuration calculates a residual vector that is added to s, instead of multiplying it. This allows adjustment of sense embedding features relative

[43]We experimented with a single FFNN to produce the concatenation of u and v, but it consistently underperformed the two-FFNN approach. Learning to project the W matrix directly required too many parameters to be practical.

to the origin, directly affecting cosine similarity.

$$\hat{s} = s + \text{FFNN}(C, s) \tag{7.8}$$

Direct Our final configuration only uses the sense embeddings as input, and directly predicts a new embedding vector for each candidate sense.

$$\hat{s} = \text{FFNN}(C, s) \tag{7.9}$$

In all configurations, the FFNN is of arbitrary depth, hidden state dimensions, and activation function.

Combining multiple representations

For multiple representation spaces $M_1 \dots M_k$, we define separate projectors $\phi_1 \dots \phi_k$ for each; with the MatrixMult and Direct formulations, this allows us to combine representations from spaces with different dimensionalities. The final projections are then calculated as an average of the outputs of individual projectors:

$$\hat{s_i} = \frac{1}{k}\left[\phi_1(C, s_i^1) + \dots + \phi_k(C, s_i^k)\right] \tag{7.10}$$

Scoring model

Using projected sense representations $\hat{s_1} \dots \hat{s_n}$, we calculate a normalized scoring distribution for the candidate senses using the vector similarity-based linear model of Sabbir et al. (2017):

$$\psi(S, T) = \underset{s_i \in S}{\text{softmax}}\left[\cos(C, s_i) * \frac{||C_{|s_i}||}{||s_i||}\right] \tag{7.11}$$

where $C_{|s_i}$ denotes the vector projection of C onto s_i, and $||s_i||$ denotes the vector norm of s_i. Note that this model requires that the context and sense vectors be of the same dimensionality.

Loss function

Given labeled triples as $\langle T, S, y \rangle$, where y is a one-hot vector indicating the correct sense from the candidates, we train our model with cross entropy:

$$\mathcal{L}(T, S, y) = -\sum_{i=1}^{n} y_i \log \psi_i \tag{7.12}$$

Under this formulation, the "label" set for cross entropy training changes from sample to sample (as it is the set of candidate senses for a given word); however, the only learned components of our model are the neural networks in ϕ, which operate on single embedding vectors and are weight-tied over the set of candidates.

7.1.3 Data

Our model requires three types of input data: sense-annotated datasets, embedded context representations, and embedded sense representations. Each are detailed in the following subsections.

Datasets

We use the WSD evaluation framework of Raganato et al. (2017a), which uses five benchmark WSD datasets for evaluation, and provides a separate large-scale corpus for model training; all corpora are annotated with WordNet 3.0 sense keys. Statistics for all datasets are provided in Table 7.1.

SemCor (Miller et al., 1994) is a large subset of the Brown corpus, annotated with WordNet senses; this dataset is used for model training only.

SensEval-2 (Edmonds and Cotton, 2001) contains sense annotations from the British National Corpus and the Wall Street Journal (WSJ) sections of the Penn Treebank.

Dataset	# Types	# Senses	# Samples
SemCor *(Train)*	22,436	33,362	226,036
SemEval-07 *(Dev)*	330	375	455
SensEval-2	1,093	1,335	2,282
SensEval-3	352	1,167	1,850
SemEval-13	751	827	1,644
SemEval-15	512	659	1,022
Test *(no SemEval-07)*	2,654	3,447	6,798

Table 7.1: Sense-annotated datasets for WSD datasets; Types denotes the number of unique lemma/POS combinations.

SensEval-3 (Snyder and Palmer, 2004) contains annotations from WSJ and the Brown corpus.

SemEval-2007 Task 17 (Pradhan et al., 2007) also contains WSJ and Brown corpus annotations; as the smallest evaluation dataset, we follow Raganato et al. (2017b) in holding this dataset out for development and using the other four sets for test.

SemEval-2013 Task 12 (Navigli et al., 2013) uses the English portions of thirteen sense-annotated multilingual news articles.

SemEval-2015 Task 13 (Moro and Navigli, 2015) contains multilingual sense-annotated data from diverse domains; we restrict ourselves to the English portions.

Context embeddings

Following Peters et al. (2018), we calculate representations of ambiguous contexts using the second bLSTM layer of a pre-trained ELMo model.[44] All datasets in the

[44]We used the 1B Word Benchmark model of Peters et al. (2018), available at `https://allennlp.org/elmo`.

WSD evaluation framework are pre-sentence chunked and tokenized; for a given ambiguous word w at position i in sentence $\mathcal{S}$, we pass $\mathcal{S}$ to ELMo and use the hidden state of the model at index i as the context representation.

Sense representations

We use four approaches to represent WordNet senses, each using a different knowledge source.

SemCor embeddings follow the method of Peters et al. (2018); we embed sentences in SemCor with ELMo, and retrieve the hidden state of the second bLSTM layer at the index of each sense annotation. For each unique word sense, we average the bLSTM states over every occurrence of that sense throughout the corpus.

Definition embeddings are created using the human-written definitions, or glosses, for senses provided in WordNet. For each unique synset, we retrieve its definition and pass it through ELMo, and then average the hidden states of the second bLSTM layer across all words in the definitions. For compatibility with dataset annotations, these are then mapped to the synset's default sense key.[45]

WordNet embeddings are derived from the graph structure of WordNet. We define the graph $G = (V, E)$, where V is the set of synsets in WordNet, and edge $(s_1, s_2) \in E$ if synsets s_1 and s_2 are connected by hypernymy, hyponymy (edge weight 1.0), meronymy, or holonymy (edge weight 0.5). We then run node2vec (Grover and Leskovec, 2016) over this graph to learn 1024-dimensional embeddings, using default values for all other hyperparameters.

[45]This is a slightly lossy mapping, as we use the key of a synset's first lemma, and not all sense keys annotated in SemCor correspond to the default lemmas.

NASARI Camacho-Collados et al. (2016) developed multilingual sense embeddings, derived from pre-trained word embeddings and global lexical statistics. We use 300-d NASARI embeddings trained on the UMBC corpus,[46] and map them from BabelNet synsets to WordNet 3.0 synsets.

7.1.4 Experiments

For all experiments, we used a 1-layer DNN with ReLU activation[47] and trained our model using minibatch gradient descent with Adam optimization and a minibatch size of 5. After each training epoch, we evaluated cumulative cross-entropy loss over development data, and used early stopping with patience to stop training. After halting, we evaluated our best-performing model from dev on test data.

We randomly sampled 10% of SemCor data (stratified by label) as held-out development data for model training; SemEval 2007 was used as a second development set for hyperparameter tuning, as it captures generalization performance better than SemCor data. We use an early stopping threshold of 1e-4, and patience of 5 epochs.

Handling unseen lemmas and senses

Two types of out-of-vocabulary (OOV) issues arose in our evaluation: unseen lexical forms and unknown senses. In prior work, the set of candidate senses for a given lemma and part-of-speech (POS) tag is determined by the training data; if a new lemma/POS combination is seen at test time, it is considered OOV and backoff to the first synset returned by WordNet is used (Raganato et al. (2017b); Iacobacci et al. (2016), *inter alia*).

[46] `http://lcl.uniroma1.it/nasari/`

[47] We compared 0-3 hidden layers, with ReLU, sigmoid, and linear activations. Residual final layer is always linear.

Embeddings	VecSim	MM	RW	Res	Dir
SemCor (SC)	56.3	62.9	62.6	60.4	62.9
Definitions (Defn)	39.6	54.3	56.3	54.5	56.3
WordNet (WN)	31.2	54.9	53.8	54.5	54.3
NASARI	–	53.2	–	–	53.0
SC+Defn	48.4	59.3	61.8	61.5	62.6
SC+WN	48.1	59.8	59.6	61.8	61.5
SC+NASARI	–	59.6	–	–	61.8
SC+Defn+WN	49.9	**64.6**	60.2	60.0	63.3
SC+Defn+NASARI	–	60.9	–	–	60.4
All	–	63.3	–	–	61.1

Table 7.2: Macro F-1 on WSD dev set (SemEval-07) senses with different PROSE configurations, including different sets of sense embeddings; the best configuration for each set is underlined, and the best overall result bolded. For brevity, we only report combinations of $n+1$ embeddings using the best choice of n embeddings. Re-weighting, Residual, and linear scorer results are not given for NASARI, due to its different vector dimensionality. VecSim=vector similarity baseline, MM=MatrixMult, RW=Re-weighting, Res=Residual, Dir=Direct.

Model	Dev	Test Datasets				Concatenation of Test				
	SE-07	SE-2	SE-3	SE-13	SE-15	Nouns	Verbs	Adj.	Adv.	All
Baselines										
WordNet First Sense (WNFS)	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5
VecSim (SemCor)	56.3	71.6	67.8	65.6	71.0	70.2	55.2	**78.2**	83.8	69.0
Word experts										
SemCor (tuned)	60.7	70.4	68.9	64.7	69.6	70.4	53.8	75.8	82.4	68.5
Iacobacci et al. (2016)	62.6	72.2	70.4	65.9	71.5	71.9	56.6	75.9	**84.7**	70.1
Papandrea et al. (2017)	62.4	72.3	69.7	66.7	71.9	72.0	56.9	76.3	83.5	70.2
Unified models										
Raganato et al. (2017b)	63.7	72.0	69.4	66.4	72.4	71.8	57.1	75.6	83.2	70.0
HCAN (OOVs omitted)	64.9	72.8	70.3	68.5	72.8	72.7	58.2	77.4	84.1	71.1
HCAN + WNFS backoff	63.5	71.7	69.2	66.9	71.2	71.4	56.6	76.5	83.5	69.8
fastSense (Uslu et al., 2018)	62.4	73.5	**73.5**	66.2	73.2	–	–	–	–	–
Kumar et al. (2019)	**67.3**	**73.8**	71.1	69.4	**74.5**	**74.0**	**60.2**	78.0	82.1	**71.8**
Our model										
PROSE MatrixMult (SC)	62.9	72.9	71.5	67.7	70.4	72.1	59.0	77.2	**84.7**	70.9
PROSE Re-weighting (SC)	62.6	72.8	71.1	69.5	71.4	73.1	58.0	77.9	84.1	71.3
PROSE Residual (SC)	60.4	71.5	69.8	67.9	71.0	72.0	57.6	75.2	82.9	70.1
PROSE MatrixMult (SC+DF+WN)	64.6	71.5	71.5	**70.4**	71.7	73.2	58.0	76.6	**84.7**	71.3
PROSE Re-weighting (SC+DF+WN)	59.8	70.4	70.2	67.4	71.6	71.4	56.9	75.7	84.1	69.8
PROSE Residual (SC+DF+WN)	60.0	70.4	69.7	68.1	70.6	71.4	56.4	76.2	83.5	69.7

Table 7.3: Macro F-scores (%) for English all-words fine-grained WSD in evaluation framework. HCAN omits OOV forms; for direct comparibility, we include their results augmented with WordNet first sense (WNFS) backoff. All other prior models use WNFS backoff for OOV forms. Uslu et al. (2018) do not include concatenated results. Best performance on each subset is marked bolded; best results from our model are underlined. VecSim=vector similarity baseline; SC=SemCor embeddings; DF=WordNet definition embeddings; WN=WordNet graph embeddings.

In our approach, the set of available senses is determined by the sense embeddings used. SemCor embeddings followed prior work; for other embeddings, we took the POS tag and default lemma for the sense as "in vocabulary". Thus, at test time, a lemma/POS combination was only OOV if it was not covered by any of our input sense embeddings; we used WordNet first sense backoff in these cases.

Finally, for any given lexical form, not all senses were covered by all of our evaluated sense embeddings. Since the combined sense projection s' is aggregated over the individual projections s'^a, we replaced any senses that are OOV for an individual embedding set with a zero vector, so that s' is derived only from the sets that include the target sense. During training, any samples that were OOV for *all* embedding sets used were discarded from both train and dev. All samples were evaluated at test time, regardless of coverage.

Experimental results

Comparing PROSE configurations across different combinations of embedding sets for SemEval-2007 WSD in Table 7.2, the MatrixMult setting most consistently yielded the best dev set results (3/6 comparisons to Re-weighting and Residual). Interestingly, SemCor senses embedded with ELMo outperformed any single set or pair of embeddings on development data,[48] but combining SemCor, definitions, and WordNet graph embeddings (with or without NASARI) increased dev performance.

Table 7.3 shows WSD performance across the four test sets and by POS tag within the concatenation of all test sets. We compared against two baselines: WordNet first sense (WNFS) and the unprojected vector similarity model of Sabbir et al. (2017), using ELMo-derived SemCor embeddings to measure performance of contextualized

[48]This also held for other single embedding sets.

features without learned projection. For word expert approaches, in which a separate model is trained for each surface form, we compared to individual sense embeddings directly tuned during training (instead of using a shared projection) and the strongest prior expert approaches (Iacobacci et al., 2016; Papandrea et al., 2017), which train one SVM per lemma/POS pair using POS tags, unigram features, and collocations of surrounding words together with summed embeddings of context words with exponential decay. Finally, for neural models, we compared with Raganato et al. (2017b), who use a bi-LSTM with attention and auxiliary objectives, Luo et al. (2018), who use a hierarchical co-attention network with word context and sense gloss information, fastSense (Uslu et al., 2018), which uses averaged word embeddings to train a multi-label feed-forward neural model that ranks candidate senses,[49] and Kumar et al. (2019), who learn bi-LSTM encoders for context and sense embeddings. All comparison methods use WNFS backoff for OOV forms,[50] except for Luo et al. (2018), who omit OOV forms[51] (we compared against augmenting their results with WNFS backoff for parity) and Kumar et al. (2019), who use sense definitions to handle OOVs.[52]

PROSE performed competitively overall, with the best configurations achieving overall F1 within 0.5 of the much more complex state-of-the-art approach, yielding

[49]Uslu et al. (2018) use SensEval-2 for hyperparameter tuning, instead of SemEval-2007; results on these datasets are thus not directly comparable.

[50]T. Uslu, personal communication

[51]F. Luo, personal communication

[52]Following Raganato et al. (2017b), we did not include Yuan et al. (2016) in our comparison, as neither their models nor their training data are publicly available for replication on the benchmark test sets.

	SE07	SE2	SE3	SE13	SE15	All
OOV Samples	20	243	107	202	98	650
Covered	16	194	66	113	74	447
OOV Senses	19	103	91	132	65	768
Covered	15	64	53	78	45	557
WNFS	75.0	**94.8**	74.2	76.1	**79.7**	80.0
DF (A)	56.3	81.4	77.3	**87.6**	**79.7**	**82.9**
SC+DF (A)	50.0	80.9	74.2	81.4	77.0	80.1
SC+DF+WN (MM)	62.5	78.9	75.8	85.8	75.7	80.3
All (MM)	**81.3**	80.9	**78.8**	85.0	77.0	81.5

Table 7.4: Generalization evaluation in WSD experiments. Macro F-1 % is reported on OOV lemma/POS combinations and gold senses (w.r.t. SemCor) covered by WordNet definition embeddings, for the best-performing PROSE configurations of embedding combinations. A=Re-weighting, MM=MatrixMult. Some samples have multiple valid senses.

state-of-the-art F1 of 70.4% on SemEval-2013, and consistently outperforming several recent neural models on multiple datasets.[53] Comparing against contextualized baselines, PROSE projection of SemCor embeddings consistently outperformed both linear scoring without projection and tuning sense embeddings directly, indicating the value of augmenting contextualized features with a learned projection model. Incorporating multiple sets of sense embeddings yielded improved dev set performance, but only sometimes improved test results: the combination of SemCor senses, WordNet definitions, and WordNet graph embeddings only outperformed SemCor alone under the MatrixMult configuration, and tied SemCor senses alone with the Re-weighting setting for the highest overall performance.

[53]Uslu et al. (2018) did not report results on the concatenated test set.

Zero-shot disambiguation

An additional strength of a sense embedding-based approach is support for zero-shot disambiguation of arbitrary lemma/POS combinations. We identified the samples in each evaluation set that are OOV with respect to SemCor lemmas, requiring backoff to WordNet in most prior supervised approaches. Of these samples, we identified the subsets covered by using different sense embedding sets with PROSE, and compared WSD performance on these subsets to the WNFS baseline.

Table 7.4 shows the results when we used WordNet definition embeddings, alone and combined with other sets. These embeddings covered over half of the OOV samples in the evaluation sets. On SemEval-2007, SensEval-3, and SemEval-2013, our model outperformed the baseline by a large margin on the covered samples, and tied it on SemEval-2015. Notably, comparing across the four test sets combined, our model beat the WordNet first sense baseline by 2.9% macro F1.

7.1.5 Analysis

We have shown that a neural projection model can effectively combine contextualized embedding features with diverse sense embeddings to achieve WSD performance on par with state-of-the-art methods, and can successfully disambiguate lexical forms not seen during training. Analysis of our results suggests some interesting areas for improving use of sense embeddings for disambiguation.

Challenges in highly ambiguous words

Not all ambiguous words are equal: the verb *discuss* has two senses in WordNet, but the verb *change* has ten to choose from. As might be expected, our model's performance tends to decrease as the number of candidate senses for a word increases.

Review of model predictions on development data highlights an interesting contributing trend: our approach tends to assign higher weights to more generic and/or literal classes. For example, for "among the problems, the one at HUD," the numeric sense of "one" is given 78% probability, but the correct sense of "a single person or thing" only 22%. Multiple generic senses can share output scores: in "I had to reach back to French 101," reaching a destination was scored 26% likely, reaching physically upward 25%, and the correct sense "to extend as far as" only 5%.

This genericity preference varies by context. One example of the verb "follow," referring to following progress, led to misprediction of "to travel behind;" another referring to a sequence of events was mispredicted as "be next;" and one referring to following someone's lead was correctly predicted by a low margin over "choose and follow [a theory]." This suggests that distinguishing literal and figurative language is an important area to improve WSD. In cases where a word is more likely used figuratively, more generic senses can be downweighted in favor of figurative ones.

Embedding synthesis and zero-shot analysis

The generalization results in Table 7.4 are strong in many cases, but demonstrate clear remaining areas of improvement. Many of the samples we successfully disambiguate have only one corresponding sense in our embeddings; by the same token, many of our incorrect predictions are due to not having an embedding for the correct sense. However, a reasonable portion of OOV samples had multiple valid candidates, and we find clearly more of these samples to be disambiguated correctly in our results. Going from one set of sense embeddings to multiple sets has surprisingly little effect on final predictions for OOV samples, but as shown in Figure 7.4, it noticeably increases the model's confidence in its predictions, both correct and incorrect.

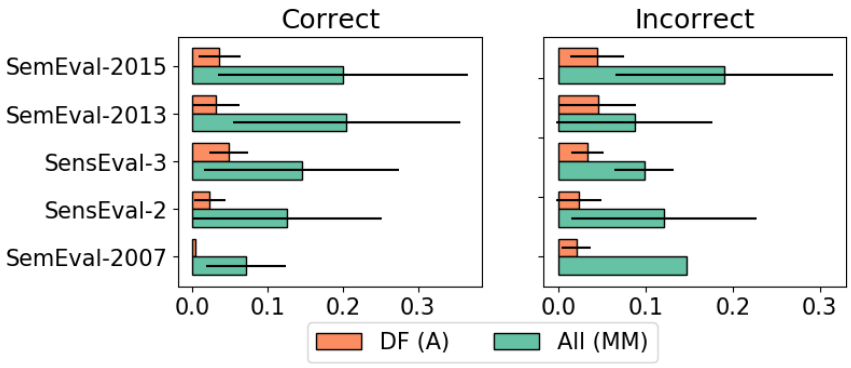


Figure 7.4: Mean success/error margins with PROSE, measured between scores assigned to correct sense and next most likely sense (for correct predictions) or correct sense and highest-scoring sense (for incorrect predictions), for OOV samples within each dataset. Error bars indicate standard deviation.

The model is also able to successfully synthesize across projected embeddings, whether or not each individual projection yields the correct prediction. As shown in Figure 7.5, over 60% of the correct predictions made by PROSE with all four sets of embeddings involved projections where only one or two of the individual embedding sets yielded the correct answer, and 6% of the time the model made the correct prediction when *none* of the individual projections did. However, the likelihood of producing an incorrect overall answer increases as the number of projected embedding sets with the correct answer decrease; 70% of the incorrect predictions made by our model involved either zero or only one of the projected embedding sets producing the correct answer.

7.1.6 Conclusion

Contextualized word embedding methods provide powerful features for word sense disambiguation, but have not systematically been leveraged in recent WSD models.

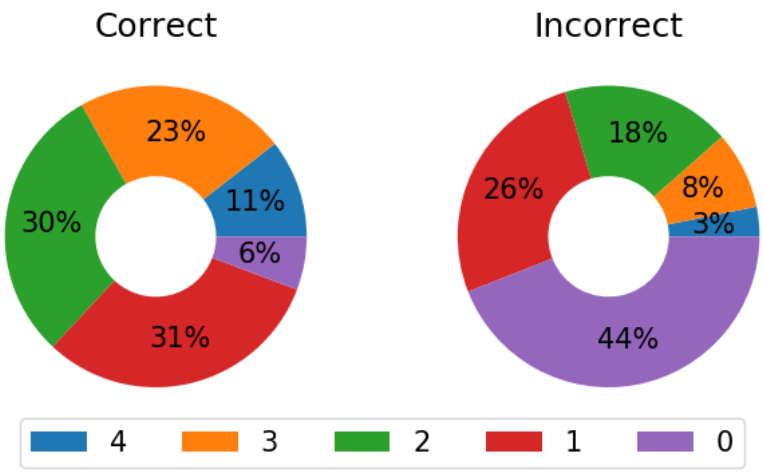


Figure 7.5: PROSE successes/failures by number of contributing embedding sets; using PROSE MatrixMult (All).

We have demonstrated that contextualized embeddings can be effectively combined with sense embeddings via a context-sensitive projection method to achieve comparable performance to complex, task-specific state-of-the-art WSD models and enable zero-shot disambiguation of unseen lemmas. Further, our demonstrated gains on both benchmark WSD datasets and a low-resource biomedical coding task illustrate the generalized value of combining contextualized embedding features with targeted disambiguation models. Analysis of our system outputs showed that our system is able to leverage sense embeddings from diverse knowledge sources, and highlighted figurative language as an outstanding challenge for improving our model's performance. Our code, trained models, and associated data will be made available online.

7.2 Application to medical concept normalization

Medical Concept Normalization (MCN) is the task of assigning canonical identifiers to text mentions of medical concepts, in order to unify different ways of referring

	n2c2 2019		SemEval-2015 Task 14	
	Train	Test	Train	Test
Documents	50	50	298	133
Samples	6,684	6,925	11,554	8,003
CUI-less	151	217	3,478	1,930
Distinct CUIs	2,330	2,578	1,355	1,143
CUI overlap	1,117		628	

Table 7.5: Datasets used for PROSE MCN experiments. Both datasets are split into training and test data at the document level; the number of documents, medical concept mentions, CUI-less mentions, and unique CUI labels are given for each, along with the number of CUIs overlapping between training and test.

to the same concept. MCN has historically been studied jointly with medical Named Entity Recognition (NER) (Elhadad et al., 2015), but some recent shared tasks have investigated NER and MCN as separate components of information extraction (Pradhan et al., 2014; Uzuner et al., 2019). In this study, we investigated the application of our PROSE model to MCN, as part of participating in the 2019 n2c2/UMass Shared Task on Clinical Concept Normalization (Uzuner et al., 2019).

7.2.1 Datasets

In addition to the dataset for the n2c2 2019 shared task, we leveraged data from the previous SemEval-2014 Task 7 challenge, in order to pretrain our normalization models on a greater diversity of data. Brief descriptions of each dataset are given below, and summary statistics are given in Table 7.5.

n2c2-2019 Track 3

The n2c2 2019 MCN dataset was originally developed by Luo et al. (2019) in order to complement the narrow focus of previous MCN datasets, which had only

included disorder-related information (Uzuner et al., 2011; Elhadad et al., 2012). The dataset of Luo et al. (2019), which they simply called MCN to reflect the task it was intended for, included 100 documents from the 2010 i2b2 challenge on medical information extraction (Uzuner et al., 2011). These documents had been annotated to identify all mentions of medical problems, treatments, or tests; Luo et al. (2019) further annotated these mentions to assign Concept Unique Identifiers (CUIs) from the 2017AB release of the Unified Medical Language System (UMLS) (Bodenreider, 2004), using the SNOMED-CT and RxNorm vocabularies. The documents were split into 50 for training and 50 for testing, maintaining similar CUI distributions between train and test.

The breadth and complexity of medical concepts in the UMLS means that MCN annotation is never straightforward. Luo et al. (2019) describe several approaches to addressing annotation challenges, two of which are particularly relevant to highlight here. First, annotators were asked to normalize each mention using only the mention text whenever possible, and incorporate context only when the mention text was not sufficient to identify a unique CUI—i.e., in the cases of ambiguous (or uninformative) strings. Second, mention texts describing a compositional concept (e.g., "left breast biopsy") were split into multiple mentions such that each mention could be normalized to a unique CUI (e.g., "left", "breast biopsy"). Thus, all mentions have exactly one CUI annotated, and mentions were only labeled as "CUI-less" when an appropriate concept could not be found in either SNOMED-CT or RxNorm.

SemEval-2015 Task 14

The ShARe corpus is a collection of 531 EHR documents collected from the Mayo Clinic, different subsets of which have been used as the source for a variety of shared

tasks in clinical NLP (Mowery et al., 2014; Pradhan et al., 2014; Elhadad et al., 2015). These documents, including discharge summaries and echocardiogram, electrocardiogram, and radiology reports, have been annotated for a variety of tasks, including MCN. The most recent shared task using these data for MCN was SemEval-2014 Task 15 (Elhadad et al., 2015), which used 431 documents for training and 100 for testing. As our use of ShARe corpus data was as additional training data for the n2c2 2019 task, we restricted our experimentation to the training data only, and used as training/development sets the splits from SemEval-2014 Task 7 (Pradhan et al., 2014), which formed the full training set for the later shared task.

These documents were annotated only for disorder-related concepts, restricted to the SNOMED-CT vocabulary (Elhadad et al., 2012). Compositional mentions and mentions that could not be assigned a disorder-related SNOMED-CT CUI were marked as "CUI-less", covering a relatively large portion of the dataset (Osborne et al., 2018). There is relatively low overlap in both strings (shown in Table 3.8) and CUIs between these data and the n2c2 2019 task data: only 405 CUIs are present in both SemEval-2015 Task 14 data and n2c2 2019 training data. Nonetheless, given the number of novel (i.e., not observed in training) CUIs in the n2c2 2019 test data, we hypothesized that these additional training data would improve generalization power for our MCN models.

7.2.2 Methods

Controlled vocabularies aim to capture diverse natural language expressions used to refer to standardized concepts, making them a powerful first step for MCN. However, they are not exhaustive—the productivity of language means that new expressions can emerge to refer to standardized concepts—and any given expression may be ambiguous (or multi-referent) between multiple distinct concepts. Thus, we can consider three contributing factors to normalization outside of controlled vocabulary match: (1) reference with novel expressions, (2) expressions that refer to multiple concepts (e.g., compositional expressions), and (3) ambiguous expressions.

Mention matching heuristics

The datasets used in our experiments addressed compositional expressions in the annotation phase, removing this problem from a modeling standpoint. For the problem of novel expressions, two contributing factors that can be addressed heuristically are institutional terminology specific to local practitioners and lexical reordering/adjustment of multi-word expressions (e.g., "gait is antalgic" vs "gait, antalgic" vs "antalgic gait"). As the shared task dataset is from a single document collection, we follow Luo et al. (2019) in using matches from the training data as a way to capture dataset-specific expressions. To help address lexical variations, we utilized MetaMap (Aronson and Lang, 2010), which includes a number of heuristics for string matching.

Neural normalization with PROSE

Finally, to help address both the problems of ambiguity and expressions that could not be normalized through the above heuristics, we utilized our PROSE model as the

final component of our system. This enabled us to use representation-based features to choose between candidate CUIs without limiting our model to string-based matches. In contrast to a pure neural classifier, however, in which candidate CUIs would be treated as orthogonal labels and all relevant information learned from supervised signal in the task, using PROSE allowed us to leverage CUI representations learned from diverse sources as an informed starting point for neural normalization.

Candidate CUI selection strategies As PROSE is formulated as a learned similarity scorer between text representations and a set of candidate sense representations, we utilized two approaches to identify the set of candidate senses (here, CUIs) to consider for a given mention. To address samples where the mention string partially corresponded to known medical terms (ambiguous strings, but also lexical edits or extra words), we queried the UMLS REST API with the mention string, using the word-level approximate search setting. This search type includes some lemmatization, and returns CUIs in decreasing similarity order to the query; we chose the top 30[54] CUIs as our candidate set.

Some mention strings did not yield any CUIs from UMLS query, due to misspellings, genericness, or string complexity. In these cases, we backed off to choose between all 434,056 CUIs in SNOMED-CT and RxNorm. This backoff strategy was only used at test time, as the time required to calculate projection-based similarity to all CUIs was impractical for model training; these samples were instead dropped from the training data.

[54]Empirically-chosen; 100 CUIs (and higher values) did not appreciably increase hit rate for including the correct CUI, and significantly decreased performance.

Embeddings	Method	Data	# CUIs
PubMed	JET	PubMed 2018 baseline	89,713
Graph	node2vec	UMLS knowledge graph	420,892
Definitions	Word averaging	UMLS definitions	192,313

Table 7.6: CUI embeddings used for n2c2 2019 shared task. The method and data used to generate the embeddings are given, along with the number of SNOMED-CT and RxNorm CUIs covered in each set. SNOMED-CT and RxNorm encompass 434,056 CUIs in total.

CUI and text representations In order to utilize different types of biomedical knowledge in our model, we provided CUI representations learned from multiple sources as inputs to PROSE; these representations are summarized in Table 7.6.

PubMed abstracts To capture information about usage of medical concepts in text, we used our JET toolkit (Newman-Griffis et al., 2018) to learn CUI representations from the 2018 PubMed baseline, using all terms from Level 0 vocabularies (plus SNOMED-CT) in the 2017AB release of the UMLS.

UMLS graph To capture hierarchical relationships between UMLS concepts, we extracted the graph structure captured by PAR and CHD relations in the UMLS, and trained CUI representations on the graph with node2vec (Grover and Leskovec, 2016).

UMLS definitions Expert-written definitions are included in the UMLS for a subset of concepts. These definitions have been utilized as a source for generating concept representations in previous work, by averaging embeddings for every word in each definition (Pakhomov et al., 2016). We follow Pakhomov et al. (2016) to create extended definitions for each CUI by concatenating the definitions of all parent, children, and

sibling CUIs. Each CUI's representation is calculated by averaging its definition and 50%-weighted extended definition (where present).

Mention representations We experimented with two different methods for representing the context of each concept mention: static and contextualized embeddings. For static features, which are less powerful but capture lexicalization, we used 300-dimensional word embeddings trained jointly with our PubMed CUI embeddings using JET, and represented each mention by averaging the embeddings of each word in a fixed context window around the mention, along with the text of the mention. For contextualized features, we used 768-dimensional clinicalBERT (Alsentzer et al., 2019) to embed the full line containing the mention and average the token-level embeddings for the mention.

We generated two versions of each of our sets of CUI embeddings, for use with each mention representation method. Definition and string CUI embeddings were generated using JET word embeddings and clinicalBERT directly; for PubMed and UMLS graph embeddings, we re-ran JET and node2vec to generate 300-dimensional and 768-dimensional embeddings.

Ensembling models with different features In order to leverage the relative strengths of static and contextualized features, and to capture information from different context window sizes, we experimented with ensembling multiple PROSE models. We utilized both simple similarity averaging across models and a learned combination, using either a feed-forward neural network to rescore or a random forest to choose which model's scores to use. Our models for ensembling included four static embedding models, using window sizes 15, 50, 75, and 100, and a BERT model (which has no variable context).

Training details All PROSE models were trained for 50 epochs, using all training set mentions with candidate CUIs retrieved using UMLS API query. As UMLS query may not include the correct CUI for some mentions among the candidates returned, meaning these samples cannot provide any training signal, we filtered out these mentions during training. We used a 1-layer PROSE architecture with the Residual configuration, with hidden layer size matching the input embedding size.

We experimented with pre-training PROSE models using the SemEval-2015 Task 14 training data. For each model configuration used on the n2c2 2019 data, we pre-trained a model with the matching configuration on SemEval-2014 Task 14 data, using the same 50 epoch training scheme. The resulting model parameters were then used to initialize the weights of the model trained on n2c2 2019 data.

PROSE-only experiments To compare our various PROSE models and ensembling strategies, we ran 5-fold cross-validation experiments on the n2c2 2019 training data, and reported micro-averaged accuracy across folds. We compared the following experimental variables:

Feature configuration: comparing static and BERT input features, and different window sizes for static representations;

Pretraining: comparing PROSE models with and without pre-training on SemEval-2015 Task 14 data;

Ensembling strategies: comparing different combinations of individual PROSE models, using score averaging and machine learning approaches.

After identifying the best combination of PROSE models for both PROSE - Mention and PROSE - All sieves, we added our first three sieves into the system for final evaluation on training data and submission of test data results for the shared task.

Normalization system overview

Our full system consisted of six sieves, each of which covered a cumulatively increasing portion of the dataset. If a given mention was covered by an earlier sieve, we used that sieve's prediction and did not further process the mention. Our sieve progression is illustrated in Figure 7.6, and described briefly below. Our first three sieves follow the described baselines used by Luo et al. (2019).

Exact Match - Training The first stage of our system compares the current sample's mention string to all mentions in the training set; if one or more matches are found that are labeled with a single unique CUI, that CUI is output. We use two stages of matching: lowercased string, and lowercased string with stopwords removed.

MetaMap Our second stage uses MetaMap (Aronson and Lang, 2010) with term processing on the mention string; if a unique CUI is produced, that CUI is output.

Exact Match - UMLS The third stage compares the sample's mention string to the strings in the UMLS MRCONSO table, following the same normalization steps as in the Exact Match - Training sieve. Again, if a unique CUI is matched, that CUI is output; otherwise, the mention string continues to the fourth stage.

PROSE - Mention In our fourth sieve, we query the UMLS API for the mention string, and pass the top 30 candidates into our trained PROSE model(s) to identify the highest-scoring candidate. If no results are returned from the API query, the mention proceeds to the final sieve.

PROSE - All Our fifth sieve is our backoff approach: the mention is passed into our trained PROSE models to score all CUIs in SNOMED-CT and RxNorm. This sieve achieves complete coverage of the dataset, but is the hardest to get right.

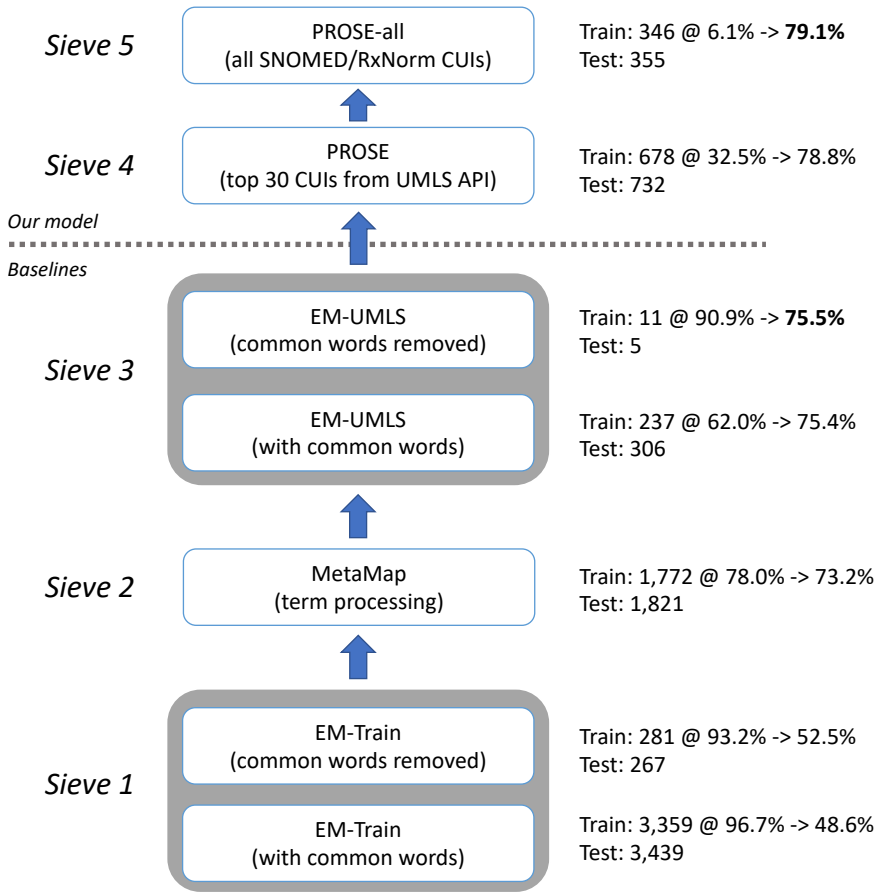


Figure 7.6: Sieve-based normalization system for n2c2 2019 MCN shared task. The column at right lists the number of training and test instances covered by each sieve, along with the accuracy of the sieve on its addressed samples and the overall cumulative accuracy of the system up to that sieve. EM-Train and EM-UMLS are our exact match heuristics.

7.2.3 Results and analysis

BERT fails in open-ended candidate selection

Figure 7.7 shows the cross-validation performance of our various PROSE models alone on the n2c2 2019 training data. BERT features yield a notable improvement over static featuers in the PROSE - Mention setup (involving choosing between up to

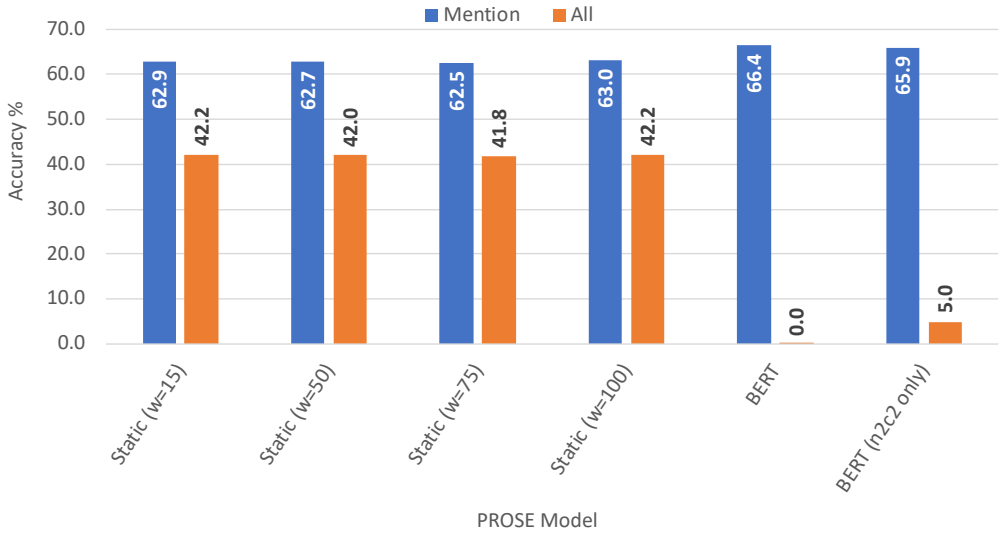


Figure 7.7: Cross-validation accuracy of individual PROSE models on n2c2 2019 training data, using API candidates or all candidates for evaluation.

30 candidates), but fail spectacularly in PROSE - All experiments (involving choosing between all 434,000 available CUIs). On PROSE - All, BERT features yield 5% accuracy or less, while static features consistently achieve a respectable 42% accuracy. While no explanation for this discrepancy is readily apparent, it is clear that static and BERT features capture some degree of complementary information for our formulation of the task.

With static representations, pretraining on SemEval-2015 Task 14 data yielded universal improvement, on both PROSE - Mention and PROSE - All experiments. With BERT features, however, we observed a slight improvement in PROSE - Mention performance from pre-training, but a significant *degradation* from 5% to 0.04% accuracy in PROSE - All experiments.

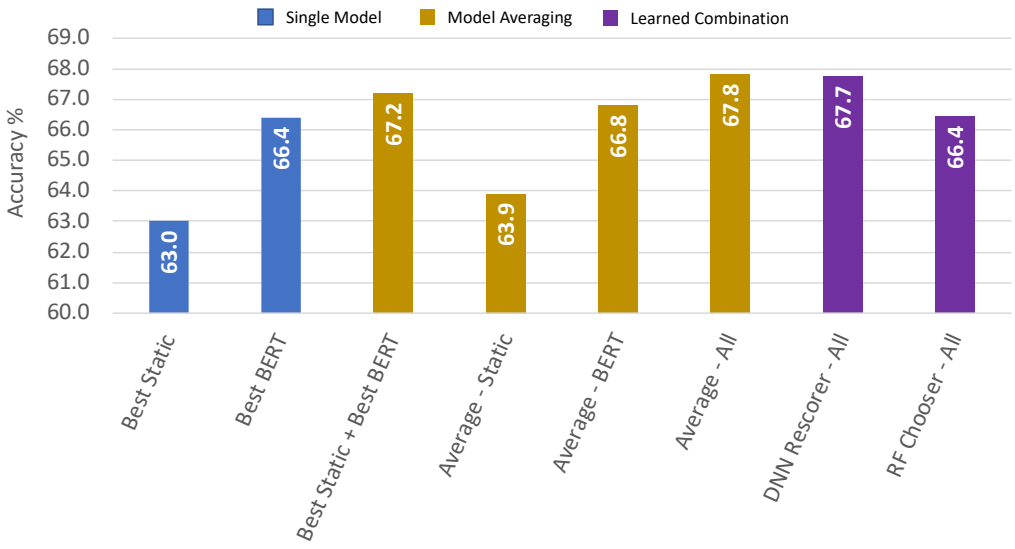


Figure 7.8: Cross-validation accuracy of ensembling strategies with PROSE models on n2c2 2019 training data, comparing single models (first two bars), averaging model scores (middle four), and combining scores with machine learning (right-hand bars). All results are for PROSE - Mention only.

Ensembling with model averaging significantly improves performance

Cross-validation performance from different ensembling strategies are shown in Figure 7.8. Combining scores from static and BERT models improves results over either by 0.8%, strengthening our observations of complementary information from the two representation methods. While we did not observe any meaningful difference between different window sizes for generating static input features (Figure 7.7), combining across window sizes nonetheless improves performance by 0.9%, indicating some complementarity between what the context windows capture. Combining our pretrained and non-pretrained BERT models also improved accuracy by 0.4%, suggesting that the two models may have learned different patterns from their different training data exposures.

Model	Training cross-validation accuracy	Test accuracy
Luo et al. (2019)	77.1	76.4
Baselines	75.5	–
Our system	79.1	78.1
Shared task mean	–	74.3
Shared task best	–	85.3

Table 7.7: Results on n2c2 2019 shared task. Baselines refers to our reimplementation of the methods used in Luo et al. (2019); test results are not given for these, as we did not submit them for evaluation in the shared task.

Combining scores from all six models yielded the best overall performance, increasing accuracy by a further 0.6% or greater over any other combination. We found that using machine learning to combine model outputs matched but did not exceed score averaging in the best case; as score averaging is more parsimonious, we therefore used this as our ensembling strategy.

Shared task performance

As shown in Table 7.7, our full, five-sieve system improved over the published baseline for the shared task dataset by 2% on cross-validation in the training data, and 1.7% on the test data. Interestingly, our reimplementation of the baselines using published details fell 1.6% short of the published performance, suggesting that with additional corrections, performance on our earlier sieves could be improved, yielding improved overall performance. In comparison to other participants in the shared task, we outperformed the mean test accuracy by nearly 4%, but placed well behind the top performing system.

Directions for further analysis

Two observations indicate specific directions for further analysis of our results. First, the behavior discrepancy between static and BERT features, including different sensitivities to pretraining and BERT's failure in the PROSE - All setting, indicate that this is likely to be a fruitful area for analyzing system outcomes. Second, as observed in Chapter 3, the n2c2 2019 dataset exhibits a very low degree of ambiguity, but does include distinct types of ambiguity; as our PROSE model is designed with normalizing ambiguity in mind, further comparison on this subset of the data and on other MCN datasets with more ambiguity is likely to clarify the strengths and weaknesses of our approach to the MCN task.

7.3 Classifying mobility activity types

The lack of standardized terminologies for functional activity makes activity normalization a natural fit for a model like PROSE, which does not require any explicit lists of known surface forms for concepts. Following our experiments on extracting mobility-related information, described in Chapter 6, we therefore applied PROSE to the next step of the processing pipeline: normalizing mobility-related activity reports to ontologically distinct activities in the ICF.

As in Section 6.2, we used an expanded version of the dataset described by Thieu et al. (2017), encompassing 400 Physical Therapy records from the NIH Clinical Center. These documents include a total of 4,528 activity reports with reference to a specified activity; these reports were assigned one of 12 ICF codes identifying different mobility activities, or an *Other* label if none of the available ICF codes

Code	Description	Frequency
d410	Changing basic body position	838
d415	Maintaining a body position	612
d420	Transferring oneself	522
d430	Lifting and carrying objects	44
d435	Moving objects with lower extremities	5
d440	Fine hand use	10
d445	Hand and arm use	66
d450	Walking	1,603
d455	Moving around	378
d460	Moving around in different locations	176
d470	Using transportation	38
d475	Driving	77
Other	–	161
	Total	4,528

Table 7.8: Label descriptions and frequencies in mobility activity normalization dataset. Descriptions given are the preferred name for the code in the ICF (World Health Organization, 2001).

applied. Table 7.8 provides descriptions of these labels and their frequencies within the dataset.

7.3.1 Methods

The only previous study on functional activity normalization was by Kukafka et al. (2006), who used hand-crafted rules to extract and code mentions of a small set of ICF codes. Thus, particularly since our set of activity labels is quite constrained in these data (thirteen, not thousands as in MCN), we wanted to investigate both *classification*-based approaches and *candidate selection*-based approaches to the task.

Classification methods

We experimented with several strong baseline methods for classifying mobility activity reports, including k-Nearest Neighbors (k-NN), Support Vector Machine

(SVM), Multi-Layer Perceptron (MLP; i.e., a feed-forward neural network), and BERT fine-tuning (Devlin et al., 2019). As many of the action codes are quite distinct from one another and likely to be discussed in very different contexts, we experimented with lexical features as well as word embedding features for representing sample activity reports. To control for activity report length, we used binary unigram features; for word embeddings, we used either the averaged static embeddings of each word in the report, or average BERT activation. BERT fine-tuning inherently uses the BERT representation of the whole sequence.

Candidate selection methods

Classification methods, while powerful, lack two important factors for practical application of concept normalization: flexibility to new concepts (as the label set is fixed) and degrees of relatedness between labels (no representation of labels means that they are orthogonal to one another). A *candidate selection* method like PROSE, however, takes as input representations of a set of candidate senses, and simply chooses which of these candidates is most representative of the sample. Thus, it can be extended to new labels (by adding their representations to the candidate set), and direct representation of labels allows for inter-label information.

PROSE is a system with multiple components. To distinguish between effects of label representation alone and effects specific to the PROSE architecture, we performed three sets of experiments: the first using cosine similarity alone to compare an activity report representation to representations of the candidate ICF codes, the second using the composite vector similarity model of Sabbir et al. (2017) (as used in PROSE), and the third being a full PROSE system.

Our concept representations were derived from the descriptions given for each code in the ICF (World Health Organization, 2001). These descriptions were passed as input to clinicalBERT (Alsentzer et al., 2019), and the token-level representations averaged over the full descriptions to generate ICF code embeddings. This approach meant that we were not able to create a representation for the *Other* label; our candidate selection models could therefore only predict one of the 12 ICF codes. We leave addressing this issue for future work. Representations of the activity reports were also generated by passing them to clinicalBERT and averaging the token representations for the portion of the report referring to the activity (included in the dataset annotations).

Experiments

We used ten-fold cross validation for all experiments (stratified by label), as the dataset did not have a pre-generated train/test split. The same splits were used for all experiments, to control for random factors in split re-generation. For feature and model selection experiments within the classification and candidate selection frameworks, we used a held-out 10% of each fold's training data for development, and reported results combined across these sets. For final experiments with the best classifier and candidate selection models, we used the full training data for each fold, and reported results combined across the test sets.

7.3.2 Results

SVMs with static embeddings are the best classifier

Table 7.9 shows the results of our feature and model selection experiments within the classifier framework. Static word embedding features yield the best results within

Features	k-NN	SVM	MLP	BERT
Unigrams	72.0	77.1	76.5	–
Static embeddings	89.0	**93.4**	93.0	–
Combination	71.1	82.8	82.8	–
BERT	84.4	91.1	91.2	77.5

Table 7.9: Mobility activity normalization results across classifier methods and features, using dev set accuracy from 10-fold cross validation.

each of k-NN, SVM, and MLP models. While unigram features provide a strong baseline in each case, the decrease in performance when adding them to word embedding features indicates that the predictive features in each set are negatively cross-correlated, suggesting that lexical signals are counterproductive for this task. Surprisingly, static embedding features outperform BERT features in all models, and BERT fine-tuning significantly underperforms the other classification models using BERT features. The best overall performance of 93.4% development accuracy is achieved with an SVM model using static word embedding features.

Transforming concept representations with PROSE is essential

Figure 7.9 presents the results of using different candidate selection models for activity normalization. Using supervised learning to learn a context-sensitive projection model for code representations with PROSE is essential for this task, yielding over 60% improvement in accuracy over cosine similarity alone. The composite vector similarity method of Sabbir et al. (2017) yields the same results as unmodified cosine similarity, indicating that the difference with PROSE comes from the learned projection.

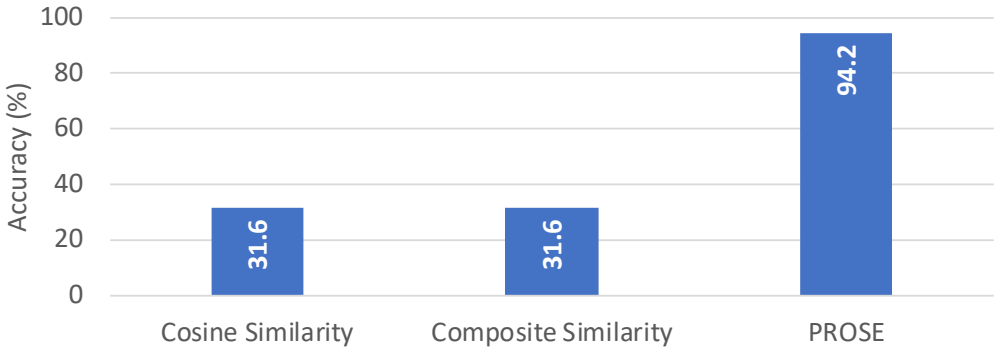


Figure 7.9: Mobility activity normalization results across candidate selection models, using development sets in cross-validation experiments.

Model	With Other		Without Other	
	Accuracy	Macro F-1	Accuracy	Macro F-1
Best classifier	93.4	85.8	94.5	87.0
Best candidate selector	89.5	71.1	92.8	77.0

Table 7.10: Mobility action normalization results with classification and candidate selection frameworks. Accuracy and macro F-1 are provided for each setting, both including the *Other* label (which the candidate selection framework cannot choose) and excluding it.

Both classification and candidate selection approaches are strong on this dataset

As shown in Table 7.10, both classification and candidate selection frameworks yield high accuracy and macro F-1 on our dataset. However, the classification approach is notably higher on both measures, even when controlling for the fact that the candidate selection approach cannot predict the *Other* label. Figure 7.10 presents the F-1 score achieved by each model on each label in the dataset; the SVM classifier

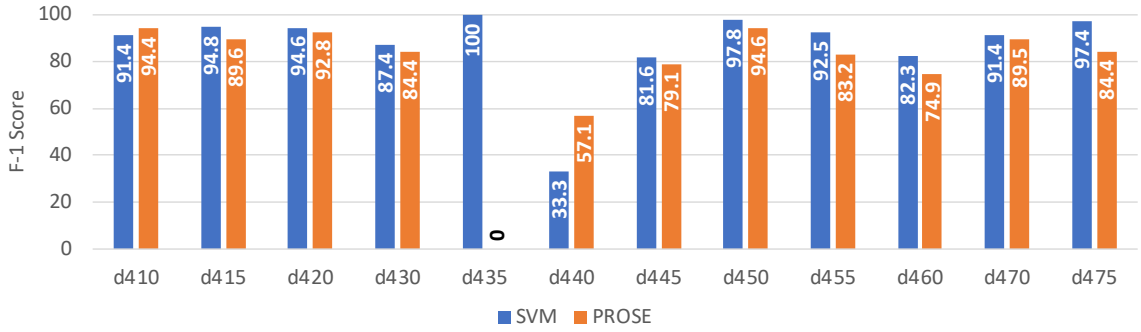


Figure 7.10: Per-label F-1 for classification and candidate selection approaches to mobility activity normalization.

outperforms PROSE in eleven of thirteen labels, though by a relatively small margin. This indicates that while our PROSE model shows clear potential for the action normalization task, and offers advantages of an expandable label set and inter-label relatedness, further experimentation is needed to see if these theoretical advantages play out in a practical setting.

7.4 Conclusions

We have demonstrated that learned representations of domain concepts are a powerful tool for disambiguation and concept normalization, yielding strong performance in open-domain word sense disambiguation benchmarks and providing context-sensitive normalization for ambiguous clinical strings. Our model's ability to combine representations learned from diverse data sources provides a significant advantage over using single representations alone, and suggests that a diversity of specialized representations has potential to improve normalization in many settings. We also show

that concept representations are most effectively used in combination with other normalization methods, in order to target distinct challenges of recognizing standard forms, identifying novel forms, and resolving ambiguity in concept normalization. Our results on identifying mobility activity types indicate clear potential for applying concept representations to the challenging FSI research space, and identify specific directions for further research in generalization to a broader set of activity types. In the next chapter, we present a related application of concept representations: leveraging the linguistic patterns they capture to study concept usage in different domains.

Chapter 8: Analyzing clinical concept usage patterns with sublanguage embeddings

Specialized domains, by their nature, exhibit patterns of language use that are highly distinctive from other focused domains and from broad-coverage samples of language. Capturing and adapting to these endemic patterns, which may reflect metatextual structure such as templates as well as idiomatic usage and genuine semantic differences, is key to successful applications of NLP within restricted domains. Identifying these patterns is one of the primary goals of sublanguage analysis, and has played a pivotal role in the development of NLP for health data, from highlighting the clear linguistic differences between biomedical literature and clinical text (Friedman et al., 2002) to supporting adaptation to multiple languages (Laippala et al., 2009). Recent studies of clinical sublanguage have taken a finer-grained approach and extended sublanguage study to the level of individual document types within an EHR, in order to improve our understanding of the syntactic and lexical differences between highly distinct documents in modern EHR systems (Feldman et al., 2016; Grön et al., 2019).[55]

[55]Portions of this chapter have previously been published in D Newman-Griffis and E Fosler-Lussier. 2019. "Writing habits and telltale neighbors: analyzing clinical concept usage patterns with sublanguage embeddings." *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 146-156.

In order to understand how semantic differences are manifested between specialized domains, whether they be different types of clinical documents or different genres entirely, it is important to be able to examine differences in usage of *concepts* of interest. Concepts are the meat and potatoes of domain semantics, representing the specialized knowledge developed among a group of speakers. As discussed in Chapter 5, concept usage is not straightforward to analyze, as any given concept may have multiple, often non-compositional surface forms that can refer to it (e.g., "ALS" and "Lou Gehrig's disease"), making them difficult to analyze using lexical occurence alone. Understanding how concept usage differs between document types and domains can not only augment recent methods for sublanguage-based text categorization (Feldman et al., 2016), but also inform the perennial challenge of concept normalization (Luo et al., 2019): for example, "depression" is much easier to disambiguate if its occurrence is known to be in a social work note or an abdominal exam.

Inspired by recent technological advances in modeling diachronic language change (Hamilton et al., 2016b; Vashisth et al., 2019), as well as the utility of concept representations for text normalization (described in Chapter 7), we characterize concept usage differences within clinical sublanguages using nearest neighborhood structures of clinical concept embeddings. We show that overlap in nearest neighborhoods can reliably distinguish between document types while controlling for noise in the embedding process. Qualitative analysis of these nearest neighborhoods demonstrates that these distinctions are semantically relevant, highlighting sublanguage-sensitive relationships between specific concepts and between concepts and related surface forms. Our findings suggest that the structure of concept embedding spaces not only captures

domain-specific semantic relationships, but can also identify a "fingerprint" of concept usage patterns within a clinical document type to inform language understanding.

8.1 Related Work

Sublanguage analysis historically focused on describing the characteristic grammatical structures of a particular domain (Friedman, 1986; Grishman, 2001; Friedman et al., 2002). As methods for automated analysis of large-scale data sets have improved, more studies have investigated lexical and semantic characteristics, such as usage patterns of different verbs and semantic categories (Denecke, 2014), as well as more structural information such as document section patterns and syntactic features (Zeng et al., 2011; Temnikova et al., 2014). Using terminologies to assess conceptual features of a sublanguage corpus was proposed by Walker and Amsler (1986), and Drouin (2004); Grön et al. (2019) used sublanguage features to expand existing terminologies, but large-scale characterization of concept usage in sublanguage has remained a challenging question.

Word embedding techniques have been utilized to describe diachronic language change in a number of recent studies, from evaluating broad changes over decades (Hamilton et al., 2016b; Vashisth et al., 2019) to detecting fine-grained shifts in conceptualizations of psychological concepts (Vylomova et al., 2019). Embedding techniques have also been used as a mirror to analyze social biases in language data (Garg et al., 2018). Similar to our work, Ye and Fabbri (2018) investigate document type-specific embeddings from clinical data as a tool for medical language analysis. However, our approach has two significant differences: Ye and Fabbri (2018) used word embeddings only, while we utilize concept embeddings to capture concepts across

multiple surface forms; more importantly, their work investigated multiple document types as a way to *control* for specific usage patterns within sublanguages in order to capture more general term similarity patterns, while our study aims to *capture* these sublanguage-specific usage patterns in order to analyze the representative differences in language use between different expert communities.

8.2 Data and preprocessing

We use free text notes from the MIMIC-III critical care database (Johnson et al., 2016) for our analysis. This includes approximately 2 million text records from hospital admissions of almost 50 thousand patients to the critical care units of Beth Israel Deaconess Medical Center over a 12-year period. Each document belongs to one of 15 document types, listed in Table 8.1.

As sentence segmentation of clinical text is often optimized for specific document types (Griffis et al., 2016), we segmented our documents at linebreaks and tokenized using SpaCy (version 2.1.6; Honnibal and Montani 2017). All tokens were lowercased, but punctuation and deidentifier strings were retained, and no stopwords were removed.

Type	Docs	Lines	Tokens	Matches	Concepts	High Confidence	
						Concepts	Consistency (%)
Case Management	967	20,106	165,608	45,306	557	111	75
Consult	98	15,514	96,515	26,109	812	0	–
Discharge Summary	59,652	14,480,154	104,027,364	30,840,589	6,381	1,599	67
ECG	209,051	1,022,023	7,307,381	2,163,682	540	14	56
Echo	45,794	2,892,069	19,752,879	6,070,772	1,233	157	65
General	8,301	307,330	2,191,618	552,789	2,559	0	–
Nursing	223,586	9,839,274	73,426,426	18,903,892	4,912	2	58
Nursing/Other	822,497	10,839,123	140,164,545	31,135,584	5,049	83	60
Nutrition	9,418	868,102	3,843,963	1,147,918	1,911	198	73
Pharmacy	103	4,887	39,163	8,935	376	0	–
Physician	141,624	26,659,749	148,306,543	39,239,425	5,538	122	57
Radiology	522,279	17,811,429	211,901,548	34,433,338	4,126	599	63
Rehab Services	5,431	585,779	2,936,022	869,485	2,239	9	62
Respiratory	31,739	1,323,495	6,358,924	2,255,725	1,039	5	63
Social Work	2,670	100,124	930,674	195,417	1,282	0	–

Table 8.1: Document type subcorpora in MIMIC-III. Tokenization was performed with SpaCy; Matches and Concepts refer to number of terminology string match instances and number of unique concepts embedded, respectively, using SNOMED-CT and LOINC vocabularies from UMLS 2017AB release. The number of high-confidence concepts identified for each document type is given with their mean consistency.

8.3 Experiments

Methods for learning clinical concept representations have proliferated in recent years (Choi et al., 2016b; Mencia et al., 2016; Phan et al., 2019), but often require annotations in forms such as billing codes or disambiguated concept mentions. These annotations may be supplied by human experts such as medical coders, or by adapting medical NLP tools such as MetaMap (Aronson and Lang, 2010) or cTAKES (Savova et al., 2010) to perform concept recognition (De Vine et al., 2014).

For investigating potentially divergent usage patterns of clinical concepts, these strategies face serious limitations: the full diversity of MIMIC data has not been annotated for concept identifiers, and the statistical biases of trained NLP tools may suppress underlying differences in automatically-recognized concepts. We therefore take a distant supervision approach, using JET (Newman-Griffis et al., 2018). JET uses a sliding context window to jointly train embedding models for words, surface forms, and concepts, using a log-bilinear objective with negative sampling and shared embeddings for context words. It leverages known surface forms from a terminology as a source of distant supervision: each occurrence of any string in the terminology is treated as a weighted training instance for each of the concepts that string can represent. As terminologies are typically many-to-many maps between surface forms and concepts, this generally leads to a unique set of contexts being used to train the embedding of each concept, though any individual context window may be used as a sample for training multiple concepts. We constrain the scope of our analysis to only concepts and strings from SNOMED-CT and LOINC,[56] two popular high-coverage clinical vocabularies.

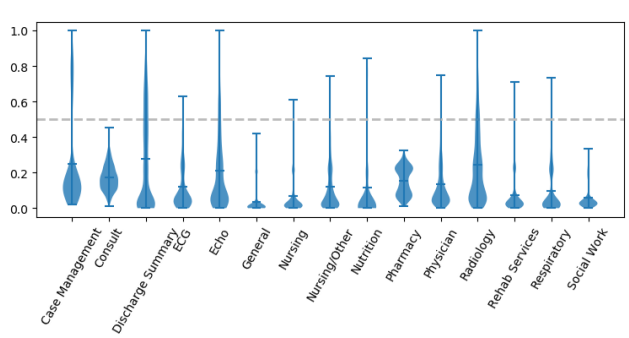
[56]We used the versions distributed in the 2017AB release of the UMLS (Bodenreider, 2004).

8.3.1 Identifying concepts for comparison

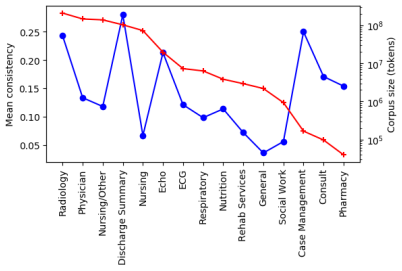
For each document type, we concatenate all of its documents (maintaining line-breaks), identify all occurrences of SNOMED-CT and LOINC strings in each line, and use these occurrences to train word, term, and concept embeddings with JET. Due to the size of our subcorpora, we used a window size of 5, minimum frequency of 5, embedding dimensionality of 100, initial learning rate of 0.05, and 10 iterations over each corpus.

Prior research has noted instability of nearest neighborhoods in multiple embedding methods (Wendlandt et al., 2018). We therefore train 10 sets of embeddings from each of our subcorpora, each using the same hyperparameter settings but a different random seed. We then use all 10 replicates from each subcorpus in our analyses, in order to control for variation in nearest neighborhoods introduced by random initialization and negative sampling. To evaluate the baseline reliability of concept embedding neighborhoods from each subcorpus, we calculated per-concept consistency by measuring, over all pairs of embedding sets within the 10 replicates, the average set membership overlap between the top 5 nearest neighbors by cosine similarity for each concept embedding.[57] As shown in Figure 8.1a, these consistency scores vary widely both within and between document types, with some document types producing no concept embeddings with consistency over 40%. Interestingly, as illustrated in Figure 8.1b, there is no linear relationship between log corpus size and

[57]We chose five nearest neighbors for our analyses based on qualitative review of neighborhoods for concepts within different document types. We found nearest neighborhoods for concept embeddings to vary more than for word embeddings, often introducing noise beyond the top five nearest neighbors; we therefore set a conservative baseline for reliability by focusing on the closest and most stable neighbors. However, using 10 neighbors, as Wendlandt et al. (2018) did, or more could yield different qualitative patterns in document type comparisons and bears exploration.



(a) Self consistency by document type; line at 50% threshold



(b) Self consistency compared to corpus size (log scale), with document types sorted by decreasing corpus size.

Figure 8.1: Self-consistency rates in concept embeddings across MIMIC document types; self-consistency measures overlap in nearest neighbors between replicate embeddings of the same concept.

mean concept consistency ($R^2 \approx 0.011$), suggesting that low consistency is not solely due to limited training data.

To mitigate concerns about the reliability of embeddings for comparison, a set of **high-confidence concepts** is identified for each document type by retaining only those with a self-consistency of at least 50%; Table 8.1 includes the number of high-confidence concepts identified and the mean consistency among this subset.[58] These embeddings capture reliable concept usage information for each document type, and form the basis of our comparative analysis.

[58]We found in our analysis that most concept consistency numbers clustered roughly bimodally, between 0-30% or 60-90%; this is reflected at a coarse level in the overall distributions in Figure 8.1a. Varying the threshold outside of these ranges did not have a significant impact on the number of concepts retained; the 50% threshold was chosen for simplicity. With larger corpora, yielding higher concept coverage, a higher threshold could be chosen for a stricter analysis.

8.3.2 Cross-corpus analysis

Our key question is what concept embeddings reveal about clinical concept usage *between* document types. To maintain a sufficient sample size, we restrict our comparison to the 7 document types with at least 50 high confidence concepts: *Case Management*, *Discharge Summary*, *Echo*, *Nursing/Other*, *Nutrition*, *Physician*, and *Radiology*. *Physician*, *ECG*, and *Nursing* were also used by Feldman et al. (2016) for their lexicosyntactic analysis, although they combined *Nursing* documents (longer narratives) and *Nursing/Other* (which tend to be much shorter) into a single set, while we retain the distinction. Interestingly, the fourth type they analyzed, *ECG*, produced only 14 high-confidence concepts in our analysis, suggesting high semantic variability despite the large number of documents.

As learned concept sets differ between document types, the first step for comparing a document type pair is to identify the set of concepts embedded for both. For reference type A and comparison type B, we identify high-confidence concepts from A that are also present in B, and calculate four distributions using this shared set:

Reference consistency: self-consistency across each of the shared concepts, using only other shared concepts to identify nearest neighborhoods in embeddings for the reference set.

Comparison consistency: self-consistency of each shared concept in embeddings for the comparison document type, again using only shared concepts for neighbors. As the shared set is based on high-confidence concepts from the reference set, this is not symmetric with reference consistency (as the high-confidence sets may differ).

Cross-type consistency: average consistency for each shared concept calculated over all pairs of replicates (i.e., comparing the nearest neighbors of all 10 reference embedding sets to the nearest neighbors in all 10 comparison embedding sets).

Consistency deltas: the difference, for each shared concept, between its reference self-consistency and its cross-type consistency. This provides a direct evaluation of how distinct the concept usage is between two document types, where a high delta indicates highly distinct usage.

Mean values for these distributions are provided for each pair of our 7 document types in Figure 8.2. Comparing Figures 8.2b and 8.2c, it is clear that high-confidence concepts for one document type are typically not high-confidence for another. Most document type pairs show fairly strong divergence, with deltas ranging from 30-60%. *Physician* notes have comparatively high cross-set consistency of around 20% for their high-confidence concepts, likely reflecting the all-purpose nature of these documents, which include patient history, medications, vitals, and detailed examination notes. Interestingly, *Case Management* and *Nutrition* are starkly divergent from other document types, with near-zero cross-set consistency and comparatively high self-consistency of over 70% in the compared concept sets, despite a relatively high overlap between their high-confidence sets and concepts learned for other document types.

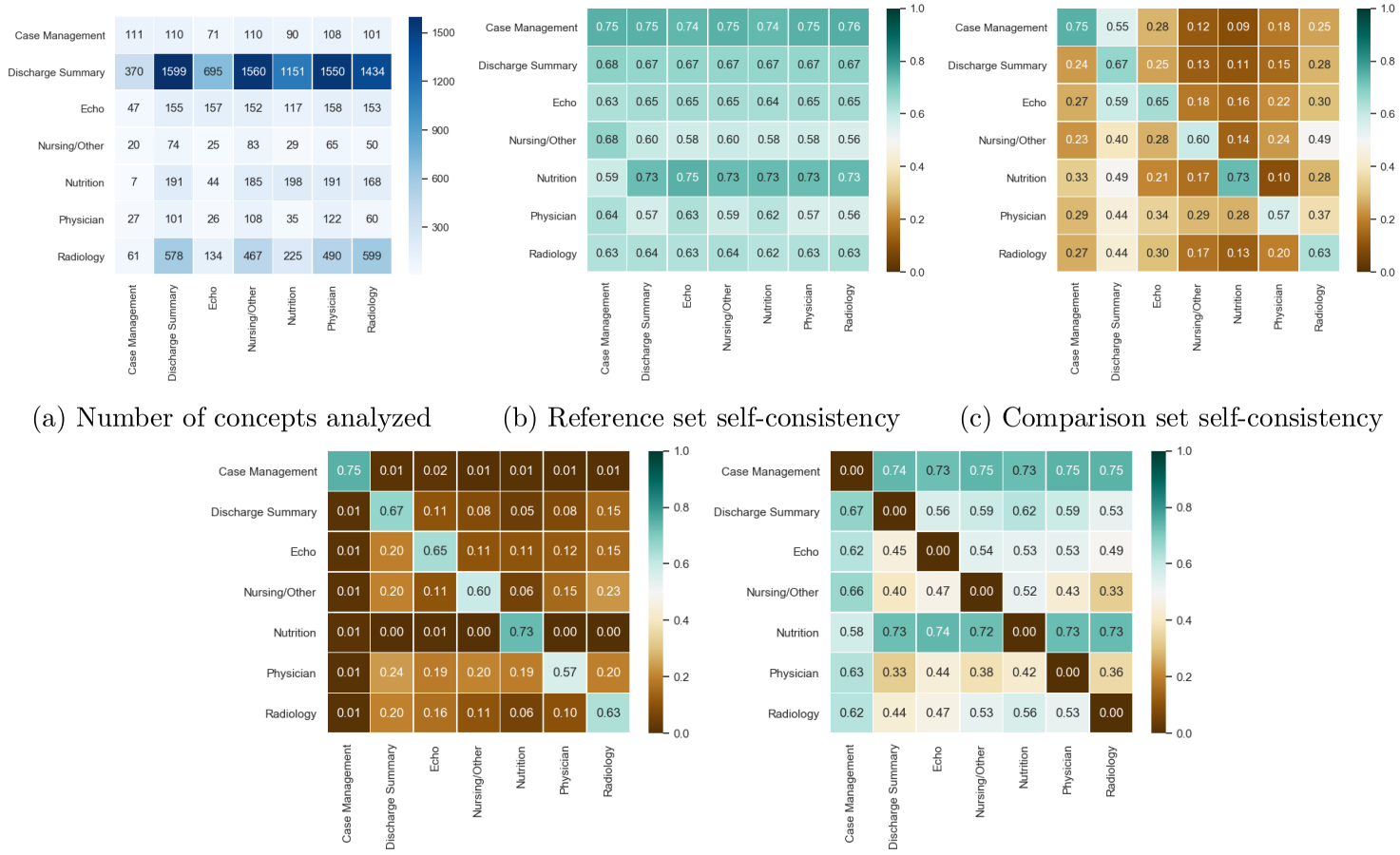
In order to control for the low overlap between high-confidence sets in different document types, we also re-ran our consistency analysis restricted to only concepts that are high-confidence in *both* the reference and comparison sets. As shown in Figure 8.3, this yields considerably smaller concept sets for comparison, with single-digit overlap for 18/42 non-self pairings. Cross-set consistency increases somewhat, most

significantly for pairings involving *Physician* or *Radiology*; however, no consistency delta falls below 20% for any non-self pair, indicating that concept neighborhoods remain distinct even within high-confidence sets.

8.3.3 Qualitative neighborhood analysis

Analysis of neighborhood consistency enables measuring divergence in the contextual usage patterns of clinical concepts; however, this divergence could be due to spurious or semantically uninformative correlations instead of clinically-relevant distinctions in concept similarities. To confirm that our methodology captures informative distinctions in concept usage, we qualitatively review example neighborhoods. To mitigate variability of nearest neighborhoods in embedding spaces, we identify a concept's *qualitative* nearest neighbors for a given document type by calculating its pairwise cosine distance vectors for all 10 replicates in that document type and taking the k neighbors with lowest average distance.

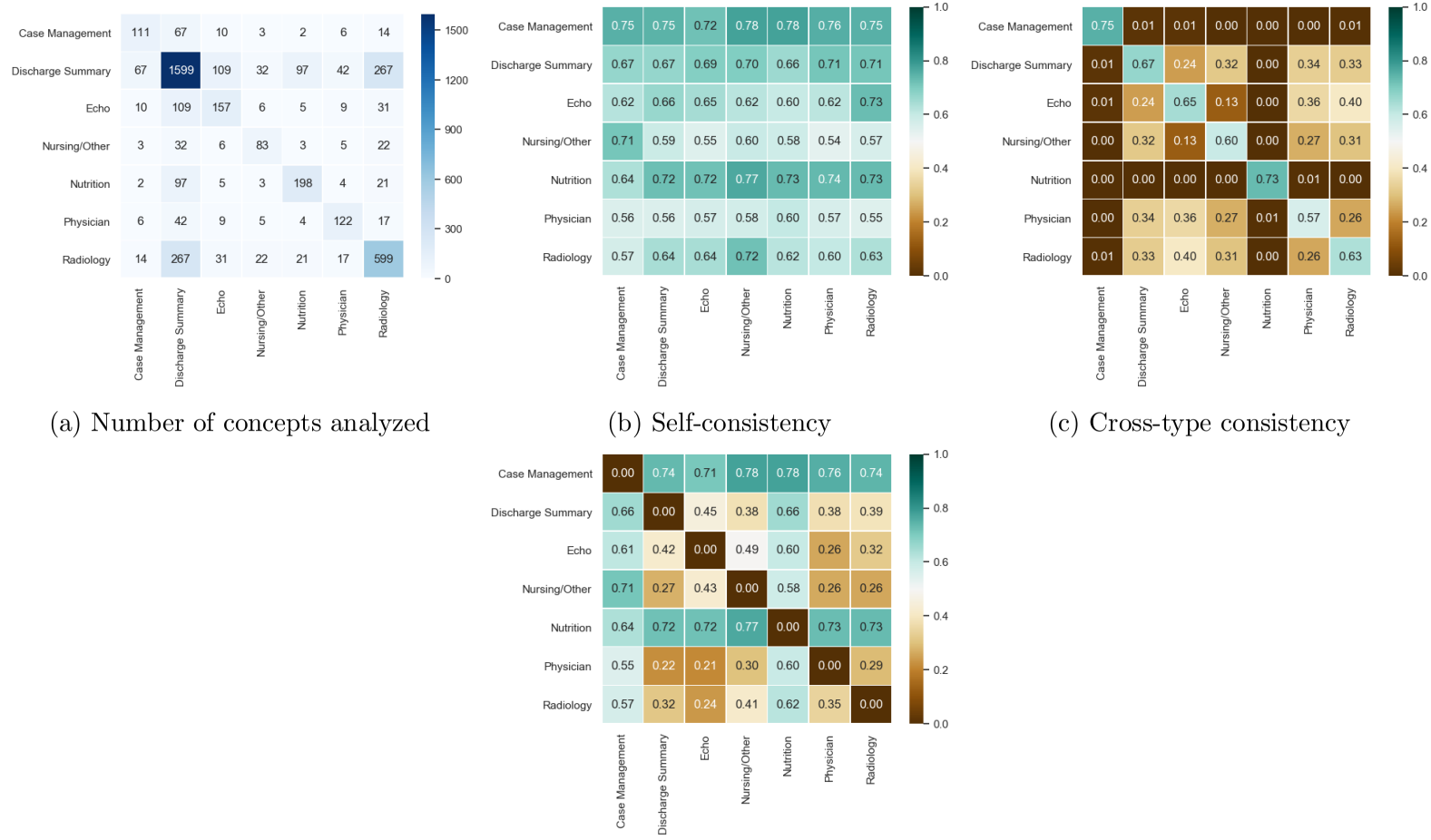
As with our consistency analyses, we focus on the neighborhoods of high-confidence concepts, although we do not filter the neighborhoods themselves. Of all high-confidence concepts identified in our embeddings, only two were high-confidence in 5 different document types, and these were highly generic concepts: C0184661 *Interventional procedure* and a corresponding LOINC code (C0945766). Seven concepts were high-confidence for 4 document types; of these, two were generic procedure concepts, two were concepts for the broad gastrointestinal category, and three were versions of body weight. For a diversity of concepts, we therefore turned to the 75 concepts that were high-confidence within 3 document types. We reviewed each of these concepts, and describe our findings for three of the most broadly clinically-relevant below.



(a) Number of concepts analyzed (b) Reference set self-consistency (c) Comparison set self-consistency

(d) Cross-type consistency (e) Consistency deltas

Figure 8.2: Comparison of concept neighborhood consistency statistics across document types, using high-confidence concepts from the reference type. Figure 8.2a provides the number of concepts shared between the high-confidence reference set and the comparison set. All values are the mean of the consistency distribution calculated over all concepts analyzed for the document type pair.



(a) Number of concepts analyzed

(b) Self-consistency

(c) Cross-type consistency

(d) Consistency deltas

Figure 8.3: Super high-confidence concept neighborhood consistency statistics across document types. Super high-confidence concepts are high-confidence in both reference and comparison sets. In this case, reference self-consistency and target self-consistency are symmetric, so only reference self-consistency is presented in Figure 8.3b.

C0011849 Diabetes Mellitus *Diabetes Mellitus* (search strings: "diabetes mellitus" and "diabetes mellitus dm") was high-confidence in *Discharge Summary*, *Nursing/Other*, and *Radiology* document types; Table 8.2 gives the top 5 neighbors from each type. These neighbors are semantically consistent across document types: more specific diabetes-related concepts, related biological factors; continuing down the nearest neighbors list yields related symptoms and comorbidities such as C0022104 *Irritable Bowel Syndrome* and C0017168 *Gastroesophageal reflux disease.*

C0751295 Memory loss *Memory loss* (search string: "memory loss") was also high-confidence in *Discharge Summary*, *Nursing/Other*, and *Radiology* documents. For brevity, its nearest neighbors are omitted from Table 8.2, as there is little variation among the top 5. However, the next neighbors (at only slightly greater cosine distance) vary considerably across document types, while remaining highly consistent within each individual type. In *Discharge Summary*, more high-level concepts related to overall function emerge, such as C0598463 *Functional status*, C0439849 *Relationships*, and C4068735 *Rambling.* *Radiology* yields more symptomatically-related neighbors: C0221470 *Aphagia* is present in both, but *Radiology* includes C0233407 *Disorientation*, C0011253 *Delusions*, and C0231686 *Gait, Unsteady.* Finally, *Nursing/Other* finds concepts more related to daily life, such as C0678446 *Cigars* and C3843228 *Multifocals*, though at a greater cosine distance than the other document types (Figure 8.4).

C0278060 Mental state *Mental state* (search strings: "mental status", "mental state") was high-confidence in *Discharge Summary*, *Echo*, and *Radiology*, and highlighted an unexpected consequence of relying on the Distributional Hypothesis (Harris, 1954) for semantic characterization in sublanguage-specific corpora.

Query	Discharge Summary	Nursing/Other	Radiology
Diabetes Mellitus (C0011849)	C0011847 *Diabetes*	C0085207 *Gestational Diabetes*	C3853134 *Poorly controlled*
	C0441730 *Type 2*	C1443036 *A2 immunologic symbol*	C0021641 *Insulin*
	C0441729 *Type 1*	C0011854 *Diabetes Mellitus, Insulin-Dependent*	C0011854 *Diabetes Mellitus, Insulin-Dependent*
	C0085207 *Gestational Diabetes*	C0015498 *Factor V*	C0011860 *Diabetes Mellitus, Non-Insulin-Dependent*
	C0011854 *Diabetes Mellitus, Insulin-Dependent*	C1443035 *A1 immunologic symbol*	C0441777 *Stage level 5*
	Discharge Summary	Echo	Radiology
Mental state (C0278060)†	C4068804 *Coherent*	C3263710 *Donor:Type:Point in time:^Patient:Nominal*	C0856054 *Mental status changes*
	C0009676 *Confusion*	C0013018 *Donor person*	C0278061 *Abnormal mental state*
	C2598168 *Respiratory status:-:Point in time:^Patient:-*	C0162297 *Respiratory arrest*	C0234425 *Level of consciousness*
	C1998827 *Respiratory status*	C1716004 *Organ donor:Type:Point in time:^Donor:Nominal*	C4050479 *Level of consciousness:Find:Pt:^Patient:Ord*
	C0278061 *Abnormal mental state*	C4281783 *Swallowing G-code*	C0026221 *Mississippi (state)*

Table 8.2: Examples of concept-level nearest neighbors across document types. Shown are 5 nearest neighbor concepts to *Diabetes Mellitus* and *Mental state* from 3 high-confidence document types, averaging cosine similarities across all replicate embedding sets within each document type. †The two nearest neighbors to *Mental state* for all three document types were two LOINC codes using the same "mental status" string; they are omitted here for brevity.

The top 5 nearest neighbors (excluding two trivial LOINC codes for the same concept, also using the "mental status" search string) are given in Table 8.2. In *Discharge Summary* documents, "mental status" is typically referred to in detailed patient narratives, medication lists, and the like, and this yields semantically-reasonable nearest neighbors such as C0009676 *Confusion* and C4068804 *Coherent*.

In *Echo* documents, however, "mental status" occurs most frequently within an "Indication" field of the "PATIENT/TEST INFORMATION" section. Two common patterns emerge in "Indication" texts: references to altered or reduced mental status, or patients who are vegetative and being evaluated for organ donor eligibility. Though "mental status" and "organ donor" do not co-occur, their consistent occurrence in the same contextual structures leads to extremely similar embeddings (see Figure 8.4). A similar issue occurs in *Radiology* notes, where the "MEDICAL CONDITION" section includes several instances of elderly patients presenting with either hypothermia or altered mental status; as a result, two hypothermia concepts (C1963170 and C0020672) are in the 10 nearest neighbors to *Mental state*.

Results from *Radiology* also highlight one limitation of distant supervision for learning concept embeddings: as the word "state" is polysemous, including a geopolitical entity, geographical concepts such as C0026221 *Mississippi* end up with similar embeddings to *Mental state*. A similar issue occurs in the neighbors for *Memory loss*; due to string polysemy, the concept C4255278 *CIGAR string - sequence alignment* ends up with a similar embedding to C0678446 *Cigars*.

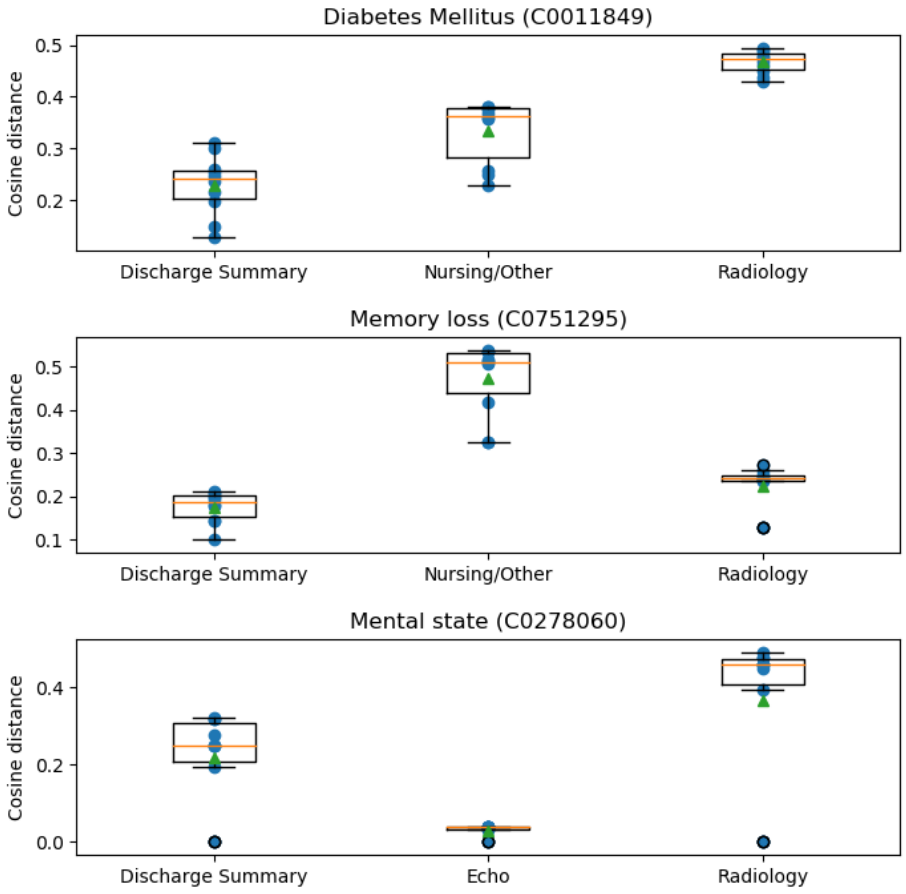


Figure 8.4: Cosine distance distribution of three concept representations to their 10 nearest neighbors, averaged across document type replicate embeddings.

8.3.4 Nearest surface form embeddings

As JET learns embeddings of concepts and their surface forms jointly in a single vector space, we also analyzed the surface forms embeddings nearest to different concepts. This enabled us both to evaluate the semantic congruence of surface form and concept embeddings, and to further delve into corpus-specific contextual patterns that emerge in the vector space. As with our concept neighborhood analysis, for each of our 10 replicate embeddings in each document type, we calculated the cosine

distance vector from each high-confidence concept to all of the term embeddings in the same replicate, and then averaged these distance vectors to identify neighbors robust to embedding noise. Table 8.3 presents surface form neighbors identified for three high-confidence clinical concepts chosen for clinical relevance and wide usage; these concepts are discussed in the following paragraphs.

Blood pressure (C0005823) *Blood pressure* is high-confidence in *Discharge Summary*, *Echo*, and *Radiology* documents. It is a key concept that is measured frequently in various settings; intuitively, it is a sufficiently core concept that it should exhibit little variance. Its neighbor surface forms indeed indicate fairly consistent use across the three document types, referencing both related measurements ("heart rate") and related concepts ("exercise" and "stress").

C0013798 Echocardiogram *Echocardiogram* is high-confidence in *Discharge Summary*, *Echo* (detailed summaries and interpretation written after the ECG), and *ECG* (technical notes taken during the procedure) documents. ECGs are common, and are performed for various purposes and discussed in varying detail. Interestingly, neighbor surface forms in *Discharge Summary* embeddings reflect specific pathologies, potentially capturing details determined post diagnosis and treatment. In *Echo*

Query	Discharge Summary	Nutrition	Case Management
C0009462 *Community*	Community	Dilute	Substance
	Health center	Social work	Monitoring
	Acquired	Surgical site	Somewhat
	Residence	In situ	Hearing
	Nursing facility	Nephritis	Speech
	Discharge Summary	Echo	ECG
C0013798 *ECG*	ECG	ECG	ECG
	EKG	Exercise	Physician
	Sinus tachycardia	Stress	Last
	Sinus bradycardia	Fair	No change
	Right bundle branch block	Specific	Abnormal
	Discharge Summary	Echo	Radiology
C0005823 *Blood pressure*	Blood pressure	Blood pressure	Blood pressure
	Heart rate	Heart rate	Heart rate
	Pressure	Rate	Rate
	Systolic blood pressure	Exercise	Method
	Rate	Stress	Exercise

Table 8.3: Examples of surface form-level nearest neighbors across document types, given for three frequent clinical concepts. Document types shown are those for which each query CUI is high-confidence.

embeddings, the neighbors are more general surface forms evaluating the findings ("fair") and relevant history/symptoms that led to the ECG ("exercise", "stress"). *ECG* embeddings reflect their more technical nature, with surface forms such as "no change" and "abnormal" yielding high similarity.

C0009462 Community *Community* is a very broad concept and a common word, and is discussed primarily in documents concerned with whole-person health; it is high confidence in *Discharge Summary*, *Nutrition*, and *Case Management* documents. Each of these document types reflects different usage patterns. The nearest surface forms in *Discharge Summary* embeddings reflect a focus on living conditions, referring to "health center", "residence", and "nursing facility". In *Nutrition* documents, *Community* is discussed primarily in terms of "community-acquired pneumonia", likely leading to more treatment-oriented neighbor surface forms. Finally, in *Case Management* embeddings, nearby surface forms reflect discussion of specific risk factors or resources ("substance", "monitoring") to consider in maintaining the patient's health and responding to their specific needs (e.g., "hearing", "speech"). Thus, *Community* reflects two distinct kinds of concept usage patterns within document type sublanguages: templating, by its association with other treatment-related terms in *Nutrition* documents; and focus on specific aspects of the concept, by its associations with living conditions in *Discharge Summary* notes and with risk factors in *Case Management* notes. These two factors are entangled in our results: learning to disentangle such different aspects of sublanguage patterns from cooccurrence information represents an intriguing direction for future research.

8.4 Discussion

Our results show that learning concept embeddings from focused clinical corpora captures distinctive characteristics of those corpora. These characteristics include both semantic differences, in terms of salient associations and conceptualizations of a given concept, and structural differences, such as template-based usage. These findings suggest that sublanguage-specific embeddings can help profile distinctive usage patterns for text categorization, offering greater specificity than latent topic distributions while not relying on potentially brittle lexical features. In addition, usage profiles within specific settings could also assist with concept normalization by providing more-informed prior probability distributions for medical vocabulary senses that are conditioned on the document or section type that they occur in.

8.4.1 Detecting deviation from baseline usage

Our experiments in this study were formulated in terms of obtaining representative snapshots of language use from different document types independently, and identifying differences that emerged. An alternative formulation, which might also mitigate our observed sensitivity to embedding noise somewhat, would be to capture deviation from a shared reference instead: i.e., to train representations on a broad-coverage baseline language sample that can be expected to inform all of the subdomains of interest, and then tune these representations on subdomain-specific corpora in order to focus in on the characteristics of each particular domain. This would then enable a more straightforward way of detecting which concepts are distinctive for a given domain, by comparing their nearest neighbor associations with the nearest neighborhoods from the reference representations; those which exhibit

consistent differences in nearest neighbors are likely to be concepts with specialized usage in the domain of interest. The shared reference would further strengthen comparative analysis between two specialized domains, by controlling for some degree of embedding noise and providing a shared base of more general-purpose language use to inform the specialized models. We highlight this as a key direction for future work.

8.4.2 Disentangling corpus features from sublanguage features

As we observed with C0278060 *Mental state*, relying on similarity in contextual patterns can lead to capturing more corpus-specific features with embeddings, as opposed to (sub)language-specific features, as target corpora become smaller and more homogeneous. The same issue emerged in analysis of C0009462 *Community*, in which some samples captured template-based information while others captured semantic associations. If a particular concept or set of concepts are always used within the same section of a document, or in the same set phrasing, the "similarity" captured by organization of an embedding space will be more informed by this writing habit endemic to the specific corpus than by clinically-informed semantic patterns that can generalize to other corpora.[59] While representation learning based on contextual information must inherently conflate these two factors, one possible direction for mitigating this entanglement is to leverage the complementarity highlighted in Chapter 7 and learn concept representations from multiple knowledge sources jointly. Nonetheless, drawing a distinction between idiosyncrasies of a particular corpus and representative characteristics of an underlying sublanguage is a central concern to analysis

[59]For further discussion, see Newman-Griffis and Fosler-Lussier (2019b).

of specialized language samples, and one which is beyond the scope of concept-level semantics alone.

8.4.3 Limitations

A few limitations of our study are important to note. The embedding method we chose offers flexibility to work with arbitrary corpora and vocabularies, but its use of distant supervision introduces some undesirable noise. The example given in Section 8.3.3 of the similar embeddings learned for the concept *cigars* and the concept of the CIGAR string in genomic sequence editing illustrates the downside of not leveraging disambiguation techniques to filter out noisy matches. On the other hand, our restriction to strings from SNOMED-CT and LOINC provided a high-quality set of strings intended for clinical use, but also removed many potentially helpful strings from consideration. For example, the UMLS also includes the non-SNOMED/LOINC strings "diabetes" and "diabete mellitus" [*sic*] for C0011849 *Diabetes Mellitus*, both of which occur frequently in MIMIC data. Misspellings are also common in clinical data; leveraging well-developed technologies for clinical spelling correction would likely increase the coverage and confidence of sublanguage concept embeddings.

At the same time, the low volume of data analyzed in many document types introduces its own challenges for the learning process. First, though JET can in principle learn embeddings for every concept in a given terminology, this is predicated on the relevant surface forms appearing with sufficient frequency. For a small document sample, many such surface forms that would otherwise be present in a larger sample will either be missing entirely or insufficiently frequent, leading to effectively "missed" concepts. While we are not aware of another concept embedding method

compatible with arbitrary unannotated corpora that could help avoid these issues, some strategies could be used to reduce the potential impact of both training noise and low sample sizes. One approach referenced above, which might also help improve concept consistency in the document types that yielded few or no high-confidence concepts, would be pretraining a shared base embedding on a large corpus such as PubMed abstracts, which could then be tuned on each document type-specific subcorpus. While this could introduce its own noise in terms of the differences between biomedical literature language and clinical language (Friedman et al., 2002), it could help control for some degree of sampling error and provide a linguistically-motivated initialization for the concept embedding models.

8.5 Conclusion

Analyzing nearest neighborhoods in embedding spaces has become a powerful tool in studying diachronic language change. We have described how the same principles can be applied to sublanguage analysis, and demonstrated that the structure of concept embedding spaces captures distinctive and relevant semantic characteristics of different clinical document types. This offers a valuable tool for sublanguage characterization, and a promising avenue for developing document type "fingerprints" for text categorization and knowledge-based concept normalization.

Chapter 9: Final Remarks

Understanding the characteristics of language within new domains, and linguistic factors for new applications, has been a key contributor to the growth of the natural language processing field. Neural representation learning techniques have provided a powerful tool for capturing patterns in natural language, enabling mathematical representation of words, phrases, and other linguistic units in dense, low-dimensional feature spaces for modeling language phenomena.

Representation learning methods draw on the distributional hypothesis (Harris, 1954), which states that words used in similar contexts have similar meaning, to produce representation spaces where vector similarity correlates with similarity in usage patterns. Representation models are learned based only on the relationships between represented items, rather than explicit features, making them highly opaque and difficult to analyze and interpret. Nonetheless, the relationships they capture provide a mirror for analyzing language use, and learning more about specialized language and how to process it in specific domains.

In the first part of this thesis, we described Functional Status Information (FSI), a type of health-related information capturing an individual's lived experience in a particular health state, with utility for both care delivery and government benefits administration. We further described characteristics of clinical text that pose significant

challenges for NLP. In the second part, we briefly reviewed the field of representation learning, and discussed how learned representations can be used to study questions about language use. We then presented a new technique for learning representations of domain-specific concepts, and demonstrated its utility for both biomedical and general-purpose applications. Finally, in the third part, we described several applications of representation learning techniques to challenging tasks in processing clinical language and FSI in particular, including automatic extraction of complex FSI reports, semantic grounding of ambiguous and/or domain-specific terms, and studying patterns in how medical concepts are discussed among different clinical specialties.

9.1 Summary of contributions

Characterization of the functional status domain *(Chapter 2)* We described how conceptual models of human function can be realized in natural language, and identified specific challenges in resources, modeling techniques, and domain knowledge required to effectively leverage informatics methods to analyze function. We further demonstrated the rehabilitation medicine, a family of medical specialties focusing on optimizing function, forms a distinct clinical sublanguage, and that functional status information presents significant challenges of semantic and syntactic complexity.

Method for jointly embedding entities and text *(Chapter 5)* We developed JET, a method for learning neural representations of domain concepts from arbitrary text corpora, without the need for direct annotations of concept mentions or any specialized domain knowledge beyond a flat terminology. We demonstrated that concept representations learned through JET correlate with human judgments of similarity and relatedness for both biomedical and encyclopedic concepts, and that

these representations capture semantically-important information beyond what word-level representations can model.

Applications of learned representations to capture FSI *(Chapter 6)* We presented two different models for extracting functional status information using learned representation features, and demonstrated that representations learned from in-domain data contribute significantly to successful extraction. We showed that a sequence-level LSTM-CRF model yields high precision extraction of mobility-focused FSI activity reports, while a word-level relevance tagging model achieves high coverage of mobility-related information in real-world clinical documents from both NIH and the U.S. Social Security Administration.

Model combining concept representations for semantic grounding of text *(Chapter 7)* We proposed PROSE, a model for learning task-specific projections of representation spaces to combine concept representations learned from different knowledge sources. We evaluated our model on two semantic grounding tasks: word sense disambiguation and medical concept normalization, using both clinical disease/treatment/test concepts and functional activity types. We demonstrated that a learned, context-sensitive projection of concept representations improves semantic grounding over vector space comparison alone, and we showed consistent improvement from combining multiple sources and methods of learning concept representations.

Demonstration of sublanguage analysis with concept representations *(Chapter 8)* Finally, we used JET to learn representations of clinical concepts from clinical text corpora representing different medical specialties, and demonstrated

that the nearest neighborhood structure of these representations captured clinically-relevant distinctions in how medical concepts were discussed between different document types, thus laying the groundwork for including concept-based representation learning in the toolkit for medical sublanguage analysis.

9.2 Future directions

Functional status information presents an attractive test bed for further developing and applying representation-based sublanguage analysis techniques, particularly as NLP for FSI expands from mobility information to other types of activity and participation. The HARE model described in Chapter 6 presents an opportunity to automatically identify FSI-related documents from other, unannotated corpora (e.g., MIMIC-III), and the work described in Chapter 8 offers a tool for asking more fine-grained questions about differences in how activity and participation concepts are used across specialties or between different institutions. In addition, learned representations can significantly contribute to efforts to automatically develop comprehensive terminologies for how FSI concepts are referred to in practice, enabling improved application of traditional clinical NLP methods in this area.

In addition to representation-based methods for analysis, expert-driven linguistic analysis of FSI will identify new strategies for model development and new research questions in how to capture this information most effectively. For example, given our observations of complex syntactic structure in activity reports, it is quite likely that analysis of syntactic dependencies within these reports, as well as semantic roles, will identify some generalizable patterns in how activity reports are constructed that can be incorporated into the design of models for FSI extraction.

In a more general modeling direction, the clear gains we demonstrate for semantic grounding tasks by combining multiple representation methods open intriguing questions for use of representation features in other tasks. Different kinds of knowledge that contribute to human language understanding within specific domains, such as the relationships between concepts and knowledge of language use patterns within local communities. This raises the question of whether a PROSE-style combination model, utilizing word or phrase representations learned from different knowledge sources, could improve performance in information extraction and text classification as well. Further, these results suggest that there are likely engineering gains to be had from learning and combining multiple highly specialized methods for concept representation, in contrast to an iterative improvement of a single best all-purpose concept embedding strategy.

Finally, as we noted in Chapter 4, one of the significant challenges in analyzing NLP models using representation features is the opacity of the representation space, and the difficulty of mapping operations in representation space back to corresponding intuitions about language. The arguments laid out in Appendix A present a possible approach towards utilizing contextualized word and sentence representations to provide a more systematic mapping back from representation space to language, which would be a powerful tool for model interpretation. Developing a practical method to approximate mapping back from arbitrary representations to meaningful language using contextualized models is an intriguing area for future investigation.

9.3 Conclusions

In this thesis, we have investigated the use of representation learning techniques at the word and concept levels to capture patterns of language use in specific domains. We further identified functional status information as a new, high-impact domain for natural language processing, and provided both descriptive analysis of the domain and empirical evidence of capturing domain information with learned representations. This work provides new techniques and clear directions for using learned representations to study questions about language, and provides a template for applying NLP to new questions and new types of information in the clinical domain.

Appendix A: Sequential representations: a homeomorphism for language?

Returning to the discussion in Section 4.1, sequence-level representations, including contextualized word representations, offer an intriguing possibility for embedding (in the mathematical sense) natural language. As discussed in Section 4.4, the semantic content of linguistic representation spaces are typically analyzed in topological terms: i.e., which representations are close to one another, what groups do they form, and what relationships exist between different groups of representations. As Section 4.5 shows, these topological characteristics also provide the signals most informative for backpropagation in machine learning, reflecting the continuous nature of most neural network transformations. Thus, it is informative to view the embedding function modeled in neural representations through a topological lens, i.e., as a homeomorphism.

A.1 A homeomorphism provides interpretability of representation space

Representing language in real space for input features (and intermediate features) in NLP applications presents the opportunity to treat these features as samples in

a continuous Euclidean space. This raises three intriguing questions regarding the relationship *back* from the feature space to the language space:

1. What does the space *between* two embeddings correspond to?

2. Does a task-specific decision function over representation space correlate with linguistic intuitions?

3. Can we define operators over the embedding space that correspond to linguistic intuitions?

For all of these questions, a homeomorphism between language and representations is a critical tool, because it provides the connection between points in the real-valued representation space and their pre-image in the domain (i.e., a word sequence).

A.1.1 Interpreting the space between embeddings

Prior work has observed semantically-correlated clustering of word and sentence representations (Xu et al., 2015; Zhai et al., 2016), and Kim and de Marneffe (2013) demonstrated that linear alignment of word representations corresponds to semantic intensity scales, and that strengthening this alignment yields improved performance on semantic tasks (Kim et al., 2016a). These results, which reflect the distributional hypothesis, suggest that continuous displacement in the representation space is likely to correspond to similarity in language. However, analyses of continuous displacement, such as the analogy completion task, utilize the Voronoi tesselation of the representation space to choose the closest candidate from the fixed vocabulary of words (Linzen, 2016). Continuous mapping back from representation space to vocabulary is impossible with non-contextualized word representations (discussed

below), necessitating this type of discontinuous decision. However, a homeomorphism to word sequences would allow for direct analysis of continuous representation displacement in terms of corresponding language, enabling analyses such as interpolation along an identified semantic intensity hyperplane or the types of errors produced by a continuous function to choose from a discrete vocabulary.

A.1.2 Interpreting decision functions in representation space

Sequence- or word-level classification models using representational features define decision functions over representation space. However, unlike engineered features, which have a mathematical value directly interpretable in light of the criteria for the feature, the meaning of representational features comes only from their correlation to linguistic inputs. Thus, interpretation of the boundary areas in learned decision functions is limited to points whose representations are known *a priori*, which may or may not be close to the boundary area. A homeomorphism would provide a direct mapping from representation points arbitrarily close to a decision boundary back to interpretable linguistic inputs, allowing for much more fine-grained analysis of whether learned decision functions correspond to linguistic intuitions or not.

A.1.3 Linguistic operators in real space

Operator functions in an embedding scenario can be defined in two ways: over the domain (yielding corresponding transformations in the image) or over the range (with corresponding transformations in the pre-image). Implementing semantic and syntactic composition of word sequences in representation space has been an area of active research (Erk and Padó, 2008; Mitchell and Lapata, 2010; Baroni and Zamparelli,

2010; Blacoe and Lapata, 2012; Fyshe et al., 2015), in which contextualized representation models are a recent entry. A homeomorphism would be a powerful new tool for this problem, enabling both linguistic analysis of Euclidean operators–e.g., investigating what multiplying two representations means in language (if anything)–and a direct analysis of the continuous outputs of custom composition functions.

A.2 Well-chosen sequential representations are homeomorphic to word sequences: proof sketch

A.2.1 Cardinality of domain and range

Defining a homeomorphism requires meeting stricter criteria than the general notion of an embedding function. For $f : X \to Y$ to be a homeomorphism, both f and its inverse must be continuous, and f must be bijective. With word-level static and sub-word representations, however, bijectivity is impossible. Natural language vocabularies are productive: neologisms emerge constantly, thus vocabularies cannot be said to be finite. However, it is reasonable to assume that any natural language vocabulary must be *countably* infinite ($\aleph_0$); it is easy to imagine an infinite dictionary mapping integers to word types, with complete coverage. The range of neural representation functions, however, is the d-dimensional real domain $\mathbb{R}^d$, which is *uncountably* infinite ($\aleph_1$). With a different cardinality between domain and range, f inherently cannot be bijective.

Sequence representation methods, however, define the domain X differently. As the representation of any sequence (or sequence element) is conditioned on the sequence of words around it, different word sequences will yield different representations; in a contextualized model, different representations will be provided for each token. The recursive nature of natural language grammars, through features such

as embedded clauses and conjunction, allow in principle for infinite-length sentences (though these are of course impossible in practice). Infinite length becomes more reasonable if we consider processing word sequences across sentence boundaries, as is frequently done in contextualized embedding models (Devlin et al., 2019; Yang et al., 2019). While some contextualized models, such as BERT, impose length limits on input sequences for processing efficiency, these limits are not required (ELMo, for example, processes arbitrary-length sequences). Thus, the space of potential inputs for sequential models is in fact the *power set* of a natural language vocabulary–and therefore uncountably infinite ($\aleph_1$). With both a domain and range of cardinality $\aleph_1$, a bijective mapping becomes possible.

A.2.2 Continuity of the representation function

Continuity between two topological spaces requires that for every open set in the range, its pre-image is an open set in the domain. If we define the domain X as the set of all unique word sequences (of length 0 to inf), then it follows that the basic open sets we want to represent in the embedding are simply each individual word sequence $x \in X$, yielding the discrete topology (i.e., every combination of zero or more word sequences is an open set). The topology of the range is therefore immaterial, as any function $f : X \to T$ mapping from a discrete topology X is inherently continuous (since all pre-images will necessarily be open sets) (Munkres, 2013). However, for practical purposes, let the topology of the range also be the discrete topology, in d-dimensional real space, generalizing the criterion that if one point is taken to correspond to a specific input sequence, all points in real space should have corresponding input sequences.

A.2.3 Meeting the bijectivity criterion

Bijectivity involves satisfying both injectivity (each item in the domain is mapped to a unique item in the range) and surjectivity (each item in the range has a corresponding pre-image in the domain).

Injectivity requires three constraints on the contextualized representation function. First, the sequence representation model must consist of a composition of injective functions; i.e., the activation functions in the neural network must be injective (sigmoid and tanh activation satisfy this requirement; the commonly-used Rectified Linear Unit (ReLU) does not). Second, the lexicalized representations used to represent each word (or wordpiece) as inputs to the sequence representation method must be unique; i.e., no two words or wordpieces may share the same lexicalized representation. These first two constraints ensure uniqueness of outputs for any sequences of the same length, i.e.

$$\forall 1 \leq t \leq \inf; x, x' \in X; x \neq x' : f(x_1 \dots x_t) \neq f(x'_1 \dots x'_t) \tag{A.1}$$

Finally, a non-zero amount of information must be carried over in the model between timesteps; i.e.

$$\forall 2 \leq t \leq \inf; x \in X : f(x_1 \dots x_{t-1}) \neq f(x_1 \dots x_t) \tag{A.2}$$

If these constraints are satisfied, it follows that

$$\forall x, x' \in X; x \neq x' : f(x) \neq f(x') \tag{A.3}$$

i.e., f is injective.

Surjectivity is somewhat simpler. As any open interval on the real line is homeomorphic to the entire real line (Munkres, 2013), it is sufficient to constrain the range

of f to an open interval:

$$\exists a, b \in \mathbb{R}^d : \forall x \in X, 1 \leq i \leq d : a < f(x)_i < b \tag{A.4}$$

For any f satisfying this criterion, a new function $f' : X \to Y$ can be trivially defined as $f'(x) = g(f(x))$, where $g : (a, b)^d \to \mathbb{R}^d$; i.e., f' is a surjective version of f. For practical purposes, the sequence representation model is therefore required to have an output activation function h whose range is an open set in the real line, such as sigmoid ($h : \mathbb{R} \to (0, 1)$) or tanh ($h : \mathbb{R} \to (-1, 1)$).

Thus, if the sequence representation function $f : X \to Y$ is chosen such that it satisfies Equations A.1, A.2, and A.4, then f is bijective. Since f is trivially continuous, as discussed in Section A.2.2, then f is a homeomorphism between the space of word sequences and the d-dimensional real space.

A.3 Homeomorphism holds when restricting to linguistically valid sentences

With the constraints and assumptions outlined in the proof sketch above, sequence representation functions, including contextualized word representations, define a homeomorphism between word sequences and real values. Of course, not all word sequences are valid in any given language: while "Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo" is a valid English sentence, "Duck duck duck duck duck" is not. However, the phenomenon of *center embedding* (e.g., recursive nesting of relative clauses) can in principle provide infinite length sentences, with infinite slots for each part of speech. Thus, if we retain our assumption of an infinite vocabulary (given the ability to coin neologisms), it is theoretically possible to produce

an uncountably infinite number of *linguistically valid* sentences within a human language. As the cardinality of this set matches the cardinality of the set of *all* word sequences, we can change our definition of the basic open sets in X to only be those word sequences that are grammatical, and the proof remains the same.

Empirically, center embedding has not been observed beyond a depth of three (Karlsson, 2007), and infinite coining of neologisms is impractical at the very least. Thus, for the kinds of practical analyses described in Section A.1, it remains virtually guaranteed that any given point in real space will not correspond to a sentence likely to be uttered in a natural setting. Nonetheless, the theoretical homeomorphism presented by sequence representation methodologies offers an intriguing avenue for further research into characterizing representation spaces.

Appendix B: Software packages and datasets contributed by this thesis

B.1 Software packages

JET: Jointly-embedded Entities and Text

- Implementation of concept-level representation learning method presented in Chapter 5.
- Originating publication: D Newman-Griffis, A M Lai, and E Fosler-Lussier, "Jointly Embedding Entities and Text with Distant Supervision". In Proceedings of the 3rd Workshop on Representation Learning for NLP, 2018.
- URL: `https://github.com/OSU-slatelab/JET`

NeuralVecmap

- Deep neural network method for learning mapping function from one set of learned representations to another.

- Originating publication: D Newman-Griffis and A Zirikly, "Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility". In Proceedings of BioNLP 2018, 2018.
- URL: `https://github.com/drgriffis/NeuralVecmap`

HARE: Highlighting Annotator for Ranking and Exploration

- Word-level relevance tagging model and web-based framework for reviewing output of information extraction models.
- Originating publication: D Newman-Griffis and E Fosler-Lussier, "HARE: a Flexible Highlighting Annotator for Ranking and Exploration". In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing: Systems Demonstrations. 2019.
- URL: `https://github.com/OSU-slatelab/HARE`

B.2 Datasets

WikiSRS: Wikipedia Similarity and Relatedness Set

- 688 pairs of Wikipedia entities, with human evaluations for similarity and relatedness.

- Originating publication: D Newman-Griffis, A M Lai, and E Fosler-Lussier, "Jointly Embedding Entities and Text with Distant Supervision". In Proceedings of the 3rd Workshop on Representation Learning for NLP, 2018.
- URL: `https://slate.cse.ohio-state.edu/WikiSRS/`

BMASS: Biomedical Analogic Similarity Dataset

- Analogical reasoning dataset derived from the Unified Medical Language System.
- Originating publication: D Newman-Griffis, A M Lai, and E Fosler-Lussier. "Insights into Analogy Completion from the Biomedical Domain." In Proceedings of the 16th Workshop on Biomedical Natural Language Processing (BioNLP), 2017.
- URL: `https://slate.cse.ohio-state.edu/BMASS/`

Bibliography

Kenneth Abbott, Yen-Yi Ho, and Jennifer Erickson. 2017. Automatic health record review to help prioritize gravely ill Social Security disability applicants. *Journal of the American Medical Informatics Association*, 24(4):709–716.

Eneko Agirre, Oier Lopez De Lacalle, and Aitor Soroa. 2009. Knowledge-based wsd on specific domains: Performing better than generic supervised WSD. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1501–1506.

Bader Al-Hablani. 2017. The Use of Automated SNOMED CT Clinical Coding in Clinical Decision Support Systems for Preventive Care. *Perspectives in health information management*, 14(Winter):1f–1f.

Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David Rosenberg, and Amanda Stent, editors. 2019. *Proceedings of the Natural Legal Language Processing Workshop 2019*. Association for Computational Linguistics, Minneapolis, Minnesota.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Barbara M. Altman. 2009. Population survey measures of functioning: strengths and weaknesses. In *Improving the Measurement of Late-Life Disability in Population Surveys: Beyond ADLs and IADLs: Summary of a Workshop*, chapter Appendix A, pages 99–156. The National Academies Press, Washington, DC.

James E Andrews, Rachel L Richesson, and Jeffrey Krischer. 2007. Variation of SNOMED CT Coding of Clinical Research Concepts among Coding Experts. *Journal of the American Medical Informatics Association*, 14(4):497–506.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, USA. Association for Computational Linguistics.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–36.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Joan S Ash, Marc Berg, and Enrico Coiera. 2004. Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors. *Journal of the American Medical Informatics Association*, 11(2):104–112.

David H Autor. 2011. The Unsustainable Rise of the Disability Rolls in the United States: Causes, Consequences, and Policy Options. Working Paper 17697, National Bureau of Economic Research.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words Are Malleable: Computing Semantic Shifts in Political and Media Discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1509–1518, New York, NY, USA. Association for Computing Machinery.

R Harald Baayen. 2001. *Word frequency distributions*. Kluwer Academic Publishers, Boston, MA.

Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. *International Conference on Intelligent Text Processing and Computational Linguistics*, 2276:136–145.

Sube Banerjee. 2015. Multimorbidity–older adults need health care that can count past one. *The Lancet*, 385(9968):587–589.

Marco Baroni, Stefan Evert, and Alessandro Lenci, editors. 2008. *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics: Bridging the gap between semantic theory and computational simulations*. Hamburg, Germany.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. *GEMS '11 Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

John R Beard, Alana Officer, Islene Araujo de Carvalho, Ritu Sadana, Anne Margriet Pot, Jean-Pierre Michel, Peter Lloyd-Sherlock, JoAnne E Epping-Jordan, G M E E Geeske Peeters, Wahyu Retno Mahanani, Jotheeswaran Amuthavalli Thiyagarajan, and Somnath Chatterji. 2016. The World report on ageing and health: a policy framework for healthy ageing. *Lancet*, 387(10033):2145–2154.

Olivier Beauchet, Cédric Annweiler, Michele L Callisaya, Anne-Marie De Cock, Jorunn L Helbostad, Reto W Kressig, Velandai Srikanth, Jean-Paul Steinmetz, Helena M Blumen, Joe Verghese, and Gilles Allali. 2016. Poor Gait Performance and Prediction of Dementia: Results From a Meta-Analysis. *Journal of the American Medical Directors Association*, 17(6):482–490.

Cosmin A Bejan, John Angiolillo, Douglas Conway, Robertson Nash, Jana K Shirey-Rice, Loren Lipworth, Robert M Cronin, Jill Pulley, Sunil Kripalani, Shari Barkin, Kevin B Johnson, and Joshua C Denny. 2018. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *Journal of the American Medical Informatics Association*, 25(1):61–71.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Asma Ben Abacha, Md. Faisal Mahbub Chowdhury, Aikaterini Karanasiou, Yassine Mrabet, Alberto Lavelli, and Pierre Zweigenbaum. 2015. Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification. *Journal of Biomedical Informatics*, 58:122–132.

Y Bengio, A Courville, and P Vincent. 2013. Representation Learning: A Review and New Perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

C Biagioli, E Francesconi, A Passerini, S Montemagni, and C Soria. 2005. Automatic Semantics Extraction in Law Documents. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, ICAIL '05, pages 133–140, New York, NY, USA. Association for Computing Machinery.

Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Daniel Bikel and Imed Zitouni. 2012. *Multilingual Natural Language Processing Applications: From Theory to Practice*, 1st edition. IBM Press.

Suzanne V Blackley, Jessica Huynh, Li Zhou, Liqin Wang, and Zfania Korach. 2019. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):324–338.

William Blacoe and Mirella Lapata. 2012. A Comparison of Vector-based Representations for Semantic Composition. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*, (July):546–556.

Katherine Blashki and Sophie Nichol. 2005. Game geek's goss: linguistic creativity in young males within an online university forum (94/\/\3 933k'5 9055oneone). *Australian journal of emerging technologies and society*, 3(2):71–80.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):D267–D270.

Olivier Bodenreider, Barry Smith, and Anita Burgun. 2004. The ontology-epistemology divide: a case study in medical terminology. In *Proceedings of the Third International Conference on Formal Ontology in Information Systems*, pages 185–195.

Sidney T. Bogardus, Virginia Towle, Christianna S. Williams, Mayur M. Desai, and Sharon Inouye. 2004. What Does the Medical Record Reveal about Functional Status? *Journal of General Internal Medicine*, 16(11):728–736.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the ACL*, 5:135–146.

Ond\vrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy.

Gemma Boleda. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6(1):213–234.

Tomas Borovicka, Marcel Jirina Jr, Pavel Kordik, and Marcel Jirina. 2012. Selecting representative data sets. *Advances in data mining knowledge discovery and applications*, pages 43–70.

Christopher R Bowie, Elizabeth W Twamley, Hannah Anderson, Brooke Halpern, Thomas L Patterson, and Philip D Harvey. 2007. Self-assessment of functional status in schizophrenia. *Journal of psychiatric research*, 41(12):1012–1018.

Rebecca T Brown, Kiya D Komaiko, Ying Shi, Kathy Z Fung, W John Boscardin, Alvin Au-Yeung, Gary Tarasovsky, Riya Jacob, and Michael A Steinman. 2017. Bringing functional status into a big data world: Validation of national Veterans Affairs functional status data. *PLOS ONE*, 12(6):e0178726.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47.

John Bryden, Sebastian Funk, and Vincent A A Jansen. 2013. Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science*, 2(1):3.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram Counts and Language Models from the Common Crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation ({LREC}'14)*, pages 3579–3584, Reykjavik, Iceland. European Language Resources Association (ELRA).

Risa B Burns, Mark A Moskowitz, Arlene Ash, Robert L Kane, Michael D Finch, and Sharon M Bak. 1992. Self-Report versus Medical Record Functional Status. *Medical Care*, 30(5):MS85–MS95.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. *J. Artif. Int. Res.*, 63(1):743–788.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge Text and Knowledge by Learning Multi-Prototype Entity Mention Embedding. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1623–1633, Vancouver, Canada. Association for Computational Linguistics.

Elise C Carey, Louise C Walter, Karla Lindquist, and Kenneth E Covinsky. 2004. Development and Validation of a Functional Morbidity Index to Predict Mortality

in Community-dwelling Elders. *Journal of General Internal Medicine*, 19(10):1027–1033.

David S Carrell, Robert E Schoen, Daniel A Leffler, Michele Morris, Sherri Rose, Andrew Baer, Seth D Crockett, Rebecca A Gourevitch, Katie M Dean, and Ateev Mehrotra. 2017. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *Journal of the American Medical Informatics Association*, 24(5):986–991.

Ming-Wei Chang, Kristina Toutanova, Kenton Lee, and Jacob Devlin. 2019. Language Model Pre-training for Hierarchical Document Representations.

Jean Charbonnier and Christian Wartena. 2018. Using Word Embeddings for Unsupervised Acronym Disambiguation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2610–2619. Association for Computational Linguistics.

Rachel Chasin, Anna Rumshisky, Ozlem Uzuner, and Peter Szolovits. 2014. Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods. *Journal of the American Medical Informatics Association*, 21(5):842–849.

Karen Chesbrough, Matt Elrod, and James J Irrgang. 2018. Systems Science in Rehabilitation Practice Realized. *Physical Therapy*, page In Press.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016a. How to Train Good Word Embeddings for Biomedical NLP. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016b. Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance. *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 1–6.

Jason P C Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional {LSTM}-{CNN}s. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016a. RETAIN: Interpretable Predictive Model in Healthcare using Reverse Time Attention Mechanism. In *NIPS*, pages 1–15.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, and Jimeng Sun. 2016b. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1495–1504, San Francisco, California, USA. ACM.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016c. Multi-layer Representation Learning for Medical Concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 1495–1504.

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016d. Learning Low-Dimensional Representations of Medical Concepts. In *AMIA Joint Summits on Translational Science Proceedings*, pages 41–50.

J J Cimino. 2012. The False Security of Blind Dates. *Appl Clin Inform*, 03(04):392–403.

James J. Cimino and Elaine J. Ayres. 2010. The clinical research data repository of the US National Institutes of Health. *Studies in Health Technology and Informatics*, 160(Pt 2):1299–1303.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What Does {BERT} Look at? An Analysis of {BERT}{'}s Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, Helsinki, Finland. Association for Computing Machinery.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Rachel Cooper, Diana Kuh, Rebecca Hardy, and Mortality Review Group. 2010. Objectively measured physical capability levels and mortality: systematic review and meta-analysis. *BMJ*, 341.

Rachel Cooper, Bjørn Heine Strand, Rebecca Hardy, Kushang V Patel, and Diana Kuh. 2014. Physical capability in mid-life and survival over 13 years of follow-up: British birth cohort study. *BMJ*, 348.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual Character-Level Neural Morphological Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological Word-Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Elizabeth A Courtney-Long, Dianna D Carroll, Qing C Zhang, Alissa C Stevens, Shannon Griffin-Blake, Brian S Armour, and Vincent A Campbell. 2015. Prevalence of Disability and Disability Type Among Adults–United States, 2013. *MMWR. Morbidity and mortality weekly report*, 64(29):777–783.

Rebecca J Crawford, Maryse Fortin, Kenneth A Weber, Andrew Smith, and James M Elliott. 2019. Are Magnetic Resonance Imaging Technologies Crucial to Our Understanding of Spinal Conditions? *Journal of Orthopaedic & Sports Physical Therapy*, 49(5):320–329.

Alan Cruse. 2004. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press, New York.

Mary F Davis, Subramaniam Sriram, William S Bush, Joshua C Denny, and Jonathan L Haines. 2013. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *Journal of the American Medical Informatics Association*, 20(e2):e334–e340.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management - CIKM '14*, CIKM '14, pages 1819–1822, Shanghai, China. ACM.

S Deerwester, S T Dumais, G W Furnas, T K Landauer, and R Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407.

Louise Deléger, Cyril Grouin, and Pierre Zweigenbaum. 2010. Extracting medical information from narrative patient records: the case of medication-related information. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):555–558.

Vincenzo Della Mea and Andrea Simoncello. 2012. An ontology-based exploration of the concepts and relationships in the activities and participation component of the international classification of functioning, disability and health. *Journal of Biomedical Semantics*, 3(1):1.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772.

Kerstin Denecke. 2014. Sublanguage Analysis of Medical Weblogs. *Studies in Health Technology and Informatics*, 205:565–569.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017a. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102, Copenhagen, Denmark. Association for Computational Linguistics.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017b. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query Expansion with Locally-Trained Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 367–377, Berlin, Germany. Association for Computational Linguistics.

Guy Divita, Shuying Shen, Marjorie E Carter, Andrew Redd, Tyler Forbush, Miland Palmer, Matthew H Samore, and Adi V Gundlapalli. 2014. Recognizing Questions and Answers in EMR Templates Using Natural Language Processing. *Studies in health technology and informatics*, 202:149–152.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *Journal of Biomedical Informatics*, 47:1–10.

Patrick Drouin. 2004. Detection of Domain Specific Terminology Using Corpora Comparison. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation ({LREC}{'}04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5. Association for Computational Linguistics.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. {S}em{E}val-2015 Task 14: Analysis

of Clinical Text. In *Proceedings of the 9th International Workshop on Semantic Evaluation ({S}em{E}val 2015)*, pages 303–310, Denver, Colorado. Association for Computational Linguistics.

Noémie Elhadad, Guergana Savova, Wendy Chapman, Glenn Zaramba, David Harris, and Amy Vogel. 2012. ShARe Guidelines for the Annotation of Modifiers for Disorders in Clinical Notes.

Peter L Elkin, Steven H Brown, Casey S Husser, Brent A Bauer, Dietlind Wahner-Roedler, S Trent Rosenbloom, and Ted Speroff. 2006. Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. *Mayo Clinic Proceedings*, 81(6):741–748.

K Erk and S Padó. 2008. A structured vector space model for word meaning in context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (October):897–906.

Letha H Etzkorn, Carl G Davis, and Lisa L Bowen. 2001. The language of comments in computer software: A sublanguage of English. *Journal of Pragmatics*, 33(11):1731–1756.

Wei Fang, Jianwen Zhang, Dilin Wang, Zheng Chen, and Ming Li. 2016. Entity Disambiguation by Knowledge and Text Jointly Embedding. In *Proceedings of the 20th SIGNLL Conference on Computational Language Learning (CoNLL)*, pages 260–269. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Safa Fathiamini, Amber M Johnson, Jia Zeng, Vijaykumar Holla, Nora S Sanchez, Funda Meric-Bernstam, Elmer V Bernstam, and Trevor Cohen. 2019. Rapamycin - mTOR + BRAF = ? Using relational similarity to find therapeutically relevant drug-gene relationships in unstructured text. *Journal of biomedical informatics*, 90:103094.

K Feldman, N Hazekamp, and N V Chawla. 2016. Mining the Clinical Narrative: All Text are Not Equal. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 271–280.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Rosa L Figueroa, Qing Zeng-Treitler, Sergey Goryachev, and Eduardo P Wiechmann. 2009. Tailoring vocabularies for NLP in sub-domains: a method to detect unused word sense. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009:188–192.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, (June):363–370.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 406–414, Hong Kong. ACM.

Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. 2016. Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214.

Zachary N Flamholz, Lyle H Ungar, and Gary Eric Weissman. 2019. Word embeddings trained on published case reports are lightweight,effective for clinical tasks, and free of protected health information. *medRxiv*.

Winthrop Nelson Francis. 1964. A Standard Sample of Present-Day English for Use with Digital Computers.

Winthrop Nelson Francis, Henry Kučera, and Andrew W Mackie. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin, Boston, MA.

Suzanne Fricke. 2018. Semantic Scholar. *Journal of the Medical Library Association : JMLA*, 106(1):145–147.

C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, and P. D. Clayton. 1995. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(01):83–108.

Carol Friedman. 1986. Automatic structuring of sublanguage information. In Ralph Grishman and Richard Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 85–102. Lawrence Erlbaum Associates.

Carol Friedman, Philip O Alderson, John H M Austin, James J Cimino, and Stephen B Johnson. 1994. A General Natural-language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: A description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Stephen Frochen and Shahla Mehdizadeh. 2017. Functional Status and Adaptation: Measuring Activities of Daily Living and Device Use in the National Health and Aging Trends Study. *Journal of Aging and Health*, 30(7):1136–1155.

Kin Wah Fung, Olivier Bodenreider, Alan R Aronson, William T Hole, and Suresh Srinivasan. 2007. Combining lexical and semantic methods of inter-terminology mapping using the UMLS. *Studies in health technology and informatics*, 129(Pt 1):605–609.

Alona Fyshe, Leila Wehbe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. 2015. A Compositional and Interpretable Semantic Space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 32–41, Denver, Colorado. Association for Computational Linguistics.

E R Gabrieli and D J Speth. 1986. Automated analysis of the discharge summary. *Journal of clinical computing*, 15(1):1–28.

Dieter Galea, Ivan Laponogov, and Kirill Veselkov. 2018. Sub-word information in pre-trained biomedical word representations: evaluation and hyper-parameter optimization. In *Proceedings of the BioNLP 2018 workshop*, pages 56–66, Melbourne, Australia. Association for Computational Linguistics.

Loïc Garçon, Chapal Khasnabis, Lloyd Walker, Yukiko Nakatani, Jostacio Lapitan, Johan Borg, Alex Ross, and Adriana Velazquez Berumen. 2016. Medical and Assistive Health Technology: Meeting the Needs of Aging Populations. *The Gerontologist*, 56(Suppl_2):S293–S302.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635—-E3644.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 36–42.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based Detection of Morphological and Semantic Relations With Word Embeddings: What Works and What Doesn't. *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Graciela H Gonzalez, Tasnia Tahsin, Britton C Goodale, Anna C Greene, and Casey S Greene. 2016. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings in Bioinformatics*, 17(1):33–42.

Graciela Gonzalez-Hernandez, Abeed Sarker, Kathryn O'Connor, and Guergana Savova. 2017. Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of Medical Informatics*, 26(1):214–227.

Jeffrey L Greenwald, Patrick R Cronin, Victoria Carballo, Goodarz Danaei, and Garry Choy. 2017. A Novel Model for Predicting Rehospitalization Risk Incorporating Physical Function, Cognitive Status, and Psychosocial Support Using Natural Language Processing. *Medical Care*, 55(3):261–266.

Mourad Gridach. 2017. Character-level neural network for biomedical named entity recognition. *Journal of Biomedical Informatics*, 70:85–91.

Denis Griffis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert M Lai. 2016. A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain. In *AMIA Summits on Translational Science Proceedings 2016*, pages 88–97. American Medical Informatics Association.

Ralph Grishman. 2001. Adaptive information extraction and sublanguage analysis. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining, Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, pages 1–4, Seattle, Washington, USA.

Ralph Grishman and Richard Kittredge, editors. 1986. *Analyzing language in restricted domains: sublanguage description and processing*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. Vancouver Welcomes You! Minimalist Location Metonymy Resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1248–1259, Vancouver, Canada. Association for Computational Linguistics.

Leonie Grön, Ann Bertels, and Kris Heylen. 2019. Leveraging Sublanguage Features for the Semantic Categorization of Clinical Terms. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 211–216, Florence, Italy. Association for Computational Linguistics.

Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA. ACM.

Stephen P Gulley, Elizabeth K Rasch, and Leighton Chan. 2011. If we build it, who will come? Working-age adults with chronic health care needs and the medical home. *Medical care*, 49(2):149–155.

Adi Gundlapalli, Guy Divita, Marjorie Carter, Shuying Shen, Miland Palmer, Tyler Forbush, Brett South, Andrew Redd, Brian Sauer, and Matthew Samore. 2013a. Extracting Surveillance Data from Templated Sections of an Electronic Medical Note: Challenges and Opportunities.

Adi V Gundlapalli, Marjorie E Carter, Miland Palmer, Thomas Ginter, Andrew Redd, Steven Pickard, Shuying Shen, Brett South, Guy Divita, Scott Duvall, Thien M Nguyen, Leonard W D'Avolio, and Matthew Samore. 2013b. Using Natural Language Processing on the Free Text of Clinical Documents to Screen for Evidence of Homelessness Among US Veterans. *AMIA Annual Symposium Proceedings*, pages 537–546.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Melissa A Haendel, Christopher G Chute, and Peter N Robinson. 2018. Classification, Ontology, and Precision Medicine. *New England Journal of Medicine*, 379(15):1452–1462.

K Haerian, D Varn, S Vaidya, L Ena, H S Chase, and C Friedman. 2012. Detection of Pharmacovigilance-Related Adverse Events Using Electronic Health Records and Automated Methods. *Clinical Pharmacology & Therapeutics*, 92(2):228–234.

Udo Hahn, Véronique Hoste, and Ming-Feng Tsai, editors. 2018. *Proceedings of the First Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics, Melbourne, Australia.

Udo Hahn, Véronique Hoste, and Zhu Zhang, editors. 2019. *Proceedings of the Second Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics, Hong Kong.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016a. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Zellig Harris. 1968. Mathematical structures of language. *Interscience Tracts in Pure and Applied Mathematics*.

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Dennis L Hart, Mark W Werneke, Daniel Deutscher, Steven Z George, Paul W Stratford, and Jerome E Mioduski. 2011. Using Intake and Change in Multiple Psychosocial Measures to Predict Functional Status Outcomes in People With Lumbar Spine Syndromes: A Preliminary Analysis. *Physical Therapy*, 91(12):1812–1825.

Zhe He, Zhiwei Chen, Sanghee Oh, Jinghui Hou, and Jiang Bian. 2017. Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach. *Journal of Biomedical Informatics*, 69:75–85.

Yvonne F Heerkens, Marjolein de Weerd, Machteld Huber, Carin P M de Brouwer, Sabina van der Veen, Rom J M Perenboom, Coen H van Gool, Huib ten Napel, Marja van Bon-Martens, Hillegonda A Stallinga, and Nico L U van Meeteren. 2018. Reconsideration of the scheme of the international classification of functioning, disability and health: incentives from the Netherlands for a global debate. *Disability and Rehabilitation*, 40(5):603–611.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726.

G E Hinton, J L McClelland, and D E Rumelhart. 1986. *Distributed Representations*, pages 77–109. MIT Press, Cambridge, MA, USA.

GE Hinton. 1986. Learning distributed representations of concepts. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 1–12.

Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

M Hopfe, G Stucki, R Marshall, C D Twomey, T B Ustun, and B Prodinger. 2016. Capturing patients' needs in casemix: a systematic literature review on the value of adding functioning information in reimbursement systems. *BMC Health Serv Res*, 16:40.

Maren Hopfe, Birgit Prodinger, Jerome E. Bickenbach, and Gerold Stucki. 2018. Optimizing health system response to patient's needs: an argument for the importance of functioning information. *Disability and Rehabilitation*, 40(19):2325–2330.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Robert Hoyt and Ann Yoshihashi. 2010. Lessons learned from implementation of voice recognition for documentation in the military electronic health record system. *Perspectives in health information management*, 7(Winter):1e–1e.

George Hripcsak, Parsa Mirhaji, Alexander F H Low, and Bradley A Malin. 2016. Preserving temporal relations in clinical data while maintaining privacy. *Journal of the American Medical Informatics Association*, 23(6):1040–1045.

Chung-Chi Huang and Zhiyong Lu. 2016. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in Bioinformatics*, 17(1):132–144.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907. Association for Computational Linguistics.

Nancy Ide and Jean Veronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1):1 – 40.

Institute of Medicine. 1991. *Disability in America: Toward a National Agenda for Prevention*. National Academy Press, Washington, DC.

Institute of Medicine. 1997. *Enabling America: assessing the role of rehabilitation science and engineering*. National Academy Press, Washington, DC.

Ganesh Jawahar, Beno\ˆ\it Sagot, and Djamé Seddah. 2019. What Does {BERT} Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Antonio Jimeno-Yepes. 2017. Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. *Journal of Biomedical Informatics*, 73:137–147.

Antonio Jimeno-Yepes and Rafael Berlanga. 2015. Knowledge based word-concept model estimation and refinement for biomedical text mining. *Journal of Biomedical Informatics*, 53:300–307.

Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics*, 12:223.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Gwyn C Jones and Kianda Bell. 2004. Adverse Health Behaviors and Chronic Conditions in Working-Age Women with Disabilities. *Family & Community Health*, 27(1).

Hayley E Jones, David I Ohlssen, and David J Spiegelhalter. 2008. Use of the false discovery rate when comparing multiple health care providers. *Journal of Clinical Epidemiology*, 61(3):232 – 240.e2.

Venkata Joopudi, Bharath Dandala, and Murthy Devarakonda. 2018. A convolutional route to abbreviation disambiguation in clinical text. *Journal of Biomedical Informatics*, 86(December 2017):71–78.

Jelena Jovanović and Ebrahim Bagheri. 2017. Semantic annotation in biomedicine: The current landscape. *Journal of Biomedical Semantics*, 8(1):1–18.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Convolutional Neural Networks for Discourse Compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126, Sofia, Bulgaria. Association for Computational Linguistics.

Fred Karlsson. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43(2):365–392.

Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii. 2002. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the {ACL}-02 Workshop on Natural Language Processing in the Biomedical Domain*, pages 1–8, Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Victoria L Keevil, Robert Luben, Shabina Hayat, Avan A Sayer, Nicholas J Wareham, and Kay-Tee Khaw. 2018. Physical capability predicts mortality in late mid-life as well as in old age: Findings from a large British cohort study. *Archives of Gerontology and Geriatrics*, 74:77–82.

J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - A semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(SUPPL. 1):180–182.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving Adjectival Scales from Continuous Space Word Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630, Seattle, Washington, USA. Association for Computational Linguistics.

Joo-Kyung Kim, Marie-Catherine de Marneffe, and Eric Fosler-Lussier. 2016a. Adjusting Word Embeddings with Semantic Intensity Orders. In *Proceedings of the 1st Workshop on Representation Learning for {NLP}*, pages 62–69, Berlin, Germany. Association for Computational Linguistics.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016b. Character-aware Neural Language Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2741–2749. AAAI Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–15.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Richard Kittredge and John Lehrberger, editors. 1982. *Sublanguage: Studies of language in restricted semantic domains*. Walter de Gruyter.

Sebastian Köhler, Nicole A Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Ségolène Aymé, Gareth Baynam, Susan M Bello, Cornelius F Boerkoel, Kym M Boycott, Michael Brudno, Orion J Buske, Patrick F Chinnery, Valentina Cipriani, Laureen E Connell, Hugh J S Dawkins, Laura E DeMare, Andrew D Devereau, Bert B.A. de Vries, Helen V Firth, Kathleen Freson, Daniel Greene, Ada Hamosh, Ingo Helbig, Courtney Hum, Johanna A Jähn, Roger James, Roland Krause, Stanley J F. Laulederkind, Hanns Lochmüller, Gholson J Lyon, Soichi Ogishima, Annie Olry, Willem H Ouwehand, Nikolas Pontikos, Ana Rath, Franz Schaefer, Richard H Scott, Michael Segal, Panagiotis I Sergouniotis, Richard Sever, Cynthia L Smith, Volker Straub, Rachel Thompson, Catherine Turner, Ernest Turro, Marijcke W M Veltman, Tom Vulliamy, Jing Yu, Julie von Ziegenweidt, Andreas Zankl, Stephan Züchner, Tomasz Zemojtel, Julius O B Jacobsen, Tudor Groza, Damian Smedley, Christopher J Mungall, Melissa Haendel, and Peter N Robinson. 2017. The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45(D1):D865–D876.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73:14–29.

Jinqiu Kuang, April F Mohanty, V H Rashmi, Charlene R Weir, Bruce E Bray, and Qing Zeng-Treitler. 2015. Representation of Functional Status Concepts from Clinical Documents and Social Media Sources by Standard Terminologies. In *AMIA Annual Symposium Proceedings 2015*, pages 795–803. American Medical Informatics Association.

Rita Kukafka, Michael E. Bales, Ann Burkhardt, and Carol Friedman. 2006. Human and Automated Coding of Rehabilitation Discharge Summaries According to the International Classification of Functioning, Disability, and Health. *Journal of the American Medical Informatics Association*, 13(5):508–515.

Casimir A Kulikowski, Edward H Shortliffe, Leanne M Currie, Peter L Elkin, Lawrence E Hunter, Todd R Johnson, Ira J Kalet, Leslie A Lenert, Mark A Musen, Judy G Ozbolt, Jack W Smith, Peter Z Tarczy-Hornoch, and Jeffrey J Williamson.

2012. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *Journal of the American Medical Informatics Association*, 19(6):931–938.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot Word Sense Disambiguation using Sense Definition Embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Veronika Laippala, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. 2009. Towards automated processing of clinical Finnish: Sublanguage analysis and a rule-based parser. *International Journal of Medical Informatics*, 78(12):e7 – e12.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

François-Michel Lang, James G Mork, Dina Demner-Fushman, and Alan R Aronson. 2017. Increasing UMLS Coverage and Reducing Ambiguity via Automated Creation of Synonymous Terms: First Steps toward Filling UMLS Synonymy Gaps.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, pages 1–8.

Alessandro Lenci. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1):151–171.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Jake Lever, Eric Y Zhao, Jasleen Grewal, Martin R Jones, and Steven J M Jones. 2019. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nature Methods*, 16(6):505–507.

Omer Levy and Yoav Goldberg. 2014a. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014c. Neural Word Embedding as Implicit Matrix Factorization. *Advances in Neural Information Processing Systems (NIPS)*, pages 2177–2185.

I-Fen Lin and Hsueh-Sheng Wu. 2014. Activity Limitations, Use of Assistive Devices or Personal Help, and Well-Being: Variation by Education. *The Journals of Gerontology: Series B*, 69(Suppl_1):S16–S25.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Nelson F Liu, Omer Levy, Roy Schwartz, Chenhao Tan, and Noah A Smith. 2018. LSTMs Exploit Linguistic Attributes of Data. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 180–186, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75:S34–S42.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018. Leveraging Gloss Knowledge in Neural Word Sense Disambiguation by Hierarchical Co-Attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411, Brussels, Belgium. Association for Computational Linguistics.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. MCN: A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, 92:103132.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

Jennifer H Madans, Barbara M Altman, Elizabeth K Rasch, Malin Synneborn, Jeremiah Banda, Margaret Mbogoni, Angela Me, and Elena DePalma. 2004. Proposed Purpose of an Internationally Comparable General Disability Measure. In *Washington Group Meeting*, Brussels, Belgium.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Mitchell P Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

John Maret-Ouda, Wenjing Tao, Karl Wahlin, and Jesper Lagergren. 2017. Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. *Scandinavian Journal of Public Health*, 45(17_suppl):14–19.

Katja Markert and Malvina Nissim. 2009. Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138.

Elaine Marsh. 1986. General semantic patterns in different sublanguages. In Ralph Grishman and Richard Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 103–127. Lawrence Erlbaum Associates, Hillsdale, NJ.

Dimitrios Mavroeidis, George Tsatsaronis, Michalis Vazirgiannis, Martin Theobald, and Gerhard Weikum. 2005. Word Sense Disambiguation for Exploiting Hierarchical Thesauri in Text Classification BT - Knowledge Discovery in Databases: PKDD 2005. pages 181–192, Berlin, Heidelberg. Springer Berlin Heidelberg.

Andrew McCallum and Wei Li. 2003. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at {HLT}-{NAACL} 2003*, pages 188–191.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

Bridget T. McInnes and Ted Pedersen. 2013. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, 46(6):1116–1124.

Bridget T. McInnes and Ted Pedersen. 2015. Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. *Journal of Biomedical Informatics*, 54:329–336.

Bridget T McInnes, Ted Pedersen, Ying Liu, Serguei V Pakhomov, and Genevieve B Melton. 2011. Using second-order vectors in a knowledge-based method for acronym disambiguation. *CoNLL 2011 - Fifteenth Conference on Computational Natural Language Learning, Proceedings of the Conference*, (June):145–153.

Bridget T McInnes, Ted Pedersen, and Serguei V S Pakhomov. 2009. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009:431–435.

Anna McPherson, Jo Durham, Nicola Richards, Hebe Gouda, Rasika Rampatige, and Maxine Whittaker. 2017. Strengthening health information systems for disability-related rehabilitation in LMICs. *Health Policy and Planning*, 32(3):384–394.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional {LSTM}. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Eneldo Loza Mencia, Gerard de Melo, and Jinseok Nam. 2016. Medical Concept Embeddings via Labeled Background Corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4629–4636. European Language Resources Association (ELRA).

Stephane Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, pages 128–144.

Stéphane M Meystre, Óscar Ferrández, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2014. Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics*, 50:142–150.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, pages 1–12.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH-2010*, pages 1045–1048. IEEE.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems 26*, NIPS '13, pages 3111–3119. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner, editors. 2019. *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*. Association for Computational Linguistics, Hong Kong, China.

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Andriy Mnih and Geoffrey E Hinton. 2008. A Scalable Hierarchical Distributed Language Model. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1081–1088. Curran Associates, Inc.

Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. 2014. A sense inventory for clinical abbreviations and acronyms created

using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *AISTATS*.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297. Association for Computational Linguistics.

Danielle L Mowery, Brett R South, Lee Christensen, Jianwei Leng, Laura-Maria Peltonen, Sanna Salanterä, Hanna Suominen, David Martinez, Sumithra Velupillai, Noémie Elhadad, Guergana Savova, Sameer Pradhan, and Wendy W Chapman. 2016. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2. *Journal of Biomedical Semantics*, 7(1):43.

Danielle L Mowery, Sumithra Velupillai, Brett R South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, and Wendy Chapman. 2014. Task 2: ShARe/CLEF eHealth Evaluation Lab 2014. In *Online Working Notes of the CLEF 2014 Evaluation Labs and Workshop*, Sheffield, United Kingdom.

T H Muneeb, Sunil Kumar Sahu, and Ashish Anand. 2015. Evaluating distributed word representations for capturing semantics of biomedical concepts. In *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015)*, pages 158–163, Beijing, China. Association for Computational Linguistics.

James R Munkres. 2013. *Topology*, second edition. Pearson Education Limited.

M. Lynne Murphy. 2010. *Lexical Meaning*. Cambridge University Press, Cambridge, UK.

Saad Z Nagi. 1965. Some conceptual issues in disability and rehabilitation. In M B Sussman, editor, *Sociology and Rehabilitation*, pages 100–113. American Sociological Association, Washington, DC.

Vivi Nastase, Benjamin Roth, Laura Dietz, and Andrew McCallum, editors. 2019. *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*. Association for Computational Linguistics, Minneapolis, Minnesota.

R Navigli and M Lapata. 2010. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.

R Navigli and P Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1086.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Comput. Surv.*, 41(2):10:1—-10:69.

Roberto Navigli, Stefano Faralli, Aitor Soroa, Oier de Lacalle, and Eneko Agirre. 2011. Two Birds with One Stone: Learning Semantic Models for Text Categorization and Word Sense Disambiguation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2317–2320, New York, NY, USA. ACM.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. {S}cispa{C}y: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Alan Newell, Stefan Langer, and Marianne Hickey. 1998. The role of natural language processing in alternative and augmentative communication. *Natural Language Engineering*, 4(1):1–16.

Denis Newman-Griffis and Eric Fosler-Lussier. 2017. Second-Order Word Embeddings from Nearest Neighbor Topological Features. *arXiv preprint arXiv:1705.08488*.

Denis Newman-Griffis and Eric Fosler-Lussier. 2019a. HARE: a Flexible Highlighting Annotator for Ranking and Exploration. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 85–90, Hong Kong, China. Association for Computational Linguistics.

Denis Newman-Griffis and Eric Fosler-Lussier. 2019b. Writing habits and telltale neighbors: analyzing clinical concept usage patterns with sublanguage embeddings. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 146–156, Hong Kong. Association for Computational Linguistics.

Denis Newman-Griffis, Albert M Lai, and Eric Fosler-Lussier. 2017. Insights into Analogy Completion from the Biomedical Domain. In *BioNLP 2017*, pages 19–28, Vancouver, Canada. Association for Computational Linguistics.

Denis Newman-Griffis, Albert M Lai, and Eric Fosler-Lussier. 2018. Jointly Embedding Entities and Text with Distant Supervision. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 195–206. Association for Computational Linguistics.

Denis Newman-Griffis, Julia Porcino, Ayah Zirikly, Thanh Thieu, Jonathan Camacho Maldonado, Pei-Shu Ho, Min Ding, Leighton Chan, and Elizabeth Rasch. 2019a. Broadening horizons: the case for capturing function and the role of health informatics in its use. *BMC Public Health*, 19(1):1288.

Denis Newman-Griffis and Ayah Zirikly. 2018. Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility. In *Proceedings of the BioNLP 2018 workshop*, pages 1–11, Melbourne, Australia. Association for Computational Linguistics.

Denis Newman-Griffis, Ayah Zirikly, Guy Divita, and Bart Desmet. 2019b. Classifying the reported ability in clinical mobility descriptions. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Francesca M Nicosia, Malena J Spar, Michael A Steinman, Sei J Lee, and Rebecca T Brown. 2019. Making Function Part of the Conversation: Clinician Perspectives on Measuring Functional Status in Primary Care. *Journal of the American Geriatrics Society*, 67(3):493–502.

Anika Oellrich, Nigel Collier, Tudor Groza, Dietrich Rebholz-Schuhmann, Nigam Shah, Olivier Bodenreider, Mary Regina Boland, Ivo Georgiev, Hongfang Liu, Kevin Livingston, Augustin Luna, Ann-Marie Mallon, Prashanti Manda, Peter N Robinson, Gabriella Rustici, Michelle Simon, Liqin Wang, Rainer Winnenburg, and Michel Dumontier. 2015. The digital revolution in phenotyping. *Briefings in bioinformatics*, (August):bbv083–.

Philip V Ogren. 2006. Knowtator: A Protégé plug-in for annotated corpus construction. In *Proceedings of the Human Language Technology Conference of the {NAACL}, Companion Volume: Demonstrations*, pages 273–275, New York City, USA. Association for Computational Linguistics.

Lucila Ohno-Machado. 2018. Sharing data from electronic health records within, across, and beyond healthcare institutions: Current trends and perspectives. *Journal of the American Medical Informatics Association*, 25(9):1113.

Michel Oleynik, Markus Kreuzthaler, and Stefan Schulz. 2017. Unsupervised Abbreviation Expansion in Clinical Narratives. *Studies in health technology and informatics*, 245:539–543.

James O'Neill, Paul Buitelaar, Cecile Robin, and Leona O'Brien. 2017. Classifying Sentential Modality in Legal Language: A Use Case in Financial Regulations, Acts and Directives. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, pages 159–168, New York, NY, USA. Association for Computing Machinery.

John D Osborne. 2015. Annotation Guidelines for Annotating CUI-less Concepts in BRAT.

John D Osborne, Matthew B Neu, Maria I Danila, Thamar Solorio, and Steven J Bethard. 2018. CUILESS2016: a clinical corpus applying compositional normalization of text mentions. *Journal of Biomedical Semantics*, 9(1):2.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. In *AMIA Annual Symposium Proceedings*, pages 572–576. American Medical Informatics Association.

Serguei V S Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B Melton. 2016. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32(23):3635–3644.

Keith Palmer, Stefania D'Angelo, E Clare Harris, Cathy Linaker, Catharine R Gale, Holly Syddall, Tjeerd Van Staa, Cyrus Cooper, Avan Aihie Sayer, David Coggon, and Karen Walker-Bone. 2016. Frailty, pre-frailty and employment outcomes in the health and employment after fifty (HEAF) study. *Occupational and Environmental Medicine*, 73(Suppl 1):A64—-A64.

Simone Papandrea, Alessandro Raganato, and Claudio Delli Bovi. 2017. SupWSD: A Flexible Toolkit for Supervised Word Sense Disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 103–108. Association for Computational Linguistics.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition LDC2011T07. *Linguistic Data Consortium*.

Talcott Parsons. 1937. *The Structure of Social Action: A Study in Social Theory with Special Reference to a Group of Recent European Writers*. McGraw-Hill Book Co., Inc., New York.

Olga Patterson and John F Hurdle. 2011. Document clustering of clinical narratives: a systematic study of clinical sublanguages. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2011:1099–1107.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Computational linguistics and intelligent text processing*, pages 241–257. Springer.

Douglas B Paul and Janet M Baker. 1992. The Design for the Wall Street Journal-Based CSR Corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 357–362, USA. Association for Computational Linguistics.

Thomas H Payne, Paul C Tang, William M Tierney, Charlotte Weaver, Charlene R Weir, Michael H Zaroukian, Sarah Corley, Theresa A Cullen, Tejal K Gandhi, Linda Harrington, Gilad J Kuperman, John E Mattison, David P McCallie, and Clement J McDonald. 2015. Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. *Journal of the American Medical Informatics Association*, 22(5):1102–1110.

Ted Pedersen. 2010. The Effect of Different Context Representations on Word Sense Discrimination in Biomedical Texts. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 56–65, New York, NY, USA. ACM.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Subashan Perera, Kushang V Patel, Caterina Rosano, Susan M Rubin, Suzanne Satterfield, Tamara Harris, Kristine Ensrud, Eric Orwoll, Christine G Lee, Julie M Chandler, Anne B Newman, Jane A Cauley, Jack M Guralnik, Luigi Ferrucci, and Stephanie A Studenski. 2015. Gait Speed Predicts Incident Disability: A Pooled Analysis. *The Journals of Gerontology: Series A*, 71(1):63–71.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized Page Rank for Named Entity Disambiguation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 238–243, Denver, Colorado. Association for Computational Linguistics.

Ahmad Pesaranghader, Stan Matwin, Marina Sokolova, and Ali Pesaranghader. 2019. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data. *Journal of the American Medical Informatics Association*, 26(5):438–446.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Minh C Phan, Aixin Sun, and Yi Tay. 2019. Robust Representation Learning of Biomedical Names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285, Florence, Italy. Association for Computational Linguistics.

Elizabeth Pisani, Peter Aaby, J Gabrielle Breugelmans, David Carr, Trish Groves, Michelle Helinski, Dorcas Kamuya, Steven Kern, Katherine Littler, Vicki Marsh, Souleymane Mboup, Laura Merson, Osman Sankoh, Micaela Serafini, Martin Schneider, Vreni Schoenenberger, and Philippe J Guerin. 2016. Beyond open data: realising the health benefits of sharing data. *BMJ*, 355.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. {S}em{E}val-2014 Task 7: Analysis of Clinical Text. In *Proceedings of the 8th International Workshop on Semantic Evaluation ({S}em{E}val 2014)*, pages 54–62, Dublin, Ireland. Association for Computational Linguistics.

Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association : JAMIA*, 22(1):143–54.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, Valencia, Spain.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92. Association for Computational Linguistics.

James Pustejovsky, Peter Anick, and Sabine Bergler. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358.

Alec Radford, Karthik Naramsihan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A Word at a Time: Computing Word Relatedness Using Temporal Semantic Analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA. ACM.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017a. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110. Association for Computational Linguistics.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017b. Neural Sequence Learning Models for Word Sense Disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167. Association for Computational Linguistics.

Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):3.

David L Ranum. 1989. Knowledge-based understanding of radiology text. *Computer Methods and Programs in Biomedicine*, 30(2):209–215.

Majid Rastegar-Mojarad, Jenna K. Lovely, Joshua Pankratz, Sunghwan Sohn, Donna M. Ihrke, Amit Merchea, David W. Larson, and Hongfang Liu. 2017. Using Unstructured Data to Identify Readmitted Patients. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–4.

Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview LSA : Representation Learning via Generalized CCA. *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL*, (1961):556–566.

T C Rindflesch and A R Aronson. 1994. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proceedings. Symposium on Computer Applications in Medical Care*, pages 240–244.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too Many) Problems of Analogical Reasoning with Word Vectors. *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in Your Embedding, And How It Predicts Task Performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703, Santa Fe, NM, USA. Association for Computational Linguistics.

S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association : JAMIA*, 18(2):181–186.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

A K M Sabbir, Antonio Jimeno-Yepes, and Ramakanth Kavuluru. 2017. Knowledge-Based Biomedical Word Sense Disambiguation with Neural Concept Embeddings. *Proceedings. IEEE International Symposium on Bioinformatics and Bioengineering*, 2017:163–170.

N Sager, I D J Bross, G Story, P Bastedo, E Marsh, and D Shedd. 1982. Automatic encoding of clinical narrative. *Computers in Biology and Medicine*, 12(1):43–56.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Guergana K Savova, Anni R Coden, Igor L Sominsky, Rie Johnson, Philip V Ogren, Piet C de Groen, and Christopher G Chute. 2008. Word sense disambiguation across two domains: Biomedical literature and clinical notes. *Journal of Biomedical Informatics*, 41(6):1088–1100.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Martijn J Schuemie, Jan A Kors, and Barend Mons. 2005. Word Sense Disambiguation in the Biomedical Domain: An Overview. *Journal of Computational Biology*, 12(5):554–565.

Douglas R. Seals, Jamie N. Justice, and Thomas J. Larocca. 2016. Physiological geroscience: Targeting function to increase healthspan and achieve optimal longevity. *Journal of Physiology*, 594(8):2001–2024.

Burr Settles. 2004. Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications ({NLPBA}/{B}io{NLP})*, pages 107–110, Geneva, Switzerland. COLING.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2016. Appraising UMLS Coverage for Summarizing Medical Evidence. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 513–524, Osaka, Japan. The COLING 2016 Organizing Committee.

Yijun Shao, April F Mohanty, Ali Ahmed, Charlene R Weir, Bruce E Bray, Rashmee U Shah, Douglas Redd, and Qing Zeng-Treitler. 2016. Identification and Use of Frailty Indicators from Text to Examine Associations with Clinical Outcomes Among Patients with Heart Failure. In *AMIA Annual Symposium Proceedings*, pages 1110–1118. American Medical Informatics Association.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. 2013. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230.

Susan M Shortreed, Andrea J Cook, R Yates Coley, Jennifer F Bobb, and Jennifer C Nelson. 2019. Challenges and Opportunities for Using Big Health Care Data to Advance Medical Science and Public Health. *American Journal of Epidemiology*, 188(5):851–861.

Rune J Simeonsson, Donald Lollar, Joseph Hollowell, and Mike Adams. 2000. Revision of the International Classification of Impairments, Disabilities, and Handicaps: Developmental issues. *Journal of Clinical Epidemiology*, 53(2):113–124.

Matthew S Simpson and Dina Demner-Fushman. 2012. Biomedical Text Mining: A Survey of Recent Progress. *Mining Text Data*, pages 465–517.

Steven J Skube, Elizabeth A Lindemann, Elliot G Arsoniadis, Mari Akre, Elizabeth C Wick, and Genevieve B Melton. 2018. Characterizing Functional Health Status of Surgical Patients in Clinical Notes. In *AMIA Joint Summits on Translational Science Proceedings 2018*, pages 379–388. American Medical Informatics Association.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

Richard Socher, Eric H Huang, Jeffrey Pennington, Andrew Y Ng, and Christopher D Manning. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. *Advances in Neural Information Processing Systems*, pages 801–809.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, Jeju Island, Korea. Association for Computational Linguistics.

Chris Sorna, Richard Steele, and Atsushi Inoue. 2009. Word prediction in assistive technologies for aphasia rehabilitation using Systemic Functional Grammar. In *NAFIPS 2009 - 2009 Annual Meeting of the North American Fuzzy Information Processing Society*, pages 1–6. IEEE.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2018. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Michelle C Specht, Michael W Kattan, Mithat Gonen, Jane Fey, and Kimberly J Van Zee. 2005. Predicting Nonsentinel Node Status After Positive Sentinel Lymph Biopsy for Breast Cancer: Clinicians Versus Nomogram. *Annals of Surgical Oncology*, 12(8):654–659.

Saskia Steinheimer, Jonas F Dorn, Cecily Morrison, Advait Sarkar, Marcus D'Souza, Jacques Boisvert, Rishi Bedi, Jessica Burggraaff, Peter Kontschieder, Frank Dahlke, Abigail Sellen, Bernard M J Uitdehaag, Ludwig Kappos, and Christian P Kamm. 2019. Setwise comparison: efficient fine-grained rating of movement videos using algorithmic support – a proof of concept study. *Disability and Rehabilitation*, pages 1–7.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for {NLP}-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the {E}uropean Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Mark Stevenson, Eneko Agirre, and Aitor Soroa. 2011. Exploiting domain information for Word Sense Disambiguation of medical documents. *Journal of the American Medical Informatics Association*, 19(2):235–240.

Mark Stevenson and Yikun Guo. 2010. Disambiguation in the biomedical domain: The role of ambiguity type. *Journal of Biomedical Informatics*, 43(6):972–981.

Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of Biomedical Informatics*, 58:S67–S77.

G Stucki, J Bickenbach, and J Melvin. 2017. Strengthening Rehabilitation in Health Systems Worldwide by Integrating Information on Functioning in National Health Information Systems. *American Journal of Physical Medicine & Rehabilitation*, 96(9):677–681.

Gerold Stucki and Jerome Bickenbach. 2017a. Functioning information in the learning health system. *European Journal of Physical and Rehabilitation Medicine*, 53(1):139–143.

Gerold Stucki and Jerome Bickenbach. 2017b. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. *European journal of physical and rehabilitation medicine*, 53(1):134–138.

Gerold Stucki, Jerome Bickenbach, Christoph Gutenbrunner, and John Melvin. 2018. Rehabilitation: The health strategy of the 21st century. *Journal of Rehabilitation Medicine*, 50(4):309–316.

Vidyalakshmi Sundar, Marcia E Daumen, Daniel J Conley, and John H Stone. 2008. The use of ICF codes for information retrieval in rehabilitation research: An empirical study. *Disability and Rehabilitation*, 30(12-13):955–962.

Svetlana Symonenko, Steven Rowe, and Elizabeth D Liddy. 2006. Illuminating Trouble Tickets with Sublanguage Theory. In *Proceedings of the Human Language Technology Conference of the {NAACL}, Companion Volume: Short Papers*, pages 169–172, New York City, USA. Association for Computational Linguistics.

Jane Taggart, Siaw-Teng Liaw, and Hairong Yu. 2015. Structured data quality reports to improve EHR data quality. *International Journal of Medical Informatics*, 84(12):1094–1098.

Nadine Tamburrini, Marco Cinnirella, Vincent A A Jansen, and John Bryden. 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40:84–89.

Yi Chern Tan and L Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13230–13241. Curran Associates, Inc.

Yu Taniguchi, Akihiko Kitamura, Yu Nofuji, Tatsuro Ishizaki, Satoshi Seino, Yuri Yokoyama, Tomohiro Shinozaki, Hiroshi Murayama, Seigo Mitsutake, Hidenori Amano, Mariko Nishi, Yutaka Matsuyama, Yoshinori Fujiwara, and Shoji Shinkai. 2018. Association of Trajectories of Higher-Level Functional Capacity with Mortality and Medical and Long-Term Care Costs Among Community-Dwelling Older Japanese. *The Journals of Gerontology: Series A*, page gly024.

Irina P Temnikova, William A Baumgartner, Negacy D Hailu, Ivelina Nikolova, Tony McEnery, Adam Kilgarriff, Galia Angelova, and K Bretonnel Cohen. 2014. Sublanguage Corpus Analysis Toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. *LREC ... International Conference on Language Resources & Evaluation : [proceedings]. International Conference on Language Resources and Evaluation*, 2014:1714–1718.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Thanh Thieu, Jonathan Camacho, Pei-Shu Ho, Julia Porcino, Min Ding, Lisa Nelson, Elizabeth Rasch, Chunxiao Zhou, Leighton Chan, Diane Brandt, Denis Newman-Griffis, Ao Yuan, and Albert M Lai. 2017. Inductive identification of functional status information and establishing a gold standard corpus: A case study on the Mobility domain. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2300–2302. IEEE.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing Text for Joint Embedding of Text and Knowledge Bases. In *Proceedings of the 2015 Conference on Empirical Methods in*

Natural Language Processing, pages 1499–1509, Lisbon, Portugal. Association for Computational Linguistics.

Debbie A Travers and Stephanie W Haas. 2006. Unified Medical Language System Coverage of Emergency-medicine Chief Complaints. *Academic Emergency Medicine*, 13(12):1319–1323.

Richard Tzong-Han Tsai, Cheng-Lung Sung, Hong-Jie Dai, Hsieh-Chuan Hung, Ting-Yi Sung, and Wen-Lian Hsu. 2006. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics*, 7(5):S11.

Samson W Tu, Csongor I Nyulas, Tania Tudorache, and Mark A Musen. 2015. A Method to Compare ICF and SNOMED CT for Coverage of U.S. Social Security Administration's Disability Listing Criteria. *AMIA Annual Symposium Proceedings*, pages 1224–1233.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Population Division United Nations,, Department of Economic and Social Affairs,. 2017. World Population Prospects: The 2017 Revision, Key Findings and Advance Tables. Technical report.

US Social Security Administration. 2008. Disability Evaluation Under Social Security.

US Social Security Administration. 2014. Consultative Examinations: A Guide for Health Professionals.

Tolga Uslu, Alexander Mehler, Daniel Baumartz, Alexander Henlein, and Wahed Hemati. 2018. {F}ast{S}ense: An Efficient Word Sense Disambiguation Classifier. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying Patient Smoking Status from Medical Discharge Records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Özlem Uzuner, Sam Henry, and Yen-Fu Luo. 2019. 2019 n2c2 Shared-Task and Workshop Track 3: n2c2/UMass Track on Clinical Concept and Normalization.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–6.

L J P Van Der Maaten and G E Hinton. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Gaurav Vashisth, Jan-Niklas Voigt-Antons, Michael Mikhailov, and Roland Roller. 2019. Exploring Diachronic Changes of Biomedical Knowledge using Distributed Concept Representations. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 348–358, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

S Velupillai, D Mowery, B R South, M Kvist, and H Dalianis. 2015. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. *Yearb Med Inform*, 24(01):183–193.

Lois M. Verbrugge. 2016. *Disability Experience and Measurement*, volume 28.

Lois M. Verbrugge, James M. Lepkowski, and Yuichi Imanaka. 1989. Comorbidity and Its Impact on Disability. *The Milbank Quarterly*, 67(3/4):450.

Karin Verspoor, Judith Cohn, Susan Mniszewski, and Cliff Joslyn. 2006. A categorization approach to automated ontological function annotation. *Protein Science*, 15(6):1544–1549.

Daniel J Vreeman and Christophe Richoz. 2015. Possibilities and implications of using the ICF and other vocabulary standards in electronic health records. *Physiotherapy research international : the journal for researchers and clinicians in physical therapy*, 20(4):210–219.

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of Semantic Change of Harm-Related Concepts in Psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Donald E Walker and Robert A Amsler. 1986. The use of machine-readable dictionaries in sublanguage analysis. In Ralph Grishman and Richard Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 69–83. Lawrence Erlbaum Associates, Hillsdale, NJ.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph and Text Jointly Embedding. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1591–1601, Doha, Qatar. Association for Computational Linguistics.

M Weeber, J G Mork, and a R Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 746–750.

Nicole Gray Weiskopf and Chunhua Weng. 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151.

Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. 2018. Factors Influencing the Surprising Instability of Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102. Association for Computational Linguistics.

Chunhua Weng and Peter J Embi. 2019. Informatics Approaches to Participant Recruitment. In Rachel L Richesson and James E Andrews, editors, *Clinical Research Informatics*, pages 109–122. Springer International Publishing, Cham.

Brendan Whitaker, Denis Newman-Griffis, Aparajita Haldar, Hakan Ferhatosmanoglu, and Eric Fosler-Lussier. 2019. Characterizing the impact of geometric properties of word embeddings on task performance. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 8–17, Minneapolis, MN. Association for Computational Linguistics.

Mary C White, Frances Babcock, Nikki S Hayes, Angela B Mariotto, Faye L Wong, Betsy A Kohler, and Hannah K Weir. 2017. The history and use of cancer registry data by public health cancer control programs in the United States. *Cancer*, 123(S24):4969–4976.

World Health Organization. 1980. *International Classification of Impairments, Disabilities, and Handicaps*. World Health Organization, Geneva.

World Health Organization. 2001. *International Classification of Functioning, Disability and Health: ICF*. World Health Organization, Geneva.

World Health Organization. 2013. *How to Use the ICF: A practical manual for using the International Classification of Functioning, Disability and Health (ICF)*. World Health Organization, Geneva.

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, Genevieve Gorrell, Angus Roberts, Matthew Broadbent, Robert Stewart, and Richard J B Dobson. 2018. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research*. *Journal of the American Medical Informatics Association*, 25(5):530–537.

Yonghui Wu, Joshua C Denny, S Trent Rosenbloom, Randolph A Miller, Dario A Giuse, Lulu Wang, Carmelo Blanquicett, Ergin Soysal, Jun Xu, and Hua Xu. 2017. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). *Journal of the American Medical Informatics Association*, 24(e1):e79–e86.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. In *arXiv preprint arXiv:1609.08144*.

Long Xia, G Alan Wang, and Weiguo Fan. 2017. A Deep Learning Based Named Entity Recognition Approach for Adverse Drug Events Identification and Extraction in Health Social Media. In *Smart Health*, pages 237–248, Cham. Springer International Publishing.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short Text Clustering via Convolutional Neural Networks.

In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, Denver, Colorado. Association for Computational Linguistics.

Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors. 2019. *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, Hong Kong, China.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. 2016. Revisiting Semi-Supervised Learning with Graph Embeddings. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 40–48, New York, New York, USA. PMLR.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Cheng Ye and Daniel Fabbri. 2018. Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. *Journal of Biomedical Informatics*, 83:63–72.

Wenpeng Yin and Hinrich Schütze. 2014. An Exploration of Embeddings for Generalized Phrases. In *Proceedings of the {ACL} 2014 Student Research Workshop*, pages 41–47, Baltimore, Maryland, USA. Association for Computational Linguistics.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised Word Sense Disambiguation with Neural Models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1374–1385. The COLING 2016 Organizing Committee.

Darin B Zahuranec, Lesli E Skolarus, Chunyang Feng, Vicki A Freedman, and James F Burke. 2017. Activity limitations and subjective well-being after stroke. *Neurology*, 89(9):944–950.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag : A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of*

the 2018 Conference on Empirical Methods in Natural Language Processing, page In Press, Brussels, Belgium. Association for Computational Linguistics.

Qing T Zeng, Doug Redd, Guy Divita, Samah Jarad, Cynthia Brandt, and Jonathan R Nebeker. 2011. Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes. *Journal of Health & Medical Informatics*, S3.

Michael Zhai, Johnny Tan, and Jinho D Choi. 2016. Intrinsic and Extrinsic Evaluations of Word Embeddings. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 4282–4283. AAAI Press.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1):52.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Wanshan Zheng, Zibin Zheng, Hai Wan, and Chuan Chen. 2019. Dynamically Route Hierarchical Structure Representation to Attentive Capsule for Text Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, {IJCAI-19}*, pages 5464–5470. International Joint Conferences on Artificial Intelligence Organization.

Ayah Zirikly, Varun Kumar, and Philip Resnik. 2016. The GW/UMD CLPsych 2016 Shared Task System. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 166–170, San Diego, CA, USA. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33.