Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility

Denis Newman-Griffis[1,2] and **Ayah Zirikly**[1]
[1]Rehabilitation Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD
[2]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH
{denis.griffis, ayah.zirikly} @nih.gov

Abstract

Functioning is gaining recognition as an important indicator of global health, but remains under-studied in medical natural language processing research. We present the first analysis of automatically extracting descriptions of patient mobility, using a recently-developed dataset of free text electronic health records. We frame the task as a named entity recognition (NER) problem, and investigate the applicability of NER techniques to mobility extraction. As text corpora focused on patient functioning are scarce, we explore domain adaptation of word embeddings for use in a recurrent neural network NER system. We find that embeddings trained on a small in-domain corpus perform nearly as well as those learned from large out-of-domain corpora, and that domain adaptation techniques yield additional improvements in both precision and recall. Our analysis identifies several significant challenges in extracting descriptions of patient mobility, including the length and complexity of annotated entities and high linguistic variability in mobility descriptions.

1 Introduction

Functioning has recently been recognized as a leading world health indicator, joining morbidity and mortality (Stucki and Bickenbach, 2017). Functioning is defined in the International Classification of Functioning, Disability, and Health (ICF; WHO 2001) as the interaction between health conditions, body functions and structures, activities and participation, and contextual factors. Understanding functioning is an important element in assessing quality of life, and automatic extraction of patient functioning would serve as a useful tool for a variety of care decisions, including rehabilitation and disability assessment (Stucki et al., 2017). In healthcare data, natural language processing (NLP) techniques have been successfully used for retrieving information about health conditions, symptoms and procedures from unstructured electronic health record (EHR) text (Soysal et al., 2018; Savova et al., 2010). As recognition of the importance of functioning grows, there is a need to investigate the application of NLP methods to other elements of functioning.

Recently, 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 suffi-

cient data to train robust word representations.

In this paper, 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.

2 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 (2015). 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 1: Synthetic document with examples of ScoreDefinition (in blue) and Mobility (in orange).

3 Data

Thieu et al. (2017) presented a dataset of 250 deidentified 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: descriptions of mobility status, which we call 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., 2012) include fairly short entities (2.1 tokens on average for i2b2), Mobility entities

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

Table 1: Named entity statistics for training, validation, and test splits of BTRIS-Mobility. 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.

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 phrases that are the meat of most NER. Figure 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 1.

3.1 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.[1] A large proportion of these documents comes from the Rehabilitation Medicine Department of the NIH Clinical Center, including Physical Therapy (PT), 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.

3.1.1 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.[2]

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., 2016); 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.

4 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 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 de-

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

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

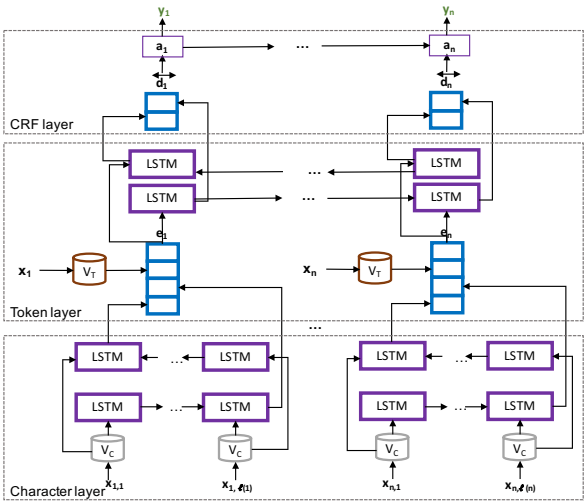

Figure 2: Bi-LSTM-CRF network architecture

scent (SGD) which showed superior performance to adaptive moment estimation (Adam) optimization technique (Kingma and Ba, 2014).

4.1 Embedding training

We use two popular toolkits for learning word embeddings: word2vec[3] (Mikolov et al., 2013) and FastText[4] (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.[5]

4.2 Domain adaptation methods

We evaluate several different methods for adapting out-of-domain embeddings to the BTRIS corpus.

Concatenation In addition to the original embeddings, 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 Recent work has found benefits from retraining learned embeddings on a target corpus (Yang et al., 2017). We pre-initialize both word2vec and FastText toolkits with embeddings learned on each of our three reference corpora, and retrain on the BTRIS corpus using an initial learning rate of 0.1. Additionally, we use the regularization-based domain adaptation approach introduced by Yang et al. (2017) as another baseline, due to its successful results in improving NER performance. Their method aims to help the model to differentiate between general and domain specific terms, using a significance function ϕ of a word w. ϕ is dependent on the definition of w's frequency, where in our implementation it is the word frequency in the target corpora.

Linear transform However, these approaches suffer from the same limitations as training BTRIS embeddings directly: a restricted vocabulary and minimal training data, both due to the size of the corpus. We therefore also investigate two methods for learning a transformation from one set of embeddings into the same space as another, based on a reference dictionary. Given an out-of-domain source embedding set and a target BTRIS embedding set, we use all words in common between source and target as our training vocabulary.[6] We adapt this to the linear transformation method successfully applied to bilingual embeddings by Artetxe et al. (2016), using this shared vocabulary as the training dictionary.

Non-linear transform As all of our embeddings are in English, but from domains that do not intuitively seem to have a linear relationship, we also extend the method of Artetxe et al. to a non-linear transformation. We randomly divide the shared vocabulary into ten folds, and train a feed-forward neural network using nine-tenths of the data, minimizing mean squared error (MSE) between the learned projection and the true embeddings. After each epoch, we calculate MSE on the held-out set, and halt when this error stops decreasing. Finally, we average the learned projections from each fold to yield the final transformation function. Following Artetxe et al. (2016), we apply this function to all source embeddings, allowing us to maintain the original vocabulary size.

Our model is a fully-connected feed-forward neural network, with the same hidden dimension as our embeddings. We evaluate with both 1 and 5 hidden layers, and use either tanh or rectified linear unit (ReLU) activation throughout. Model structure is denoted in the result; for example, "5-layer ReLU" refers to nonlinear mapping using a 5-layer network with ReLU activation. We train with Adam optimization (Kingma and Ba, 2014) and a minibatch size of 5.[7]

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

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

[5] 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.

[6] We evaluated using subsets of 1k, 2k, or 10k shared words most frequent in BTRIS, but the best downstream performance was achieved using all pivot points.

[7] Source implementation available at `github.com/drgriffis/NeuralVecmap`

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 2: Exact and token-level match results on BTRIS-Mobility, using 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.

5 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 3), 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 $_{w2v}$.

5.1 Embedding corpora

Exact and token-level match results for both Mobility and ScoreDefinition entities are given for embeddings from each corpus in Table 2. By and large, the in-domain BTRIS and PT-OT embeddings yield higher precision than out-of-domain embeddings, though this comes at the expense of recall. word2vec embeddings consistently achieve better NER performance than FastText embeddings from the clinical corpora, although this was reversed with PubMed, suggesting that further research is needed on the strengths of different embedding methods in biomedical data. The unusually 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).

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. 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 domain.

Token-level results follow the same trends as exact match, with clinical embeddings achieving highest precision, while PubMed embeddings yield better recall. As many entity-level errors are only off by a few tokens, token-level scores are generally 15-20 absolute points higher than their corresponding entity-level scores. At the token level, it is clear that ScoreDefinition entities are effectively solved in this dataset, with all F1 scores are above 97.4%. This is primarily due to the regularity of ScoreDefinition strings: they typically consist of a sequence of single numbers followed by explanatory strings, as shown in Figure 1.

5.2 Mapping methods

Table 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

Target	Source	Concat			Preinit			Linear			5-layer tanh		
		Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1
BTRIS$_{FT}$	WikiNews$_{FT}$	**72.2**	**65.3**	**68.6**	55.0	59.2	57.0	65.1	61.9	63.5	69.3	64.2	66.7
	PubMed$_{FT}$	**69.5**	65.8	**67.6**	64.2	**66.5**	65.4	65.6	60	62.7	66.1	64.5	65.3
	PubMed$_{w2v}$	65.3	65.3	**65.3**	64.8	**65.4**	65.1	70.3	65.8	68	**66.3**	62.6	64.4
	MIMIC$_{FT}$	35.0	10.4	16.0	37.8	15.5	22.0	63.7	**62.9**	63.3	**70.3**	61.3	**65.5**
	MIMIC$_{w2v}$	67.4	**67.6**	**67.5**	68.5	64.6	66.5	66.8	60.3	63.4	**69.2**	64.3	66.7
PT-OT$_{FT}$	WikiNews$_{FT}$	67.5	**63.9**	65.6	54.5	57.9	56.1	68.9	63.8	66.2	**68.5**	63.4	**65.8**
	PubMed$_{FT}$	62.8	**65.1**	**63.9**	61.3	50.2	55.2	62.6	62.6	62.6	**68.3**	60.1	**63.9**
	MIMIC$_{w2v}$	64.1	**66.1**	**65.1**	59.9	61.8	60.8	57.9	54.1	55.9	**67.3**	63.2	**65.1**

Table 4: Exact match precision and recall for Mobility entities with word embeddings mapped from each source to BTRIS$_{FT}$ embeddings, using four selected domain adaptation methods. The best-performing embeddings from each source corpus were also mapped to PT-OT$_{FT}$ embeddings. The best precision, recall, and F1 achieved with each source/target pair is marked in bold.

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 3: Comparison of mapping methods, using 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.

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 percentage points.

These differences are fleshed out further in Table 4, comparing four domain adaptation methods across several source/target pairs. Concatenation typically achieves the best overall performance among the adaptation methods, but nonlinear mappings yield highest precision in 6 of the 8 settings shown. Concatenation is also more sensitive to noise in the source embeddings, as shown with MIMIC$_{FT}$ results, and preinitialization varies widely in its performance. By contrast, linear and nonlinear mapping methods are less affected by the choice of source embeddings, yielding more consistent results than preinitialization or concatenation for a given target corpus. Nonlinear mappings exhibit this stability most clearly, producing very similar results across all settings. The

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 5: Best precision, recall, and F1 (exact) for test set Mobility mentions, with the source/target pair and domain adaptation method used.

regularization-based domain adaptation method of Yang et al. (2017) consistently yielded similar results to preinitialization: for example, an F1 score of 65% when PubMed$_{w2v}$ embeddings are adapted to BTRIS, as compared to 65.4% using pre-initialization with word2vec. We therefore omit these results for brevity.

Comparing both Tables 3 and 4 to the performance of unmodified embeddings shown in Table 2, we see a surprising lack of overall performance improvement or degradation. While the different adaptation methods exhibit consistent differences between one another, only 12 of the 32 F1 scores in Table 4 represent improvements over the relevant unmapped baselines. Many adaptation results achieve notable improvement in precision or recall individually, suggesting that different methods may be more useful for downstream applications where one metric is emphasized over the other. However, several of our results indicate failure to adapt, illustrating the difficulty of effectively adapting embeddings for this task.

5.3 Source/target pairs

Table 5 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.

There was no clear trend in the domain-adapted results as to whether word2vec or FastText embeddings led to the best downstream performance: it varied between pairs and adaptation methods. word2vec embeddings were generally more consistent, but as seen in Tables 4 and 5, FastText embeddings often achieved the highest performance.

5.4 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:[8]

```
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 5.1.

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`.

5.4.1 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 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 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`.

5.4.2 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.

5.4.3 Adaptation method effects

The most striking difference we observe when comparing different domain adaptation methods is that preinitialization universally leads to longer

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

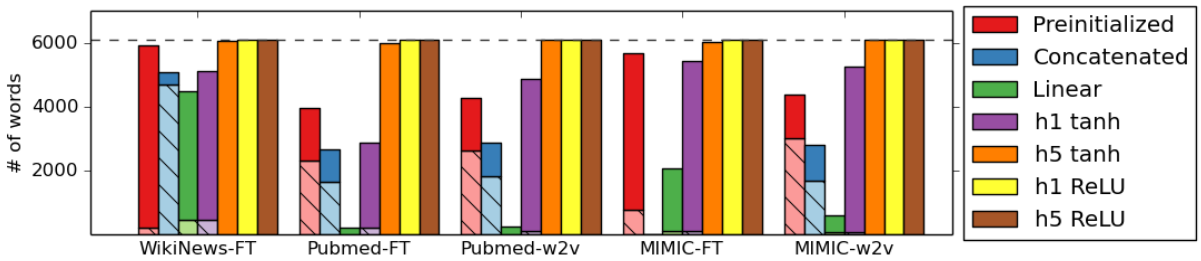

Figure 3: Number of words in shared vocabulary with different nearest neighbors in source and domain-adapted embeddings, using BTRIS$_{FT}$ as target. Light hatched bars indicate the number of words whose new nearest neighbor matches BTRIS$_{FT}$. The dashed line indicates shared vocabulary size.

Source set	Source	Target	Preinit	Concat	Linear	h1 tanh	h5 tanh
PubMed$_{FT}$	ambulating ambulate crutches	ambulating ambulate ambulatory	ambulating ambulate walker	ambulating ambulate crutches	ambulating ambulate crutches	ambulating ambulate crutch	worsening wearing complaints
WikiNews$_{FT}$	ambulating ambulate extubation	ambulating ambulate ambulatory	pos 76 acuity	ambulating ambulate ambulatory	cardiopulmonary neurosurgical resuscitation	robotic overhead ambulating	respiratory sclerotic acupuncture

Table 6: Top 3 nearest neighbors of *ambulation* in embeddings mapped to BTRIS$_{FT}$ using different adaptation methods. Source and Target are neighbors in the original source and BTRIS$_{FT}$ embeddings.

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 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 PTOT$_{w2v}$). 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 Embedding analysis

To further evaluate the effects of different domain adaptation methods, we analyzed the nearest neighbors by cosine similarity of each word before and after domain adaptation. We only considered the words present both in the dataset and in each of our original sets of embeddings, yielding a vocabulary of 6,201 words. We then took this vocabulary and calculated nearest neighbors within it, using each set of out-of-domain original embeddings and each of its domain-adapted transformations.

Figure 3 shows the number of words whose nearest neighbors changed after adaptation, using BTRIS$_{FT}$ as the target; all other targets display similar results. We see that in general, the neighborhood structure of target embeddings is well-preserved with concatenation, sometimes preserved with preinitialization, and completely disposed of with the nonlinear transformation. Interestingly, this reorganization of words to something different from both source and target does not lead to the performance degradation we might expect, as shown in Section 5.

We also qualitatively examined nearest neighbors before and after adaptation. Table 6 shows nearest neighbors of *ambulation*, a common Mobility word, for two representative source/target pairs. Preinitialization generally reflects the neighborhood structure of the target embeddings,

but can be noisy: in WikiNews$_{FT}$/BTRIS$_{FT}$, other words such as *therapy* and *fatigue* share *ambulation*'s less-than-intuitive neighbors.

Reflecting the changes seen in Figure 3, the linear transformation preserves source neighbors in the biomedical PubMed corpus, but yields a neighborhood structure different from source or target with highly out-of-domain WikiNews embeddings. Nonlinear transformations sometimes yield sensible nearest neighbors, as in the single-layer tanh mapping of PubMed$_{FT}$ to BTRIS$_{FT}$. More often, however, the learned projection significantly shuffles neighborhood structure, and observed neighbors may bear only a distant similarity to the query term. In several cases, large swathes of the vocabulary are mapped to a single tight region of the space, yielding the same nearest neighbors for many disparate words. This occurs more often when using a ReLU activation, but we also observe it occasionally with tanh activation.

7 Conclusions

We have conducted an experimental analysis of recognizing descriptions of patient mobility with a recurrent neural network, and of the effects of various domain adaptation methods on recognition performance. We find 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 often improve recognition performance over both in- and out-of-domain baselines, though such improvements are difficult to achieve consistently. Simpler methods such as preinitialization and concatenation achieve better performance gains, but are also susceptible to noise in source embeddings; more complex methods yield more consistent performance, but with practical downsides such as decreased recall and a non-intuitive projection of the embedding space. 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).

To our knowledge, this is the first investigation into automatically recognizing descriptions of patient functioning. 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. It is our hope that these initial findings, along with further research refining the appropriate framework for representing and approaching the recognition problem, will spur further research into this complex and important domain.

Acknowledgments

The authors would like to thank Elizabeth Rasch, Thanh Thieu, and Eric Fosler-Lussier for helpful discussions, the NIH Biomedical Translational Research Information System (BTRIS) for their support, and our anonymous reviewers for their invaluable feedback. This research was supported in part by the Intramural Research Program of the National Institutes of Health, Clinical Research Center and through an Inter-Agency Agreement with the US Social Security Administration.

References

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the ACL*, 5:135–146.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to Train Good Word Embeddings for Biomedical NLP. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174.

Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

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(PART 1):1299–1303.

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(Aug):2493–2537.

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.

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.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

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.

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.

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.

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-Volume 3*, pages 1–8. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

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.

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-Volume 4*, pages 188–191. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, pages 1–12.

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(August):btw529.

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 : JAMIA*, 17(5):507–513.

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*, pages 104–107. Association for Computational Linguistics.

Steven J Skube, Elizabeth A Lindemann, Elliot G Arsoniadis, Elizabeth C Wick, and Genevieve B Melton. 2018. Characterizing Functional Health Status of Surgical Patients in Clinical Notes. In *2018 AMIA Summit on Clinical Research Informatics*. American Medical Informatics Association.

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.

G Stucki, J Bickenbach, and J Melvin. 2017. Strengthening Rehabilitation in Health Systems Worldwide by Integrating Information on Functioning in National Health Information Systems. *Am J Phys Med Rehabil*, 96(9):677–681.

Gerold Stucki and Jerome Bickenbach. 2017. 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.

Thanh Thieu, Jonathan Camacho, Pei-Shu Ho, Diane Brandt, Julia Porcino, Denis Newman-Griffis, Ao Yuan, Min Ding, Lisa Nelson, Elizabeth Rasch, Chunxiao Zhou, Albert M Lai, and Leighton Chan. 2017. Inductive identification of functional status information and establishing a gold standard corpus A case study on the Mobility domain. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2300–2302.

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.

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. In *BMC Bioinformatics*, volume 7, page S11. BioMed Central.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2012. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–6.

WHO. 2001. *International Classification of Functioning, Disability and Health: ICF*. World Health Organization.

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.

Wei Yang, Wei Lu, and Vincent Zheng. 2017. A simple regularization-based algorithm for learning cross-domain word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2898–2904.