Inductive identification of functional status information and establishing a gold standard corpus

A case study on the Mobility domain

Thanh Thieu[*‡], Jonathan Camacho[*], Pei-Shu Ho[*],
Julia Porcino, Min Ding, Lisa Nelson,
Elizabeth Rasch, Chunxiao Zhou, Leighton Chan[§]
Rehabilitation Medicine Department
National Institutes of Health Clinical Center
Bethesda, MD, USA
thanh.thieu@nih.gov

Denis Newman-Griffis[◊]
Computer Science and Engineering Department
Ohio State University
Columbus, OH, USA

Diane Brandt
Social Security Advisory Board
Washington, DC, USA

Ao Yuan[◊]
Biostatistics, Bioinformatics and Biomathematics Dept.
Georgetown University
Washington, DC, USA

Albert M. Lai[◊]
Institute for Informatics
Washington University
St. Louis, MO, USA

* These authors contributed equally to this work
‡ Corresponding author
◊ Also affiliated with the NIH Clinical Center
§ Principal investigator

Abstract—**The importance of functional status information (FSI) has become increasingly evident in recent years [1, 2]. However, implementation, application, and normalization of FSI in health care and Electronic Health Records (EHRs) have been largely underexplored. The World Health Organization's International Classification of Functioning, Disability and Health (ICF) [3] is considered to be the international standard for describing and coding function and health states. Nevertheless, the ICF provides only a limited vocabulary for recognizing FSI descriptions, since its purpose is to organize concepts related to functioning rather than to provide a comprehensive terminology or a complete set of relations between concepts. While the free text portion of EHRs might provide a more complete picture of health status, treatment, and progress, current Natural Language Processing (NLP) methods largely focus on extracting medical conditions (e.g. diagnoses and symptoms, etc.). The absence of a standardized functional terminology and incompleteness of the ICF as a vocabulary source makes it challenging to build a NLP system to extract FSI from EHR free text.**

Our work takes the first step towards extraction of FSI from free text by systematically identifying the structure of FSI related to Mobility, a key domain of the ICF and an important domain in the determination of work disability. Our interdisciplinary research group inductively evaluated examples extracted from over 1,200 Physical Therapy (PT) notes from the Clinical Center of the National Institutes of Health (NIH). This extensive work resulted in a nested entity structure comprised of 2 entities, 3 sub-entities, 8 attributes, and 21 attribute values. Furthermore, we have manually curated the first gold standard corpus of 200 double-annotated and 50 triple-annotated PT notes. Our inter-annotator agreement (IAA) averages 97% F1-score on partial textual span matching and from 0.4 to 0.9 Siegel & Castellan's kappa on attribute value matching. Such a rich semantic corpus of Mobility FSI is valuable and a promising resource for future statistical learning. Our method is also adaptable to other domains of the ICF.

Keywords-functional status information; functioning; ICF; natural language processing; manual curation; annotation; physical therapy

I. INTRODUCTION

Free text in Electronic Health Records (EHRs) provides not only medical information (e.g. diseases and disorders), but is also rich in patient's functional status information (FSI). FSI is an important component of healthcare that has been under-developed when utilizing EHRs for clinical, administrative, and research purposes. To more fully understand the influence of health conditions and impairments in daily life, the interaction between functional ability and environmental demand should be considered. It is also possible that two patients with similar diagnostic codes (e.g. ICD-9 codes) may exhibit different levels of functioning.

The WHO's ICF [3] describes function from the whole person perspective and includes many interrelated

components (body functions and structures, activities and participation, and environmental factor). To tackle the complexity of this model, we started by assessing the Mobility domain, part of the Activities and Participation component of the ICF and an important and relatively observable aspect of functioning. In this work, we identified a semantic structure of Mobility information, performed manual annotation on the entities, and assigned either ICF codes or semantic values to related attributes. To the best of our knowledge, our work is the first attempt at comprehensive, semantic analysis of a domain of the ICF, accompanied by a gold standard corpus for future automatic extraction of FSI.

II. Related work

Sophisticated NLP methods often rely on high-quality, human-annotated datasets as gold standard corpora to train a statistical model on a specific Information Extraction (IE) task. There have been efforts to create human-annotated datasets in biomedical and occupational disciplines.

Bada [4] created the Colorado Richly Annotated Full-Text (CRAFT) corpus by manually annotating biomedical entities and relations from 97 full-text biomedical articles. They used all terms from selected vocabularies of the Open Biomedical Ontologies (OBO) library. Lindemann [5] relied on the Occupational Data for Health (ODH) data model to annotate 868 sentences for occupational information from six clinical sources. They derived an annotation schema comprised of six types of occupational entities and sub-entities, and measured inter-rater reliability on 10% of the sentences. Kuang [6] attempted to match patient-reported functional status terms modeled on Short Form-36 Health Survey (SF-36) to the Unified Medical Language System (UMLS). They manually collected 2,763 terms from 800 clinical documents from the Veterans Administration Informatics and Computing Infrastructure (VINCI) database and 350 posts from three online discussion forums for cardiovascular diseases and atrial fibrillation.

It was through meticulous human annotation that the author groups identified shortcomings of either their target ontology or terminology.

Bada [4] encountered various difficulties while using the OBOs for sematic annotation, so they recommended six high-level desiderata that address overlapping terms, ambiguities, mid-level ontologies, non-canonical instances, and expansion of relations. Lindemann [5] noticed that some additions to the ODH model were needed to fully represent the broad spectrum of occupational information in clinical text. Kuang [6] showed that most functional terms (85.9%) did not have exact UMLS matches, while the partial matches did not capture the terms' exact semantics.

Regarding function and health, there were fewer efforts to identify FSI, and they were limited to a few specific ICF codes or dependent on ad-hoc mapping tables to circumvent the absence of FSI terminology.

Mahmoud [7] constructed a MySQL database to pair ICF codes with descriptions. Users then chose the description of a patient to retrieve the corresponding ICF codes from the database. To supplement the simplicity of their ICF code retrieval system, the authors manually compared Functional Independence Measure (FIM) scores to the retrieved ICF qualifiers. They found that FIM scores were comparable to the average of retrieved ICF qualifiers before and after rehabilitation. Kukafka [8] investigated automatic coding of five ICF codes using the rule-based MedLEE [9] NLP system. The five ICF codes were selected from Mental Functions, Mobility, and Self-Care domains of the ICF. Free text data were acquired from 251 inpatient rehabilitation discharge summaries. The system assigns an ICF code with performance (first digit after the decimal point), and capacity (second digit after the decimal point) qualifiers to each sentence. To do so, it first mapped the language in the sentence to a normalized language using a lexicon table, and then mapped the normalized language to an ICF code with qualifiers using an encoding table. Results showed that automatic encoding of ICF codes is better than non-expert human coding, but worse than expert human coding.

Our work differs from previous studies in several dimensions: (1) We identify textual span of FSI entities in the context of clinical notes, thus avoiding the fragmentation issue of sentence segmentation; (2) We capture the phrase linked to an ICF code in a separate Action entity; (3) We clarify the performance qualifier by splitting it into two sub-entities called Assistance and Quantification; and (4) We utilize all three-digit ICF codes of the Mobility domain of Activities and Participation component. As a result of this comprehensive approach, our method is generalizable to other domains of the ICF.

III. Method

An interdisciplinary research group consisting of scientists in physical therapy, rehabilitation, health services, medicine, statistics, and computer science identified the most salient components of Mobility functional information across clinical, administrative, and statistical learning uses. Free text data were collected from Physical Therapy (PT) medical records provided by the Office of Biomedical Translational Research Information System (BTRIS) [10], an intramural resource at the National Institutes of Health (NIH) that stores clinical research data. For annotation purpose, we focused on PT notes only, since they contain the most amount of mobility related information.

While evaluating examples extracted from over 1,200 PT notes, we modeled several frameworks and coding schemes including the predicate-argument structure of a sentence, the performance and capacity qualifiers of the ICF, and the Functional Independence Measure (FIM). Learning from the pros and cons of each model, we synthesized a semantic entity framework that consistently and objectively captures the structure of Mobility functional information. A Mobility entity (a self-contained, well-defined description of physical functional status information) thus encapsulates three types of sub-entities: (1) Action entity captures information about the activities, (2) Assistance entity captures information about dependence on another person or object when performing the activity, and (3) Quantification entity captures information regarding measurement values of the activity. In addition, the framework defines Score Definition

entity to capture specific measurement scales related to Mobility activity. Each type of entity has a set of attributes and corresponding values that reflect the semantics of the context in which the entity instance was identified.

In parallel with identifying the structure of Mobility functional information, we created detailed annotation guidelines and later used the guidelines to manually curate 250 PT notes as the first gold standard corpus of FSI. Of these 250 notes, 200 were double annotated and 50 were triple annotated.

We selected GATE developer [11], an open source text processing software, as a tool for semantic annotation and created an annotation schema comprised of entity types and attributes identified in the Mobility entity framework.

IV. Results

The Mobility entity framework consists of component entities associated with attributes that provide details of the context, such as timing of the activity, performing agents, and 3-digit ICF codes. Statistically, Mobility entity has 3 attributes with 7 values, Action entity has 2 attributes with 16 values, Assistance entity has 2 attributes with 6 values, Quantification entity has 1 attribute with 4 values, and Score Definition entity does not have any attribute.

Within the gold standard corpus, Table I presents average number of entities across annotators, and Table II presents inter-annotator agreement (IAA) computed as F1 score [12] of the overlap of textual span of entities.

V. Discussion

Our inductive method to identify the structure of FSI is currently only validated on the Mobility domain of the ICF. Such a manual process is labor-intensive, but once mastered, can be applied to other domains of the ICF. A topic of future research is to automatically identify salient components of an unknown functional domain, and use them as a starting point for human experts to refine. Another research direction is to utilize active learning to produce quality annotation instead of the labor-intensive quantity annotation.

TABLE I. Number of entities

Sub-dataset	Number of entities				
	Mob	***Act***	***Asst***	***Quant***	***Sdef***
200 dbl	2318	2216	1333	931	126
50 trp	634	644	343	305	31

TABLE II. Inter-annotator agreement (IAA)

Sub-dataset	F1 score of overlapping textual span				
	Mob	***Act***	***Asst***	***Quant***	***Sdef***
200 dbl	0.980	0.980	0.960	0.982	0.980
50 trp	0.964	0.954	0.945	0.978	1

"dbl" = double-annotated; "trp" = triple-annotated; "Mob" = Mobility; "Act" = Action; "Asst" = Assistance; "Quant" = Quantification; "Sdef" = Score Definition

VI. Conclusion

We developed an inductive process to identify the structure of FSI applied to Mobility domain of the ICF. Based on the Mobility structure, we created extensive annotation guidelines and manually curated the first gold standard corpus. The semantic corpus is suitable for future machine learning and our inductive method is adaptable to other ICF domains.

Acknowledgment

This work was supported by the Intramural Research Program of the National Institutes of Health and the US Social Security Administration.

References

[1] Hopfe, M., B. Prodinger, J.E. Bickenbach, and G. Stucki, "Optimizing health system response to patient's needs: an argument for the importance of functioning information". Disabil Rehabil, 2017: p. 1-6.

[2] Stucki, G. and J. Bickenbach, "Functioning: the third health indicator in the health system and the key indicator for rehabilitation". Eur J Phys Rehabil Med, 2017. 53(1): p. 134-8.

[3] WHO, "International Classification of Functioning, Disability and Health (ICF)". 2001, Geneva: World Health Organization.

[4] Bada, M. and L. Hunter, Desiderata for ontologies to be used in semantic annotation of biomedical documents, in J Biomed Inform. 2011, 2010 Elsevier Inc: United States. p. 94-101.

[5] Lindemann, E.A., E.S. Chen, S. Rajamani, N. Manohar, Y. Wang, and G.B. Melton. "Representation of occupation information in clinical texts: an analysis of free-text clinical documentation in multiple sources" in AMIA Joint Summits on Translational Science. 2017. San Francisco.

[6] Kuang, J., A.F. Mohanty, V.H. Rashmi, C.R. Weir, B.E. Bray, and Q. Zeng-Treitler, "Representation of functional status concepts from clinical documents and social media sources by standard terminologies". AMIA Annual Symposium Proceedings, 2015. 2015: p. 795-803.

[7] Mahmoud, R., N. El-Bendary, H.M.O. Mokhtar, and A.E. Hassanien, "ICF based automation system for spinal cord injuries rehabilitation". 2014 9th International Conference on Computer Engineering & Systems (ICCES), 2014: p. 192-197.

[8] Kukafka, R., M.E. Bales, A. Burkhardt, and C. Friedman, "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 : JAMIA, 2006. 13(5): p. 508-515.

[9] Friedman, C., P.O. Alderson, J.H. Austin, J.J. Cimino, and S.B. Johnson, "A general natural-language text processor for clinical radiology". J Am Med Inform Assoc, 1994. 1(2): p. 161-74.

[10] Cimino, J.J., E.J. Ayres, L. Remennik, S. Rath, R. Freedman, A. Beri, et al., "The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date". J Biomed Inform, 2014. 52: p. 11-27.

[11] Cunningham, H., D. Maynard, and K. Bontcheva, "Text processing with GATE". 2011: Gateway Press CA. 588.

[12] Hripcsak, G. and A.S. Rothschild, "Agreement, the F-Measure, and Reliability in Information Retrieval". Journal of the American Medical Informatics Association : JAMIA, 2005. 12(3): p. 296-298.