Article:
Thieu, T., Maldonado, J.C., Ho, P.-S. et al. (2021) A comprehensive study of mobility functioning information in clinical notes: Entity hierarchy, corpus annotation, and sequence labeling. International Journal of Medical Informatics, 147. 104351. ISSN: 1386-5056

https://doi.org/10.1016/j.ijmedinf.2020.104351

International Journal of Medical Informatics 147 (2021) 104351

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

A comprehensive study of mobility functioning information in clinical notes: Entity hierarchy, corpus annotation, and sequence labeling

Thanh Thieu [a,*,1], Jonathan Camacho Maldonado [b,1], Pei-Shu Ho [b,1], Min Ding [c], Alex Marr [b], Diane Brandt [d], Denis Newman-Griffis [b,e], Ayah Zirikly [b], Leighton Chan [b], Elizabeth Rasch [b]

[a] Oklahoma State University, Stillwater, OK, United States
[b] National Institutes of Health Clinical Center, Bethesda, MD, United States
[c] National Institute of Standards and Technology, Gaithersburg, MD, United States
[d] Social Security Advisory Board, Washington, DC, United States
[e] Ohio State University, Columbus, OH, United States

ARTICLE INFO

Keywords:
Functioning information
Mobility
Clinical notes
Natural language processing
Text mining
Named entity recognition

ABSTRACT

Background: Secondary use of Electronic Health Records (EHRs) has mostly focused on health conditions (diseases and drugs). Function is an important health indicator in addition to morbidity and mortality. Nevertheless, function has been overlooked in accessing patients' health status. The World Health Organization (WHO)'s International Classification of Functioning, Disability and Health (ICF) is considered the international standard for describing and coding function and health states. We pioneer the first comprehensive analysis and identification of functioning concepts in the Mobility domain of the ICF.

Results: Using physical therapy notes at the National Institutes of Health's Clinical Center, we induced a hierarchical order of mobility-related entities including 5 entities types, 3 relations, 8 attributes, and 33 attribute values. Two domain experts manually curated a gold standard corpus of 14,281 nested entity mentions from 400 clinical notes. Inter-annotator agreement (IAA) of exact matching averaged 92.3 % F1-score on mention text spans, and 96.6 % Cohen's kappa on attributes assignments. A high-performance Ensemble machine learning model for named entity recognition (NER) was trained and evaluated using the gold standard corpus. Average F1-score on exact entity matching of our Ensemble method (84.90 %) outperformed popular NER methods: Conditional Random Field (80.4 %), Recurrent Neural Network (81.82 %), and Bidirectional Encoder Representations from Transformers (82.33 %).

Conclusions: The results of this study show that mobility functioning information can be reliably captured from clinical notes once adequate resources are provided for sequence labeling methods. We expect that functioning concepts in other domains of the ICF can be identified in similar fashion.

1. Introduction

1.1. Overview

Clinical natural language processing (NLP) has been well-explored on three application areas: disease studies, drug-related studies, and workflow optimization [1]. Community shared tasks such as i2b2/n2c2 challenges [2–11], CLEF eHealth [12–19], and SemEval [20–23] addressed various NLP questions including de-identification, concept extraction, and temporal information on disease-specific datasets. However, the analysis of human functioning within medical EHRs, in the presence of health conditions (diseases and drugs) and demands of the environment has been largely un-explored. Function has been increasingly perceived as an important health indicator in addition to mortality and morbidity [24,25].

The World Health Organization (WHO)'s International Classification of Functioning, Disability and Health (ICF) [26] is the international standard for coding function and health states. Components of the ICF (Fig. 1) encompass Body Functions and Structures, tasks performed by an individual (Activities), societal interaction (Participation), and

* Corresponding author at: 230 Math Sciences, Stillwater, OK, 74078, United States.
E-mail address: tthieu@okstate.edu (T. Thieu).
[1] Equal contribution to this work.

https://doi.org/10.1016/j.ijmedinf.2020.104351
Received 11 March 2020; Received in revised form 10 August 2020; Accepted 22 November 2020
Available online 18 December 2020

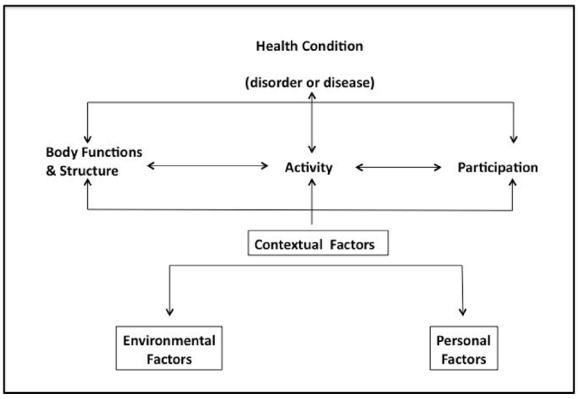

Fig. 1. Diagram of the International Classification of Functioning, Disability and Health (ICF) model of function. Reproduced by permission of World Health Organization (WHO), from ICF [26], p18.

Contextual Factors.

1.2. Objective

Our study focuses on Mobility domain, in Activities and Participation chapter, of the ICF. Mobility is a more well-defined and observable construct of human functioning. Our goal is to set a foundational step to utilize mobility information in clinical NLP [27–31]. We systematically induced an entity hierarchy, annotated a gold standard corpus, and trained NER models. Our work significantly expands the previous, preliminary report [32].

1.3. Existing work on functioning information

Information extraction in clinical text were possible due to standardized vocabulary and annotated corpora [33–35]. In functioning domain, the lack of a standardized ontology [27] and the incompleteness of the ICF as a vocabulary source [36], made existing work to rely on application specific dictionary [37] collected through focus groups [28,30] or manual chart reviews [38]. The lack of annotated resources and a consensus representation of functioning concepts led existing methods to rely on heuristic rules [27–30], manual mapping tables [29], or manual conversion of truncated phrases [39]. Recent work focused on Mobility domain of the ICF and systematically argued for the need to capture and standardize functioning information [40], created an annotated corpus [32], compared word embeddings [41], and classified a coarse qualifier [42].

2. Methods

2.1. Data collection

We sampled 1,554 Physical Therapy (PT) notes from the Rehabilitation and Medicine Department at the NIH Clinical Center using databases of the NIH Biomedical Translational Research Information System [43]. The sample included 950 PT Initial Assessment notes, 320 PT Reassessment notes, 278 PT Assessment and Discharge notes, and 6 PT Discharge notes (Appendix A.3).

2.2. Annotation

An interdisciplinary team comprised of computational linguists, health scientists, and statisticians analyzed components of mobility concepts and developed annotation guidelines similar to a previous work [44]. Among the team, two researchers in health sciences were assigned annotator roles. Annotation process was divided into three phases. In phase 1, a seed batch of 100 PT Initial Assessment notes was analyzed by the interdisciplinary team. At the end of this phase, a hierarchical representation of mobility-related entities was constructed alongside with an initial schema and annotation guidelines. In phase 2, the two annotators consolidated the results of phase 1 on the remaining 1,454 PT notes. In phase 3, consensus annotation was performed to create a gold standard corpus of 400 PT notes. All annotation was done on GATE Developer [45]. (Appendix A.4 and Fig. 2)

2.3. A hierarchy of mobility-related entities

In contrast to past work that either stored functioning phrases as strings [28–30,38] or involved manual conversion [39], we captured a consistent representation of mobility concepts over 1,554 PT notes. Given a sentence “The patient ambulates with modified independence for 300 ft”, the head verb “ambulates” is a predicate modified by two prepositional phrases “with modified independence” and “for 300 ft”. We generalized the predicate to become an Action, accompanied by two types of modifiers: Assistance and Quantification respectively. Such generalization neutralized both grammatical roles and the predicate-argument structure. For example, an Action could be a phrase, and an Assistance or a Quantification could have no association to any specific Action. Generalization allowed the concepts to be flexibly

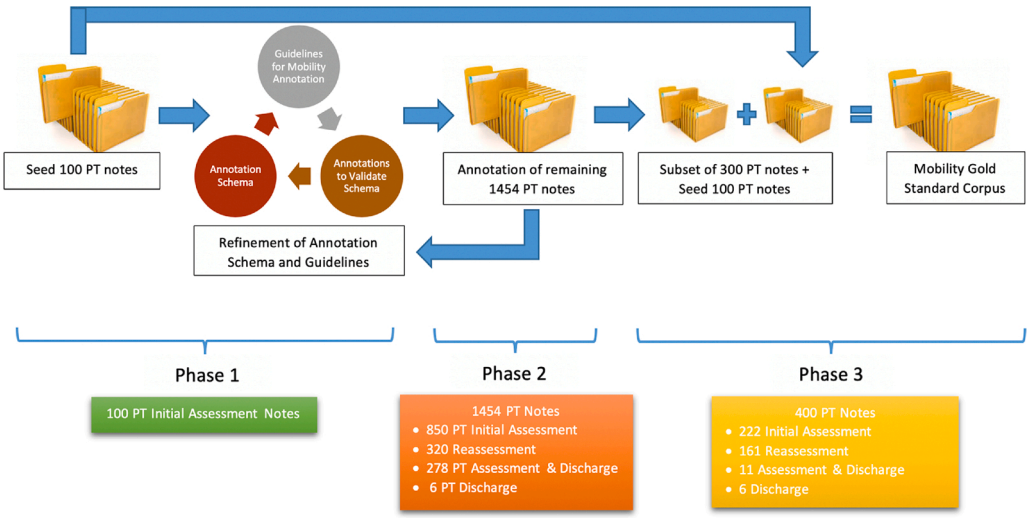

Fig. 2. The annotation process: from guidelines and schema development to creation of the gold standard corpus.

Table 1
Mobility-related entities.

Entity	Definition	*Example*
Mobility	A self-contained, well-defined description of physical functional status information.	*Patient able to ambulate 40 ft. with rolling walker*
Action	Captures the type of activity as well as an individual's ability to perform said activity.	*ambulate*
Assistance	Information about the use and the source of needed assistance (e.g., another person or object) to perform an activity.	*with rolling walker*
Quantification	Information regarding measurement values of the activity.	*40 ft.*
Score Definition	A standardized assessment of functional status. Often represented as numerical values that provide a calibrated scale of functional status	*1=Totally dependent, 2=Requires assistance of a person (with or without appliance), 3=Requires appliances, orthosis or prosthesis for independence, 4=Totally independent (indoors/outdoors)*

conveyed in the complex clinical narratives.

To align with mainstream NLP, we modeled each component as a named entity (Table 1). As a result, Mobility became a nested entity that encapsulated three sub-entities: Action, Assistance, and Quantification. In addition, we observed that Quantification entities occasionally referred to numerical scales either by name (e.g. NIHFA, FIM) or by elaboration in a series of short phrases (Table 1, Score Definition *Example*). We captured these elaborated scales in Score Definition entities. For example, a PT note included: *"(1=dependent, 2=requires person to assist, 3=requires assistive device, 4=independent) Transfers score: 4/4 Ambulation score: 3/4 Wheelchair score: 4/4*". Here, the Quantification scales "4 /4" and "3 / 4" referred to the Score Definition elaborated within the parentheses. In term of implicit relations between entities, we denoted that a Score Definition entity Calibrated subsequent Quantification entities. While within the same Mobility instance, Assistance entities Enabled and Quantification entities Measured the extension of the Action entity. The hierarchy of entities and their implicit relations are presented in Fig. 3 - Entities layer.

In addition, we recorded values of eight types of contextual attributes (Fig. 3, Attributes layer) that accompany the component entities. These attributes provided additional layers of semantics. We captured 3-digit ICF codes because such granularity improved data density and it was widely used in clinical applications relating to health outcome evaluation. Granularity of other attributes were captured at a coarse level to provide grouping and avoid redundancy.

2.4. Analysis and evaluation

We used descriptive statistics to analyze the distribution of annotation results in three annotation phases.

We used F1 score [46] to measure inter-annotator agreement (IAA) of entity mention spans, and Cohen's kappa (κ) [47] to measure agreement of the contextual attributes. We report IAAs on both exact matching similar to CoNLL evaluation [48] and partial matching similar to MUC evaluation [49,50].

2.5. Ensemble identification of nesting mobility-related entity mentions

We split the gold standard corpus (GSC) of 400 annotated notes into five-fold cross-validation. Each fold comprised of 240 notes for training, 80 notes for evaluation, and 80 notes for testing (Fig. 4). Our method included two stages. In stage one, we trained and optimized hyperparameters of three popular, base NER methods (Section 2.5.4) using the

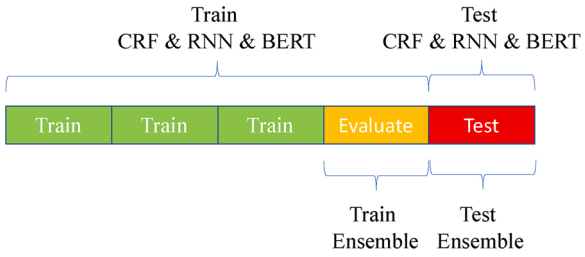

Fig. 4. Distribution of the corpus in one of the five-fold cross-validation. Each rectangle corresponds to 80 clinical notes.

Table 2
Rules of the tokenizing post-processor.

Error Description	*Erroneous Token*	Correction
Splitting combined tokens with concatenators such as forward slash, backward slash, and hyphen.	*"driving/transportation"*	"driving", "/", "transportation"
Splitting abbreviated measure of functioning ability comprising of both letter and digits.	*"x400 ft"*	"x", "400", "ft"
Special PT abbreviation such as "A." denoting mobility assistance at the end of a sentence. The tokenizer mistakenly recognized it as an abbreviated name, thus losing end-of-sentence semantics.	*"A."*	"A", "."
Recognizing end-of-sentence even without a space after a full stop.	*"discuss.The"*	"discuss", ".", "The"

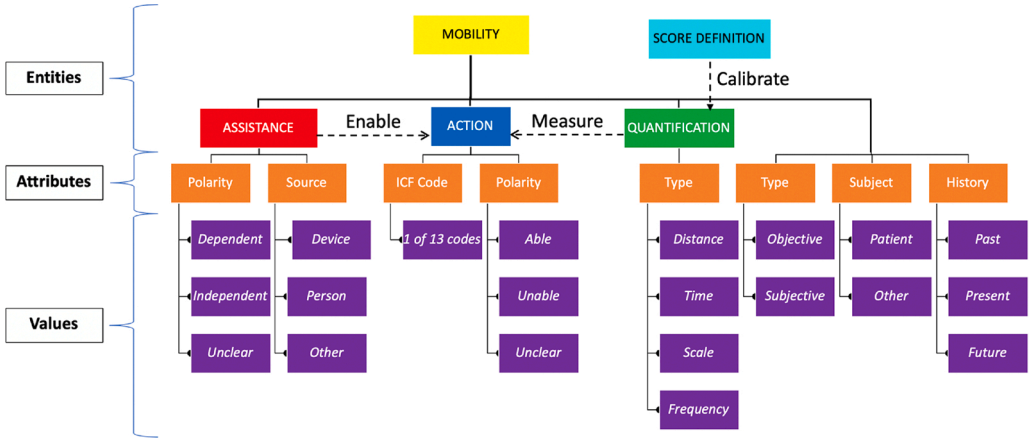

Fig. 3. A hierarchy of mobility-related entities, attributes, attribute values, and relations. Implicit relations between entities are expressed in dashed arrows.

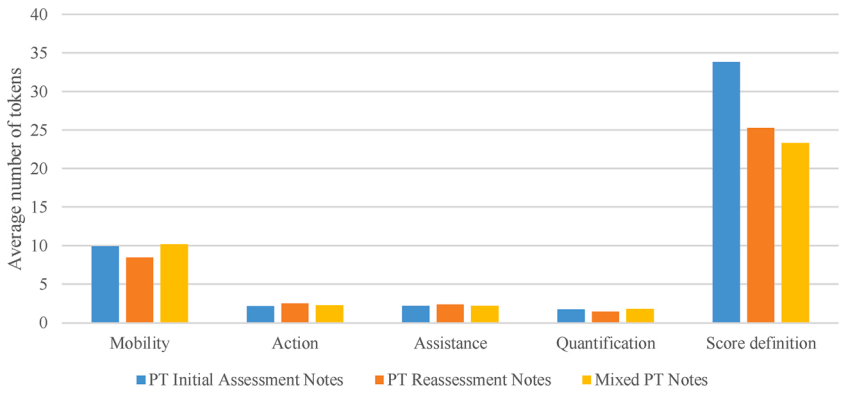

Fig. 5. Average number of tokens by entity types and PT note types.

training and evaluation sets. In stage two, we used the base models to predict NER tags (Section 2.5.2) of the evaluation set and used the prediction as raw input to generate features for our Ensemble method (Section 2.5.5). Next we trained our Ensemble model with the generated features while using human annotated tags on the evaluation set as labels. Finally, we compared performance of our Ensemble model against the base models on the held-out test set.

2.5.1. Tokenization

We used tokenizer of Stanford CoreNLP [51] to split a clinical note into tokens and associated character indices. We implemented a rule-based post-processor to correct tokenizing errors on common PT scribing patterns (Table 2).

2.5.2. Modelling entity mention recognition task

We modeled the nested NER task using joined label tagging [52]. Specifically, each token was assigned a tag, and the sequence of tags encoded entity mentions. We used the common BIO tagging scheme (Appendix A.5). Performance difference between BIO tagging compared to other schemes such as BIOES was inconclusive [53,54].

2.5.3. Tagging granularity

We observed that PT notes contained noisy end-of-sentence signals. These noises made algorithmic sentence segmentation inaccurate and the downstream fragmented sentences perturbed human annotators. We decided to annotate the GSC on the whole clinical note, without sentence segmentation. Sequence tagging on a lengthy document is a harder structure prediction problem, while noisy sentence segmentation might trim away useful context of an entity mention. To thoroughly investigate the accuracy of NER models, we conducted NER on two levels of granularity: document level, and sentence level.

In document-level tagging, the NER classifier took each entire PT note as one example. In sentence-level tagging, we used Stanford CoreNLP [51] to split a PT note into sentences. Sequence tagging models were trained and decoded on the sentences. After that, predicted tags of sentences were concatenated to form tagging of the whole PT note. At evaluation, we measured entity-level performance of both document-level tagging and sentence-level tagging using the same script.

2.5.4. Base classifiers

We used three popular NER methods: Conditional Random Field (CRF), Recurrent Neural Networks (RNN), and Bidirectional Encoder Representations from Transformers (BERT).

CRF is a probabilistic graphical model [55] and we used Stanford NER implementation [51,56]. We kept the original feature set including lexical, morphological, n-gram, and word shape features.

The RNN we used is a bi-directional long-short term memory neural networks (Bi-LSTM) [57] with a CRF decoding layer [58]. We parametrized the Bi-LSTM-CRF with 0.005 learning rate, gradient clipping at 5.0, and a dropout rate of 0.5. We also experimented with two sets of pre-trained word vectors: (a- Wikipedia) GloVe 300 dimensional vectors embedded from 6B tokens of Wikipedia 2014 and Gigaword 5 [59], and (b- PubMed) word2vec [60] 200 dimensional vectors embedded from 5B tokens of PubMed abstracts and PubMed Central full-text articles [61].

We experimented three pre-trained models of bidirectional transformers: BERT (base + large) [62] and BioBERT [63]. Both BERT (base + large) models were pre-trained on BooksCorpus (800 M tokens) [64] and English Wikipedia (2,5B tokens). BERT base had 110 M parameters while BERT large had 340 M parameters. BioBERT was BERT base additionally pre-trained on 4.5B tokens PubMed abstracts and 13.5B tokens PubMed Central full-text articles. We fine-tuned BERT models to do NER with 5 epochs, a batch size of 32, and a dropout rate of 0.1.

2.5.5. Ensemble learning

Our method employed ensemble stacking that combines outputs of multiple classifiers. We used Scikit-learn [65] to stack outputs of CRF, RNN, and BERT under two combiners: (a) Softmax, and (b) Error-Correcting Output Code (ECOC) model [66] with Support Vector Machine [67,68]. At each tag position, we extracted a symmetric feature

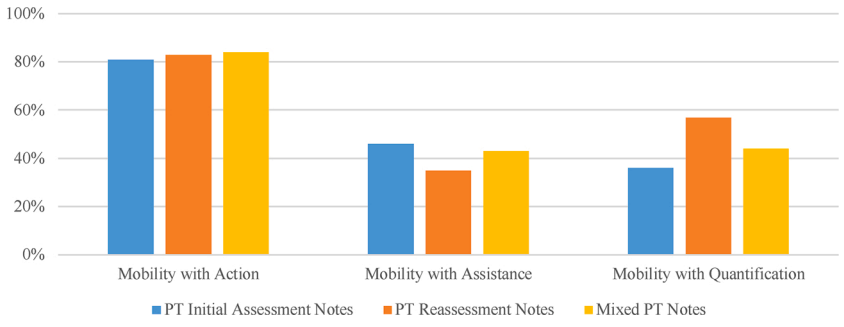

Fig. 6. Distribution of Mobility mentions by entity types and PT note types.

Table 3
Number of entity mentions annotated by entity types, annotators, and annotation phases.

	Total	Mob	Act	Asst	Quant	ScDf
D - A1	13,236	4,387	4,190	2,256	2,101	302
D - A2	14,169	4,597	4,490	2,485	2,292	305
C - A1	14,010	4,653	4,441	2,412	2,202	302
C - A2	14,263	4,623	4,525	2,509	2,303	303
Gold	14,281	4,631	4,527	2,517	2,303	303

Notes: D = double annotation, C = cross-adjudication, Gold = gold standard (consensus adjudication), A1 = first annotator, A2 = second annotator, Mob = Mobility, Act = Action, Asst = Assistance, Quant = Quantification, ScDf = Score Definition.

Table 4
Best performing models on each entity type. Average performance and standard deviation are computed on exact matching over five-fold cross-validation.

Parameters **F-1 score** Precision Recall	Mobility	Action	Assistance	Quantification	Score Definition
CRF	*sent* **71.26 ± 1.66** 78.25 ± 2.03 65.43 ± 1.50	*sent* **81.04 ± 1.93** 87.57 ± 2.83 75.46 ± 2.07	*sent* **68.89 ± 2.42** 76.79 ± 1.49 62.55 ± 3.56	*sent* **86.92 ± 2.47** 93.69 ± 2.93 81.09 ± 2.61	*doc* **93.91 ± 7.69** 97.97 ± 2.33 90.99 ± 12.37
RNN (Bi-LSTM-CRF)	*doc, pubmed* **73.04 ± 2.91** 74.36 ± 3.36 71.77 ± 2.52	*doc, pubmed* **83.89 ± 1.97** 84.23 ± 2.69 83.61 ± 2.30	*sent, wiki* **71.46 ± 3.54** 74.15 ± 3.03 69.00 ± 4.30	*sent, wiki* **87.95 ± 2.89** 89.79 ± 3.24 86.22 ± 3.13	*doc, wiki* **92.74 ± 6.77** 95.44 ± 2.11 90.96 ± 11.53
BERT	*sent, large* **74.17 ± 1.43** 73.28 ± 2.05 75.09 ± 1.21	*sent, large* **86.00 ± 1.29** 85.04 ± 1.94 87.00 ± 1.13	*sent, large* **70.29 ± 3.67** 71.48 ± 4.23 69.23 ± 4.08	*sent, large* **88.79 ± 4.11** 87.84 ± 6.33 89.92 ± 2.12	*sent, bio* **92.40 ± 7.52** 96.00 ± 2.67 90.08 ± 12.74
Ensemble	*ECOC, w=15* **78.02 ± 1.63** 79.68 ± 1.86 76.42 ± 1.55	*ECOC, w=3* **87.67 ± 0.91** 87.78 ± 1.90 87.60 ± 0.65	*ECOC, w=9* **74.74 ± 2.45** 78.36 ± 1.16 71.55 ± 4.24	*Softmax, w=1* **89.65 ± 3.56** 90.15 ± 5.08 89.23 ± 2.29	*ECOC, w=11* **94.41 ± 7.07** 97.97 ± 1.52 91.87 ± 11.74

Notes: doc/sent = tagging at document/sentence level, wiki/pubmed = Wikipedia/PubMed pre-trained word embedding, base/large/bio = types of BERT models, w = size of feature window.

window comprising of tags produced by the base classifiers. For example, to predict a tag at position *k* with a feature window of size 3, we extracted tags produced by individual classifiers at positions *k-1, k,* and *k+1* into a feature vector:

$$v = (tag_{k-1}^{CRF}, tag_{k}^{CRF}, tag_{k+1}^{CRF}, tag_{k-1}^{RNN}, tag_{k}^{RNN}, tag_{k+1}^{RNN}, tag_{k-1}^{BERT}, tag_{k}^{BERT}, tag_{k+1}^{BERT})$$

We experimented with feature windows of odd sizes ranging from 1 to 39 to fully encapsulate all entity types based on average-lengths (Fig. 5). Our ensemble classifier aggregated prediction outputs of all CRF and RNN models. For BERT models, we only aggregated outputs on sentence tagging because BERT document level tagging performed badly.

3. Results

3.1. Corpus characteristics

The gold standard corpus consists of 400 PT notes across three subsets: 200 PT initial assessment notes, 150 PT reassessment notes, and 50 mixed PT notes. The corpus has 274,165 tokens with 13,814 unique tokens, and each PT note on average has 685 tokens with 316 unique tokens (Appendix A.6).

Fig. 5 shows the average number of tokens per entity type. Fig. 6 shows the co-occurrence of Mobility mentions with sub-entity mentions. There is also a small portion (≈ 0.1 %) of Mobility mentions that do not contain any sub-entity mentions. Table 3 shows distribution of mentions as the annotation process transitioned across phases. The variation in number of mentions indicates the two annotators making effort to come to a consensus agreement.

We computed IAAs for both text span agreement and attribute values (Appendix A.7), together with two sample proportion significance tests p-values of the change in precision and recall of entity text spans (Appendix A.8). Based on a conservative significance level at p-value < 0.002, most entity types exhibit statistically significant improvement in IAA when moving from an earlier to a later annotation phase (Appendix A.9).

3.2. Named entity recognition

Table 4 presents the best NER results of our Ensemble method compared to the best results of three base classifiers. All results are averages over five cross-validation folds. Generally, the level of conservation (precision) decreased from CRF, RNN, to BERT, while the level of aggressiveness (recall) increased from CRF, RNN, to BERT. The uncorrelation of the base classifiers is a prerequisite for Ensemble learning. As a result, our Ensemble method outperformed all base classifiers in F1-score on all entity types. Our RNN model alone yielded higher performance on Mobility mentions compared to a prior work [41]. Our Ensemble method thus established a strong baseline to benchmark mobility-related entity recognition.

Performance differences between classifiers reveals interesting properties of each entity type. Mobility and Action required more context to identify correctly, so they were better recognized at document level for RNN. They also shared commonality with biomedical text, as evidenced by Pubmed embedding in RNN. On the contrary, Assistance and Quantification were more independent on context and shared commonality with news-wire text. In overall, Ensemble was able to rely on base classifiers' outputs with relative short window size. Using larger window size than Assistance's average length implied Ensemble had difficulty in identifying Assistance entities.

Looking across entity types, Action and Quantification were short in textual length and that made them easier to detect than Mobility. Score Definition was the longest type of entity but having the highest detection accuracy due to its rather uniformed wording. Assistance was the opposite with short textual length but low detection accuracy. This was due to Assistance mentions expressed more textual variation and their quantity was only about half of Mobility quantity. Beside the difference in quantity, we hypothesized that Mobility was better detected than Assistance because it relied on signals from the more accurate Action and Quantification sub-entities.

4. Discussion

4.1. Comparison to related works

Annotation of gold standard datasets and benchmarking NER performance were prevalent in general English [69] and biomedical sublanguage [70]. Recent reviews [69,70] summarized 17 popular English NER corpora and 39 popular biomedical NER corpora. A typical corpus

Summary Table

What was already known on the topic

- Functioning terminology is underpopulated in electronic health records and underrepresented in the Unified Medical Language System (UMLS) [27].
- Use of functioning information has been incoherent, unorganized [28,30,38], incomplete [27–30,39], and relies on manually-built mapping tables [29].
- Recent work focuses on one domain (i.e. Mobility) of the ICF and systematically argued for the need to capture and standardize functioning information [40], created an annotated corpus [32], compare several choices of word embeddings [41], and attempt to classify a coarse qualifier [42].

What this study added to our knowledge

- This study provides a complete process to analyze a new clinical domain for natural language processing. It significantly extends a previous summary [32] by providing comprehensive analysis of entity hierarchy, annotation procedures, corpus characteristics, and irregular challenges.
- This study demonstrates that complex and nested functioning concepts can be accurately identified given an adequate training corpus. This is an advancement to a previous approach using manual mapping tables [29].
- This study is the first that analyzes the strength and weaknesses of the conditional random field and recurrent neural networks in information extraction of functioning concepts. It subsequently builds a state-of-the-art ensemble model for mobility-related named entity recognition.
- This study is the first comprehensive analysis of an entire domain of the ICF, including entity analysis, annotation, quality control, and machine sequence labeling.

in these domains contained thousands of abstracts and a dozen entity types. NER performance typically reached more than 0.90 F1-score in English [62] and more than 0.80 F1-score in biomedicine [71].

In the medical/clinical domain, English corpora containing annotated concepts coupled with attributes and/or relations were sparse and NER benchmarking was infrequent. Popular datasets include 2009 i2b2 [9] with 1,243 discharge notes, 2010 i2b2/VA [4] with 1,748 discharge notes, ShARe/MIMIC-II corpus [72] used in SemEval [21,73] with 531 clinical notes, MiPACQ [74] with 13,091 sentences, and CLEF [44] with 150 clinical notes. Besides, ACL conferences published several corpora with 5,000 abstracts [75], 5,160 clinical notes [76], and 300 discharge notes [77]. Recent clinical NER performance reached ≈0.85 F1-score [78,79]. Our work is the first in the functioning sublanguage of clinical domain that incorporated three components: (i) a semantically annotated corpus, (ii) a compact entity hierarchy to represent a rather complex sublanguage, and (iii) a strong baseline for benchmarking NER. Our new Ensemble NER performance of 0.849 F1-score was close to the top NER performance in clinical NLP. Our annotated corpus of 400 clinical notes was humble but approximately equal in size to other well-known corpora such as ShARe/MIMIC-II, MiPACQ, and CLEF.

Existing NLP works in the functioning sublanguage either collected shallow phrases [38], or involved manual conversion of clinical text [39]. Our work carried deeper semantics than grouping of phrases and provided a fully automatic method to extract mobility concepts. Our focus on the entire Mobility domain was comprehensive and our annotation process was systematic similar to prior works [44,75]. The impact of our corpus has already been demonstrated by recent analysis [42,80]. Unlike others, irregularities existed in this new mobility sublanguage (Appendix A.10).

4.2. Limitations

Both the hierarchical order of mobility-related entities and the annotated corpus were derived from rehabilitation patient records at the NIH Clinical Center; thus, they reflected regional language idiosyncrasies. Our representation was limited to a single domain of the ICF and did not capture cross-domain interaction. We simplified the definition of an entity as a contiguous span of text, and our annotation lacked deeper semantic layers such as co-references and event annotation. Our ensemble NER accuracy is still well under human IAA performance, thus leaving space for NER model research. Despite a recent attempt [42], entity attribute grounding tasks are mostly open for the scientific community.

4.3. Future directions

We plan to expand the gold standard corpus to claimants' clinical notes at the Social Security Administration (SSA). We are also interested in applying our method to publicly available datasets such as i2b2 and MIMIC. Our entity representation would also benefit from further research on combining representation across multiple ICF domains.

5. Conclusion

Our work contributed three folds to clinical NLP community: (i) created a hierarchical entity representation that consistently captured the entire Mobility domain of the ICF, (ii) annotated a semantic corpus of mobility-related concepts and attributes, and (iii) established a strong baseline to benchmark mobility NER in clinical notes. We expect this pioneer work to proliferate research in this important yet underexplored area.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Source code of our method is freely available at: https://bitbucket.org/LanguageAndIntelligence/mobilityconcepts.

The datasets generated and analysed during the current study are not publicly available due to NIH privacy restriction on clinical records at the NIH Clinical Center. However, we are investigating the option of publicly releasing the pre-trained models, subject to NIH Clinical Center

privacy guidelines.

Funding

Supported by the Intramural Research Program of the National Institutes of Health and the US Social Security Administration.

CRediT authorship contribution statement

Thanh Thieu: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Jonathan Camacho Maldonado:** Data curation, Methodology, Validation, Writing - original draft, Writing - review & editing. **Pei-Shu Ho:** Data curation, Methodology, Validation, Writing - original draft, Writing - review & editing. **Min Ding:** Software, Formal analysis, Visualization. **Alex Marr:** Software, Formal analysis. **Diane Brandt:** Methodology, Writing - review & editing. **Denis Newman-Griffis:** Methodology. **Ayah Zirikly:** Visualization. **Leighton Chan:** Resources, Funding acquisition. **Elizabeth Rasch:** Project administration, Funding acquisition.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgement

We are especially thankful to Julia Porcino, Liansheng Tang, Chunxiao Zhou, Ao Yuan, Lisa Nelson, Albert Lai, and Jamil Hashmi for their critical reviews and insightful recommendations.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2020.104351.

References

[1] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, et al., Clinical information extraction applications: a literature review, J. Biomed. Inform. 77 (2018) 34–49.
[2] Ö Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (5) (2007) 550–563.
[3] Ö Uzuner, Recognizing obesity and comorbidities in sparse data, J. Am. Med. Inform. Assoc. 16 (4) (2009) 561–570.
[4] Ö Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556.
[5] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, J. Am. Med. Inform. Assoc. 20 (5) (2013) 806–813.
[6] A. Stubbs, C. Kotfila, Ö Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1, J. Biomed. Inform. 58 (2015). S11-S9.
[7] S. Henry, K. Buchan, M. Filannino, A. Stubbs, O. Uzuner, 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records, J. Am. Med. Inform. Assoc. 27 (1) (2020) 3–12.
[8] Ö Uzuner, I. Goldstein, Y. Luo, I. Kohane, Identifying patient smoking status from medical discharge records, J. Am. Med. Inform. Assoc. 15 (1) (2008) 14–24.
[9] Ö Uzuner, I. Solti, E. Cadag, Extracting medication information from clinical text, J. Am. Med. Inform. Assoc. 17 (5) (2010) 514–518.
[10] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, B.R. South, Evaluating the state of the art in coreference resolution for electronic medical records, J. Am. Med. Inform. Assoc. 19 (5) (2012) 786–791.
[11] A. Stubbs, C. Kotfila, H. Xu, Ö Uzuner, Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2, J. Biomed. Inform. 58 (2015) S67–S77.
[12] S. Pradhan, N. Elhadad, B.R. South, D. Martinez, L.M. Christensen, A. Vogel. Task 1: ShARe/CLEF Ehealth Evaluation Lab 2013, CLEF (Working Notes), 2013.
[13] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D.L. Mowery, et al., Overview of the share/clef ehealth evaluation lab 2014, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2014.
[14] L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, et al., Overview of the CLEF eHealth evaluation lab 2015, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2015.
[15] A. Névéol, K.B. Cohen, C. Grouin, T. Hamon, T. Lavergne, L. Kelly, Clinical information extraction at the CLEF eHealth evaluation lab 2016, in: CEUR Workshop Proceedings, NIH Public Access, 2016.
[16] L. Goeuriot, L. Kelly, H. Suominen, A. Névéol, A. Robert, E. Kanoulas, et al., CLEF 2017 eHealth evaluation lab overview, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017.
[17] H. Suominen, L. Kelly, L. Goeuriot, A. Névéol, L. Ramadier, A. Robert, et al., Overview of the CLEF eHealth evaluation lab 2018, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2018.
[18] L. Kelly, H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, et al., Overview of the CLEF eHealth evaluation lab 2019, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019.
[19] H. Suominen, L. Kelly, L. Goeuriot, M. Krallinger, CLEF eHealth evaluation lab 2020, in: European Conference on Information Retrieval, Springer, 2020.
[20] I. Segura Bedmar, P. Martínez, M. Herrero Zazo. Semeval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (Ddiextraction 2013), Association for Computational Linguistics, 2013.
[21] N. Elhadad, S. Pradhan, S. Gorman, S. Manandhar, W. Chapman, G. Savova, SemEval-2015 task 14: analysis of clinical text. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015.
[22] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, M. Verhagen, SemEval-2016 task 12: clinical TempEval, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California: Association for Computational Linguistics, 2016. Jun.
[23] S. Bethard, G. Savova, M. Palmer, J. Pustejovsky, SemEval-2017 task 12: clinical TempEval, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada: Association for Computational Linguistics, 2017. Aug.
[24] M. Hopfe, B. Prodinger, J.E. Bickenbach, G. Stucki, Optimizing health system response to patient's needs: an argument for the importance of functioning information, Disabil. Rehabil. (2017) 1–6.
[25] G. Stucki, J. Bickenbach, Functioning: the third health indicator in the health system and the key indicator for rehabilitation, Eur. J. Phys. Rehabil. Med. 53 (1) (2017) 134–138.
[26] WHO, International Classification of Functioning, Disability and Health, World Health Organization, Geneva, 2001.
[27] J. Kuang, A.F. Mohanty, V.H. Rashmi, C.R. Weir, B.E. Bray, Q. Zeng-Treitler, Representation of functional status concepts from clinical documents and social media sources by standard terminologies, AMIA Annual Symposium Proceedings (2015) 795–803.
[28] J.L. Greenwald, P.R. Cronin, V. Carballo, G. Danaei, G. Choy, A novel model for predicting rehospitalization risk incorporating physical function, cognitive status, and psychosocial support using natural language processing, Med. Care (2016).
[29] R. Kukafka, M.E. Bales, A. Burkhardt, C. Friedman, Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health, J. Am. Med. Inform. Assoc. 13 (5) (2006) 508–515.
[30] R. Mahmoud, N. El-Bendary, H.M.O. Mokhtar, A.E. Hassanien, ICF based automation system for spinal cord injuries rehabilitation, 2014 9th International Conference on Computer Engineering & Systems (ICCES) (2014) 192–197.
[31] A.B. Abacha, A.G.S.D. Herrera, K. Wang, L.R. Long, S. Antani, D. Demner-Fushman, Named Entity Recognition in Functional Neuroimaging Literature, IEEE BIBM, Kansas City, MO, USA, 2017.
[32] T. Thieu, J.C. Maldonado, P.-S.-S. Ho, J. Porcino, M. Ding, L. Nelson, Inductive identification of functional status information and establishing a gold standard corpus: a case study on the mobility domain, in: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, 2017.
[33] M. Bada, L. Hunter, Desiderata for ontologies to be used in semantic annotation of biomedical documents, J. Biomed. Inform. 44 (2011) 94–101. United States: 2010 Elsevier Inc.
[34] S.V. Pakhomov, A. Coden, C.G. Chute, Developing a corpus of clinical notes manually annotated for part-of-speech, Int. J. Med. Inform. 75 (6) (2006) 418–429.
[35] D. Albright, A. Lanfranchi, A. Fredriksen, W.F. Styler, C. Warner, J.D. Hwang, et al., Towards comeprehensive syntactic and semantic annotations of the clinical narrative, J. Am. Med. Inform. Assoc. 20 (5) (2013) 922–930.
[36] S.W. Tu, C.I. Nyulas, T. Tudorache, M.A. Musen, A method to compare ICF and SNOMED CT for coverage of U.S. Social security administration's disability listing criteria, AMIA Annual Symposium Proceedings (2015) 1224–1233.
[37] E.A. Lindemann, E.S. Chen, S. Rajamani, N. Manohar, Y. Wang, G.B. Melton, Representation of Occupation Information in Clinical Texts: An Analysis of Free-Text Clinical Documentation in Multiple Sources, AMIA Joint Summits on Translational Science, San Francisco, 2017.
[38] S. Skube, E. Lindemann, E. Arsoniadis, M. Akre, E. Wick, G. Melton, Characterizing Functional Health Status of Surgical Patients in Clinical Notes, AMIA Informatics Summit, San Francisco, 2018.
[39] A.P. Ruggieri, S.V. Pakhomov, C.G. Chute, A corpus driven approach applying the "frame semantic" method for modeling functional status terminology, Stud. Health Technol. Inform. 107 (Pt 1) (2004) 434–438.
[40] D. Newman-Griffis, J. Porcino, A. Zirikly, T. Thieu, J. Camacho Maldonado, P.-S. Ho, et al., Broadening horizons: the case for capturing function and the role of health informatics in its use, BMC Public Health 19 (1) (2019) 1288.
[41] D. Newman-Griffis, A. Zirikly, Embedding transfer for low-resource medical named entity recognition: a case study on patient mobility, in: Proceedings of the BioNLP

2018 Workshop, Melbourne, Australia: Association for Computational Linguistics, 2018. Jul.
[42] D. Newman-Griffis, A. Zirikly, G. Divita, B. Desmet, Classifying the reported ability in clinical mobility descriptions, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy: Association for Computational Linguistics, 2019. Aug.
[43] J.J. Cimino, 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. 52 (2014) 11–27.
[44] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, et al., Building a semantically annotated corpus of clinical texts, J. Biomed. Inform. 42 (5) (2009) 950–966.
[45] H. Cunningham, D. Maynard, K. Bontcheva, Text Processing With GATE, Gateway Press, CA, 2011.
[46] G. Hripcsak, A.S. Rothschild, Agreement, the F-Measure, and Reliability in Information Retrieval, J. Am. Med. Inform. Assoc. 12 (3) (2005) 296–298.
[47] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.
[48] E.F.T.K. Sang, F.D. Meulder, 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; Edmonton, Canada, 1119195: Association for Computational Linguistics, 2003, pp. 142–147.
[49] N. Chinchor, MUC-4 evaluation metrics, in: Proceedings of the 4th Conference on Message Understanding; McLean, Virginia, 1072067: Association for Computational Linguistics, 1992, pp. 22–29.
[50] N. Chinchor , B. Sundheim , MUC-5 Evaluation Metrics 1993.
[51] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit, Association for Computational Linguistics, 2014.
[52] B. Alex, B. Haddow, C. Grover, Recognising nested named entities in biomedical text, in: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic, 1572404: Association for Computational Linguistics, 2007, pp. 65–72.
[53] J. Yang, S. Liang, Y. Zhang, Design challenges and misconceptions in neural sequence labeling, 27th International Conference on Computational Linguistics (COLING) (2018).
[54] N. Reimers, I. Gurevych. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-Networks for Sequence Tagging, Association for Computational Linguistics, 2017.
[55] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, 655813: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.
[56] J.R. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; Ann Arbor, Michigan, 1219885: Association for Computational Linguistics, 2005, pp. 363–370.
[57] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[58] F. Dernoncourt, J.Y. Lee, P. Szolovits. NeuroNER: an Easy-to-Use Program for Named-Entity Recognition Based on Neural Networks, Association for Computational Linguistics, 2017.
[59] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics, 2014. Oct.
[60] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013.
[61] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, S. Ananiadou, Distributional semantics resources for biomedical text processing. 5th Languages in Biology and Medicine Conference (LBM 2013), 2013.
[62] J. Devlin, M.-W.- Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, 2019. Jun.
[63] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics. 36 (4) (2019) 1234–1240.
[64] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, et al., Aligning books and movies: towards story-like visual explanations by watching movies and reading books, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV): IEEE Computer Society (2015) 19–27.
[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[66] T.G. Dietterich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes, J. Artif. Intell. Res. 2 (1) (1995) 263–286.
[67] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory; Pittsburgh, Pennsylvania, USA, 130401: ACM, 1992, pp. 144–152.
[68] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[69] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Trans. Knowl. Data Eng. (2020), 1-.
[70] M.-S. Huang, P.-T. Lai, P.-Y. Lin, Y.-T. You, R.T.-H. Tsai, W.-L. Hsu, Biomedical named entity recognition and linking datasets: survey and our recent development, Brief. Bioinformatics (2020).
[71] H. Cho, H. Lee, Biomedical named entity recognition using deep neural networks with contextual information, BMC Bioinformatics 20 (1) (2019) 735.
[72] S. Pradhan, N. Elhadad, B.R. South, D. Martinez, L. Christensen, A. Vogel, et al., Evaluating the state of the art in disorder recognition and normalization of the clinical narrative, J. Am. Med. Inform. Assoc. 22 (1) (2014) 143–154.
[73] S. Pradhan, W. Chapman, S. Man, G. Savova, Semeval-2014 task 7: analysis of clinical text, in: Proc of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Citeseer, 2014.
[74] D. Albright, A. Lanfranchi, A. Fredriksen, W.F. Styler, C. Warner, J.D. Hwang, et al., Towards comeprehensive syntactic and semantic annotations of the clinical narrative, J. Am. Med. Inform. Assoc. 20 (5) (2013) 922–930.
[75] B. Nye, J.J. Li, R. Patel, Y. Yang, I. Marshall, A. Nenkova, A Corpus with multi-level annotations of patients, in: Interventions and Outcomes to Support Language Processing for Medical Literature, Association for Computational Linguistics, 2018.
[76] P. Patel, D. Davey, V. Panchal, P. Pathak, Annotation of a large clinical entity corpus, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics, 2018 oct nov.
[77] N. Alnazzawi, P. Thompson, S. Ananiadou, Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi), 2014.
[78] Y. Wu, M. Jiang, J. Xu, D. Zhi, H. Xu, Clinical named entity recognition using deep learning models, AMIA Annual Symposium Proceedings AMIA Symposium 2017 (2018) 1812–1819.
[79] G. Xu, C. Wang, X. He, Improving clinical named entity recognition with global neural attention, in: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, Springer, 2018.
[80] D. Newman-Griffis, E. Fosler-Lussier, HARE: a flexible highlighting annotator for ranking and exploration. Conference on Empirical Methods in Natural Language Processing: Systems Demonstrations, 2019.