Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200129

Clinical Concept Normalization on Medical Records Using Word Embeddings and Heuristics

João Figueira SILVA [a], Rui ANTUNES [a], João Rafael ALMEIDA [a,b], and Sérgio MATOS [a,1]
[a] *DETI/IEETA, University of Aveiro, Portugal*
[b] *Department of Computation, University of A Coruña, Spain*

Abstract. Electronic health records contain valuable information on patients' clinical history in the form of free text. Manually analyzing millions of these documents is unfeasible and automatic natural language processing methods are essential for efficiently exploiting these data. Within this, normalization of clinical entities, where the aim is to link entity mentions to reference vocabularies, is of utmost importance to successfully extract knowledge from clinical narratives. In this paper we present sieve-based models combined with heuristics and word embeddings and present results of our participation in the 2019 n2c2 (National NLP Clinical Challenges) shared-task on clinical concept normalization.

Keywords. natural language processing, clinical information extraction, clinical concept disambiguation, word embeddings, sieve-based model

1. Introduction

Electronic health records (EHRs) hold great value for the correct understanding of patient trajectories since they combine the multimodality in medical data with the temporal aspect that is embedded in the clinical history, rendering it as a vital component to correctly understand how symptoms, findings, diagnoses and diseases evolve [1].

Textual data is highly relevant and can be found in EHRs in (semi-)structured and unstructured formats, the latter being commonly referred to as free text. Medical narratives (*e.g.* discharge or admission reports) are stored as free text since natural language provides a flexible convoy for physicians to track and report each medical situation, overcoming the limitations from ambiguous and unspecific terms that physicians can find when using coding standards such as RxNorm or SNOMED-CT. Owing to that, it is reckoned that free text frequently contains plenty of information otherwise not obtainable from other data sources [2]. However, it is unfeasible to manually analyze large scale medical datasets. Despite the intricacies inherent to free text that make it very challenging to process and interpret, the process of automatically annotating clinical narratives is

[1] Corresponding Author: Sérgio Matos, DETI/IEETA, Universidade de Aveiro, Campus Universitário do Santiago, Aveiro 3810-193, Portugal; E-mail: aleixomatos@ua.pt.

an important way to summarize or extract relevant data from medical text. This typically requires the text processing tasks of named entity recognition (NER) and named entity normalization (NEN) to extract relevant concepts and to standardize their referencing. The latter is a necessary step to combat ambiguities since clinical text contains many abbreviations, misspellings, and domain-specific expressions.

In this paper we describe an approach for normalization of clinical entity mentions using sieve-based models combined with heuristics and word embeddings. The proposed method was used in the 2019 n2c2 shared task on clinical concept normalization[3], where the aim was to link clinical entities to SNOMED CT or RxNorm vocabularies through UMLS Concept Unique Identifiers (CUIs) [3].

2. Materials and Methods

The aim of the n2c2 clinical concept normalization task was to normalize medical entities to standard medical vocabularies. For simplicity purposes, the task focused only in NEN, bypassing the NER step by providing relevant clinical mentions identified *a priori*.

In this section, we describe the dataset used as well as the three different methodologies applied, namely a method based on concept embeddings, an improved version of a baseline sieve approach, and a final rule-based method.

2.1 Dataset

The Medical Concept Normalization (MCN) corpus proposed by Luo *et al.* [1] was used. It comprises a wider set of clinical concepts in contrast to previous NEN challenges that only evaluated the normalization of disease mentions [4,5].

Each annotated clinical entity is associated with a single CUI from the UMLS 2017AB version. For example, "hypertension" and "HTN", or "blood pressure" and "BP" are two examples of expressions that refer to the same concepts and are therefore identified by the same CUIs (C0020538 and C0005824, respectively). Although UMLS encompasses several vocabularies, only two were used for annotation. RxNorm was used to annotate clinical drugs and medications, whereas SNOMED-CT, an extensive vocabulary of clinical terminology, was used for normalizing the remaining concepts (disorders, procedures, body structures, and others).

Table 1. Detailed dataset statistics.

	Training	Test	Total
Number of clinical records	50	50	100
Number of annotated entities	6 684	6 925	13 609
Number of unique CUIs	2 331	2 579	3 792

The dataset contains a total of 100 annotated discharge summaries and is split in training and test subsets (Table 1). This allowed model development in the training set and a blind official challenge evaluation using withheld test data (the gold standard annotations for the test data were only available after the official evaluation).

2.2 Developed Methodologies

2.2.1 Concept Embeddings Similarity

This method involves two sequential steps: (1) text pre-processing and (2) similarity computation. In the first step, certain text rewrite rules were handcrafted by inspecting the clinical named entities in the training set (Table 2) with the aim of cleansing the surface representation of these mentions. In addition to the text replacements made, HTML entities and other superfluous symbols were also discarded.

Table 2. Examples of text rewrite rules handcrafted according to the training set.

b/l	bilateral	po	oral	10%	partial
co2	carbon dioxide	trop	troponin	¼	fourth
e.	escherichia	u/s	ultrasound scan	x2	double
iv	intravenous	vit	vitamin	2L	two liters
mso4	morphine sulphate	w/u	workup	3 of 6	iii/vi

In the second step we represented the clinical named entities (from training andtest sets) and UMLS concept names by using pre-calculated biomedical word embeddings. We employed the publicly available BioWordVec model [6] that was created applying the fastText library [7] and was generated from over 30 million documents from PubMed articles and clinical notes from the MIMIC-III database.

From these representations we created a direct mapping between terms and CUIs where each term was defined by the embedding vector average of its constituent words [8]. This mapping was built using (i) text mentions from the train set and (ii) concept names from the UMLS CUIs corresponding to the SNOMED-CT and RxNorm vocabularies.

At last, for making the predictions in the test set, the cosine similarity was calculated between each test mention embedding vector and all pre-calculated term embeddings [8]. The CUI corresponding to the most similar term was chosen.

A simplified overview of this method is shown in Figure 1. Due to its ability to correctly capture text mentions with similar semantics, this method was also used as a final sieve to predict remaining (unclassified) entities in the next two pipelines.

2.2.2 Improved Sieve-Based Approach

The sieve-based approach by Luo *et al.* [1] was firstly deployed, containing exact matching with the training set annotations and with UMLS, along with MetaMap sieves [9]. This pipeline was performed in two stages, firstly using raw text mentions andsecondly using clean text mentions. This approach yielded a 5-fold accuracy on the training set of 0.783. An error analysis showed that MetaMap generated too many incorrect identi- fiers, limiting the number of unclassified mentions passing to the second sieve stage, thus capping the maximum attainable accuracy.

The pipeline was reworked by shifting the MetaMap sieve from stage 1 to the end of the pipeline, right before the MetaMap sieve from stage 2. Further improvements involved performing a Monte Carlo simulation to select an optimal threshold for CUI classification with MetaMap, greatly reducing false positives. Remaining unclassified men- tions were classified using the concept embeddings approach described in Section 2.2.1. The final pipeline is presented in Figure 1.

2.2.3 An Orchestrator for Combining Simple and Complex Rules

This approach applies a 4-step pipeline combining rules and dictionary lookup, followed by previously computed concept embeddings.

The initial step processes mentions containing ambiguous terms that map to multiple CUIs in the training set. For each mention, several rules are applied that consider specific words from the respective sentence and section heading if applicable. In steps 2 and 3, the system performs an exact match using dictionaries created from the training set and the UMLS database, respectively. This three step pipeline is invoked twice: in a first phase it is applied to raw mentions, which was essential to detect certain patterns that could match concepts annotated in the training set; in a second phase, the remaining mentions are pre-processed as described in Section 2.2.1 before the pipeline is applied.

Finally, we combined this workflow with the concept embeddings from Section 2.2.1 by having a final stage where concepts that completed the previous workflow without an attributed CUI were processed using the concept embeddings.

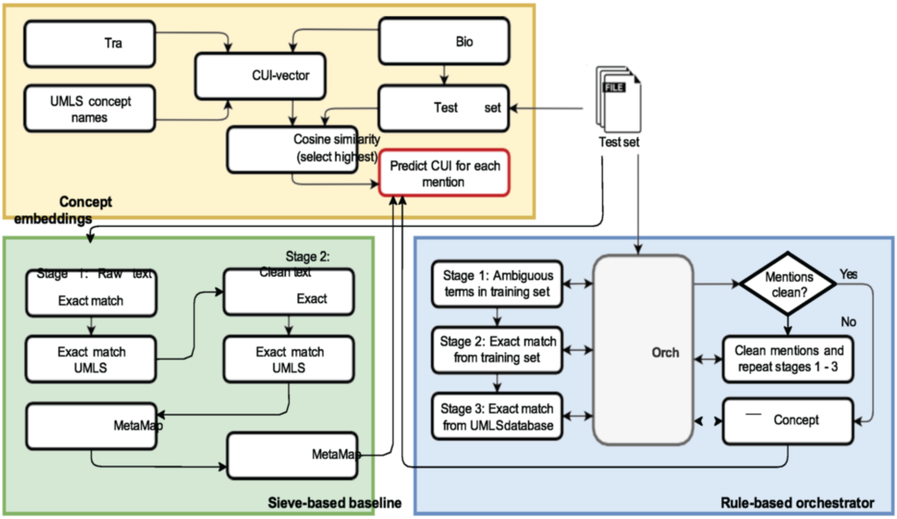

Figure 1. Final system architecture composed of three different approaches.

3. Results and Discussion

Table 3 shows the results, obtained in the training and test sets, of the three previously described methodologies.

Concept embeddings achieved an accuracy of 0.812 and 0.801 in the training and test sets. However, a posterior evaluation without using handcrafted replacements proved that the created text patterns were biased into the training set and led to overfitting, since simpler text pre-processing resulted in a lower training accuracy (0.807) and similar test accuracy (0.800).

The rule-based approach combined with the concept embeddings method obtained the highest accuracy in the test set with a value of 0.806, corresponding to a 4 percentage points improvement compared to the baseline sieve-based model proposed by Luo et al. [1] that obtained an accuracy of 0.764 in the test set.

Table 3. Obtained accuracy results with the three distinct methodologies. Results in the training set achieved by 5-fold cross-validation (10 repetitions in the concept embeddings method).

	Concept embeddings	Sieve-based	Rule-based
Training	0.812	0.806	—
Test	0.801	0.791	0.806

4. Conclusions and Future Work

In this paper, we proposed three methodologies to normalize clinical concepts from patient clinical reports automatically. The development of such methods applied in healthcare data migration pipelines is essential to increase data value and have harmonized datasets.

As future work, we intend to improve each methodology individually by using information from the surrounding context and section heading of each mention. Also we intend to incorporate UMLS concept definitions to enrich the semantic knowledge regarding each concept. These improvements may reduce concept ambiguity issues. As a final refinement, we aim to apply deep neural network models to predict the correct CUI embedding vector.

5. Acknowledgments

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. João Figueira Silva, Rui Antunes, and João Rafael Almeida are funded by the FCT - Foundation for Science and Technology (national funds) under the grants PD/BD/142878/2018, SFRH/BD/137000/2018, and SFRH/BD/147837/2019 respectively.

References

[1] Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. MCN: a comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, 92:103132, April 2019.

[2] Kasper Jensen, Cristina Soguero-Ruiz, *et al.* Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports*, 7(46226), April 2017.

[3] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Suppl 1):D267, January2004.

[4] Sameer Pradhan, Noemie Elhadad, *et al.* Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In *CLEF (Working Notes)*, Valencia, Spain, September 2013. CEUR Workshop Proceedings.

[5] Sameer Pradhan, Noémie Elhadad, *et al.* SemEval-2014 Task 7: analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[6] Qingyu Chen, Yifan Peng, and Zhiyong Lu. BioSentVec: creating sentence embeddings for biomedical texts. *arXiv:1810.09302*, October2018.

[7] Piotr Bojanowski, Edouard Grave, *et al.* Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(1):135–146, 2017.

[8] Tomas Mikolov, Kai Chen, *et al.* Efficient estimation of word representations in vector space.

[9] *arXiv:1301.3781*, January 2013.

[10] Alan R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, pages 17–21, Washington, DC, USA, November 2001. American Medical Informatics Association.