A Light-weight & Robust System for Clinical Concept Disambiguation

Dirk Weissenborn, Roland Roller, Feiyu Xu and **Hans Uszkoreit**
Language Technology Lab, DFKI
Alt-Moabit 91c, Berlin, Germany
{dirk.weissenborn, roland.roller, feiyu, uszkoreit}@dfki.de

Enrique Garcia Perez
SAP Innovation Center
Konrad-Zuse-Ring 10, Potsdam, Germany
enrique.garcia.perez@sap.com

Abstract

This paper presents a system for the normalization of concept mentions in clinical narratives. We evaluate and compare it against a popular, open-source solution that is frequently used for natural language processing of clinical text. The evaluation is based on a manually annotated dataset of 72 discharge summaries taken from the i2b2-corpus. Besides the demonstration and evaluation of our system we provide an in-depth corpus analysis that guided the development of the system. Our focus lies on the task of concept disambiguation, for which we combine two unsupervised approaches that are easy to implement and computationally inexpensive. We show that some ambiguities can only be resolved by adapting to annotation guidelines and preferences which we solve via the introduction of heuristics. Finally, we present an online-demo that gives insights into the individual parts of the normalization pipeline.

1 Introduction

Recognizing and disambiguating clinical concepts plays a central role in many information extraction tasks within the clinical domain. It requires the identification of concept mentions in clinical narratives and the disambiguation of their respective surface strings (normalization). In recent years, many tasks have focused on the normalization of clinical concepts, such as the i2b2 challenge (Uzuner et al., 2011), ShARe/CLEF (Pradhan et al., 2013) and SemEval (Elhadad et al., 2015).

Traditionally, disambiguation systems rely on supervised (Martinez and Baldwin, 2011), semi-supervised (Preiss and Stevenson, 2013) or unsupervised (Agirre et al., 2010) methods. Each of those techniques has its advantages, however, as seen in different disambiguation tasks, simple methods (and their combination) can achieve very good results, such as the generation of rules and heuristics from the training data (Afzal et al., 2015), the usage of similarity measures (Pathak et al., 2015) or the inclusion of Information Content (Leal et al., 2015).

In this work we develop a light-weight solution to the problem of clinical concept normalization, that is easy to implement and does not require expensive computations and is therefore particularly suited for industrial application. The approach is mainly unsupervised and does not require large amounts of training data. In particular, the disambiguation is based on a densest-subgraph algorithm to ensure contextual compatibility among the normalized concepts and the string similarity between the surface string and the preferred labels of a respective concept. We achieve very good performance with this setup on a manually annotated dataset. An web-application was developed for demonstration purposes and to debug the normalization pipeline[1].

2 Clinical Concept Normalization

The concept normalization task requires a well defined target vocabulary. A useful resource is the Unified Medical Language System (UMLS), which defines biomedical concepts with various names, spellings and abbreviations. Concepts within UMLS are defined by so called *concept unique identifiers* (CUI) that represent concepts across different biomedical vocabularies, such as NCI, NDF-RT or RxNorm. However, natural language is highly variable and surface strings can have different meanings depending on the context.

[1] http://clinical-ta.dfki.de

Concept-Type (Source)	**#Annotations**
Symptoms (NCI)	1434
Disease (NCI)	1370
Medication (RxNorm)	1190
Diagnostic Procedure (NCI)	647
Therapeutic procedure (NCI)	644
Anatomy (NCI)	593
Laboratory Tests (NCI)	458

Table 1: Concept annotations by type in our dataset.

Class	**% of mentions**
ambiguous	18
ambiguous given type	12
not ambiguous	49
no candidates	28
correct candidate not found	33

Table 2: Ambiguity classes and their relative frequency in the dataset. *ambiguous* - mentions with more than one candidate including the correct; *ambiguous given type* - subset of ambiguous that remains ambiguous after removing candidates of wrong type; *not ambiguous* - only one, correct candidate.

The task of normalizing surface strings to unique concepts of a given vocabulary such as UMLS can be subdivided into three partial tasks: Mention Recognition, Candidate Search and Disambiguation. Given an input text, the mention recognition subtask identifies text-spans that are potential mentions of a medical concept. Subsequently, the candidate search is responsible for finding candidate concepts for the surface strings of each mention. Finally, the disambiguation step selects the candidate that fits best into the mentions context, i.e., it resolves the ambiguity among its candidates. The work focuses on the disambiguation task.

2.1 Data

In our experiments we used a part of the i2b2[2]-corpus (Uzuner et al., 2011) that was manually re-annotated[3]. It consists of 72 discharge summaries. Overall, the dataset contains 6336 annotations. Table 1 lists annotation types and their corresponding number of annotations. The corpus was split into 2 distinct subsets, each covering half of the documents. The first set was used for system development and the second half for testing.

We also analyzed the ambiguities within the corpus based on our candidate search (§3.2). Table 2 lists different ambiguity classes and their fraction in the dataset. It shows that ambiguity arises only in 18% of mentions. Candidate search fails in about a third of all cases for which the correct candidate is not found. For most of those cases no candidate is found at all. This shows that the currently employed dictionary lookup has to be refined. However, this work addresses the problem of disambiguation. Thus, only 18% of all cases are non-trivial and are useful for evaluating disambiguation.

3 System Architecture

3.1 Mention Recognition

Because of the focus on disambiguation our demo system employs a simple approach to mention recognition. Given a tokenized input document all word n-grams up to a predefined n are extracted. This guarantees high recall. In the subsequent candidate search step we eliminate all extracted mentions for which no candidates are found.

3.2 Candidate Search

We find concept candidates for each recognized mention via a string lookup to a given dictionary. The dictionary maps surface strings to concepts. Those were extracted from a predefined subset of vocabularies in the UMLS, namely RxNorm for medications and NCI for anatomical concepts, diseases, therapeutic procedures, diagnostic procedures, laboratory tests and symptoms. The surface strings of the dictionary were expanded by including additional lexical variations.

3.3 Disambiguation

The most crucial part of the concept normalization pipeline is the concept disambiguation. Given a set of candidates for each recognized mention it selects the concept which fits best to the mention of interest. The disambiguation is guided by two algorithms, that are explained in the following.

String-Edit-Distance Each concept in UMLS may include a set of synonyms containing a range of variations and spellings. Not all of those string variations are likely to represent a concept in free text. However, a small subset of strings are indicated as *preferred labels* for a concept. In a corpus analysis, we found that many ambiguities can be

[2] `https://www.i2b2.org/`

[3] Note, the re-annotation took place within an industrial use case and was not carried out by one of the authors. The data and the dictionaries we used were already given.

resolved by selecting the candidate concept whose preferred labels contains a close match with the mention string. We further found that preferred labels of distinct UMLS concepts are usually mutual exclusive. Thus, we employ a string-edit-distance (ED) algorithm, namely Levenshtein-distance, between the preferred labels L_c of all candidates c_i^m and the mention string x_m. We use the minimum of those distances to define the ED-score of a candidate concept.

$$s_{ed}(c_i^m) = \max_{l \in L_{c_i^m}} \frac{1}{distance(x_m, l) + 1}$$

Densest-Subgraph We employ a densest-subgraph algorithm similar to Moro et al. (2014) or Weissenborn et al. (2015) to account for the context of a mention. First we construct a graph that consists of all candidates c_m^i for all mentions m of a document. These are the vertices of the graph. We connect candidate concepts from different mentions with each other, whenever they co-occurred at least once together in MEDLINE, a repository of abstracts from biomedical publications. This information is annually summarized by the National Institutes of Health (NIH)[4]. Given the concept graph $G = (V, E)$ of a document, we iteratively select a mention with the most remaining candidates and remove its least connected candidate until each mention has at most a predefined number of candidates left[5]. Given the pruned graph $G^* = (V^*, E^*)$ we score each remaining candidate by the product of its number of connections to other mention candidates and other mentions, i.e., number of mentions that have at least one connected candidate concept.

$$s_{ds}^u(c_i^m) = \left|\{c_j^{m'} | (c_i^m, c_j^{m'}) \in E^*\}\right| \cdot \left|\{m' | \exists j : (c_i^m, c_j^{m'}) \in E^*\}\right|$$

$$s_{ds}(c_i^m) = \frac{s_{ds}^u(c_i^m)}{\sum_j s_{ds}^u(c_j^m)}$$

We tried different combinations of both scores and found the disambiguation via s_{ds} with a fallback to s_{ed} to work best. I.e., we select always the candidate for each mention with the highest s_{ds} and apply s_{ed} in case there are more than one candidate with the same score.

[4] https://mbr.nlm.nih.gov/MRCOC.shtml

[5] We use 5 in our system, which performs slightly better or equal to other configurations.

3.4 Rule-based disambiguation

A problem of unsupervised disambiguation is the inability of learning corpus-specific patterns which depend on annotation guide-lines and the personal perspective of the annotators themselves. Based on our observations the following set of simple rules are defined and used to support both disambiguation techniques:

Active Substance: If the given mention is a tradename (e.g., Tylenol), in most of the cases its active substance (e.g., Acetaminophen) is annotated. Therefore we map all concepts that refer to a tradename to its active substance: This information is taken from the UMLS Metathesaurus relation *has-tradename*.

Structure of: If a mention '*M*' (e.g. 'left foot') includes two candidate concepts, one containing the preferred label 'structure of *M*' and the other one 'entire *M*', the second concept is removed from the list of candidates.

Abbreviation validation: Abbreviations tend to be highly ambiguous (Kim et al., 2011) and are difficult to disambiguate. However, in many cases those candidates are selected, whose preferred labels fit the mentioned abbreviation. To address this issue, abbreviations are firstly identified using the UMLS Lexical Tools. Next, candidates whose preferred labels are not valid long forms of a mentioned abbreviation are removed during pre-processing. Valid long forms of abbreviations have to fulfill the following criterion: The first letter of the abbreviation must match the first letter of the text, and the remainder of the abbreviation, i.e., the abbreviation without its first letter, must be an abbreviation for the either the remaining text or the remaining words, excluding the first.

4 Online Demo

The web interface of the online demo[6] is based on the BRAT NLP-tool[7] to visualize the implemented candidate search and disambiguation. Figure 1 presents the output of our Demo after processing a clinical narrative. The upper part '*Candidate Search*' displays the text including mentions with their respective concept candidates. Different colors indicate different types of concepts. In the given example, red refers to anatomy, green to

[6] http://clinical-ta.dfki.de

[7] http://brat.nlplab.org/

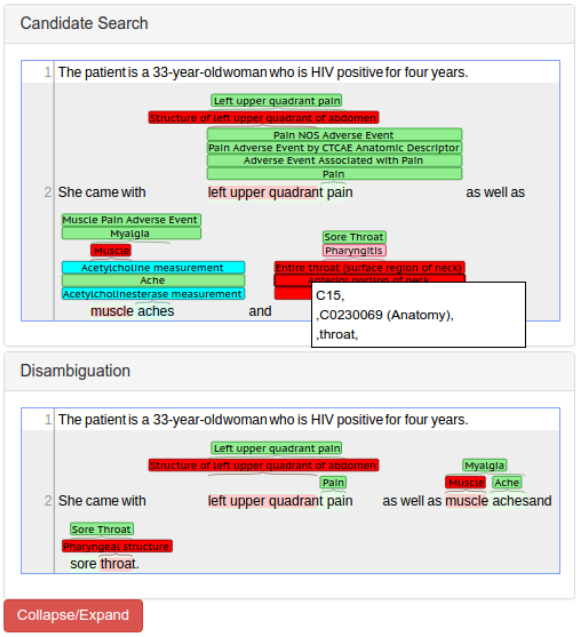

Figure 1: Annotations comprising candidate and disambiguated view.

symptom, pink to disease and turquoise to laboratory test. Moving the mouse courser over a candidate mention, the GUI shows the vocabulary origin and its concept unique identifier.

5 Experiments

5.1 Setup

We evaluated our system on the test part of the dataset with different configurations. More specifically, we compare the performance of the individual disambiguation algorithms, namely string-edit-distance (ED) and densest-subgraph (DS), and their combination, as well as a widely used reference system called cTAKES[8] (Savova et al., 2010) in combination with the disambiguation component YTEX (Garla et al., 2011). We make use of a gold-standard mention recognizer that extracts only annotated mentions in the experiments. When comparing to cTAKES, we make use of its internal mention extraction and candidate search in combination with our disambiguation to guarantee a fair comparison. Additionally, our post-processing heuristics were applied to the output of both our system and cTAKES.

5.2 Results

Table 3 shows the results on the entire testset. We achieve a high precision of over 85% which we attribute to the performance of disambiguation. Our system performs also better than cTAKES[9] with the same pre-processing (mention recognition and candidate search). The main problem in general lies in the low recall, which is mainly due to failing candidate search. This is also a major concern in future work.

System	Pre-processing	P	R	F1
ED	Gold-standard	0.850	0.592	0.698
DS	Gold-standard	0.850	0.592	0.698
DS+SE	Gold-standard	**0.857**	**0.597**	**0.703**
ED	cTAKES	0.777	0.522	0.624
DS	cTAKES	0.766	0.514	0.615
DS+ED	cTAKES	**0.780**	**0.524**	**0.627**
cTAKES	cTAKES	0.743	0.499	0.597

Table 3: Normalization results in Precision (P), Recall (R) and F1-score (F1) for all mentions in testset.

System	Pre-processing	#Mentions	P
ED	Gold-standard	502	0.751
DS	Gold-standard	502	0.751
DS+SE	Gold-standard	502	**0.781**
ED	cTAKES	270	0.730
DS	cTAKES	270	0.659
DS+ED	cTAKES	270	**0.767**
cTAKES	cTAKES	270	0.481

Table 4: Precision (P) for all non-trivial mentions in testset, i.e., mentions with at least 2 candidates containing the correct one.

As mentioned in §2.1, only a fraction of mentions can be considered non-trivial with respect to the disambiguation. Table 4 shows the performance of our system and cTAKES for all non-trivial mentions. The observations are similar to the previous results. We can see that the precision of our system is quite robust and much better than the performance of cTAKES.

6 Conclusion

We presented a light-weight disambiguation system for the normalization of clinical concept mentions. The system is mainly unsupervised and utilizes string similarity metrics as well as information from concept co-occurrences. We demonstrate its robustness with respect to disambiguation and compared it to cTAKES, a popular open-source system for clinical NLP. In addition, we give examples where our unsupervised approach fails because of annotation guidelines and preferences. This problem is solved by the introduction

[8] https://ctakes.apache.org/

[9] standard configuration for YTEX disambiguation

of simple heuristics. Finally, our system can be accessed via a web-application.

Acknowledgements

This research was partially supported by SAP, the German Federal Ministry of Economics and Energy (BMWi) through the project MACSS (01MD16011F), and by the German Federal Ministry of Education and Research (BMBF) through the project BBDC (01IS14013E).

References

Zubair Afzal, Saber A. Akhondi, Herman van Haagen, Erik M. van Mulligen, and Jan A. Kors. 2015. Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2010. Graph-based Word Sense Disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. SemEval-2015 Task 14: Analysis of Clinical Text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, Denver, Colorado, June. Association for Computational Linguistics.

Vijay Garla, Vincent Lo Re, Zachariah Dorey-Stein, Farah Kidwai, Matthew Scotch, Julie Womack, Amy Justice, and Cynthia Brandt. 2011. The Yale cTAKES extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association*, 18(5):614–620.

Youngjun Kim, John Hurdle, and Stéphane M Meystre. 2011. Using UMLS lexical resources to disambiguate abbreviations in clinical text. *AMIA Symposium*, 2011:715722.

André Leal, Bruno Martins, and Francisco Couto. 2015. ULisboa: Recognition and Normalization of Medical Concepts. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 406–411. Association for Computational Linguistics.

David Martinez and Timothy Baldwin. 2011. Word sense disambiguation for event trigger word detection in biomedicine. *BMC Bioinformatics*, 12(2):1–8.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni, Kinjal Dani, Amrish Patel, and Narayan Choudhary. 2015. ezDI: A Supervised NLP System for Clinical Narrative Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 412–416. Association for Computational Linguistics.

Sameer Pradhan, Noémie Elhadad, Brett R. South, David Martínez, Lee M. Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana K. Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*

Judita Preiss and Mark Stevenson. 2013. DALE: A Word Sense Disambiguation System for Biomedical Documents Trained using Automatically Labeled Examples. In *Proceedings of the 2013 NAACL HLT Demonstration Session*, pages 1–4, Atlanta, Georgia, June. Association for Computational Linguistics.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Dirk Weissenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. 2015. Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities. *Proc. of ACLIJCNLP, Beijing, China*, pages 596–605.