MindLab-UNAL: Comparing Metamap and T-mapper for Medical Concept Extraction in SemEval 2014 Task 7

Alejandro Riveros, Maria De-Arteaga, Fabio A. González and **Sergio Jimenez**
Universidad Nacional de Colombia
Ciudad Universitaria
Bogotá, Colombia
[lariverosc,mdeg,fagonzalezo, sgjimenezv]@unal.edu.co

Henning Müller
Univ. of Applied Sciences
Western Switzerland, HES-SO
Sierre, Switzerland
henning.mueller@hevs.ch

Abstract

This paper describes our participation in task 7 of SemEval 2014, which focuses on analysis of clinical text. The task is divided into two parts: recognizing mentions of concepts that belong to the UMLS (Unified Medical Language System) semantic group *disorders*, and mapping each disorder to a unique UMLS CUI (Concept Unique Identifier), if possible. For identifying and mapping disorders belonging to the UMLS meta thesaurus, we explore two tools: Metamap and T-mapper. Additionally, a Named Entity Recognition system, based on a maximum entropy model, was implemented to identify other disorders.

1 Introduction

Clinical texts are unstructured data that, when processed properly, can be of great value. Extracting key information from these documents can make medical notes more suitable for automatic processing. It can also help diagnose patients, structure their medical histories and optimize other clinical procedures and research.

The task of identifying mentions to medical concepts in free text and mapping these mentions to a knowledge base was recently proposed in ShARe/CLEF eHealth Evaluation Lab 2013, attracting the attention of several research groups worldwide (Pradhan et al., 2013). The task 7 in SemEval 2014 (Pradhan et al., 2014) elaborates in that previous effort focusing on the recognition and normalization of named entity mentions belonging to the UMLS semantic group *disorders*.

The paper is organized as follows: in section 2 we briefly present the data, section 3 contains the description of the methods and tools used in our system. Later, on sections 4 and 5 we provide the details of the three submitted runs and expose the official results. Finally, sections 6 and 7 include discussions on variations that could be done to improve performance and conclusions to be drawn from our participation in the task.

2 Data Description

The training data for SemEval 2014 Task 7 consists of the ShARe (Shared Annotation Resource) corpus, which contains clinical notes from MIMIC II database (Multiparameter Intelligent Monitoring in Intensive Care). The data were manually annotated for disorder mentions, normalized to a UMLS Concept Unique Identifier when possible, and marked as CUI-less otherwise.

Four types of reports where found in the corpus: 61 discharge summaries, 54 ECG reports, 42 ECHO reports and 42 radiology reports, for a total of 199 training documents, each containing several disorder mentions.

3 Methods Used

3.1 Named-Entity Recognition

Using the Java libraries Apache OpenNLP[1] and Maxent[2], a maximum entropy model was implemented for Named Entity Recognition (NER). Two types of classifiers were built: the first one using the library's default configuration, and a second one including additional features. The default model includes the following attributes: target word, two words of context at the left of the target word, two words of context at the right of the target word, type of token for target word (capitalized word, number, hyphen, commas, etc.), and type of token for words in the context.

[1] http://opennlp.apache.org
[2] http://maxent.sourceforge.net/about.html

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 424–427,
Dublin, Ireland, August 23-24, 2014.

For the enhanced model, we included n-grams at character level extracted from the target word, going from two to five characters.

OpenNLP uses the BIO tagging scheme, which marks each token as either beginning a chunk, continuing it, or not in a chunk, therefore, this model cannot identify discontinuous terms. Given this, we excluded discontinuous term annotations from the training data, and trained the model with the resulting corpus.

During the experiments, we also considered POS (Part of Speech) tags obtained with the OpenNLP library, POS tags obtained with the Stanford Java library and the number of characters in each token. However, we decided not to include any of these because accuracy decreased when using them.

3.2 Weirdness Measure

According to preliminary experiments, the chosen enhanced NER method exhibited low precision, i.e. a high number of false positives. To deal with this problem we calculated a measure for the specificity of a candidate named entity with respect to a specialized corpus, this quantity is based on the weirdness (Ahmad et al., 1999) of the candidate words. Having a general corpus C_g and a specialized corpus C_s, where w_g and w_s refer to the number of occurrences of a word w in each corpus and t_s and t_g to the total count of words in each corpus, the weirdness of a word is defined as follows:

$$Weirdness(w) = \frac{w_s}{t_s} / \frac{w_g}{t_g}$$

Those words that are common to any domain will very likely have a low weirdness score, while those with a high weirdness score indicate w is not used in the general corpus as much as in the specialized one, meaning it probably corresponds to specialized vocabulary.

Using around 1000 books from the Guttenberg Project as the general corpus, and the terms in UMLS as the specialized corpus, we applied the weirdness measure to those words that, according to the NER model, are disorders. By keeping only those with high weirdness measures, we prevent our system from tagging words that are not even medical vocabulary, thus reducing the amount of false positives.

3.3 Metamap

For identifying and mapping disorders included in the UMLS meta thesaurus to its corresponding CUI, we explored two tools. Both of them find candidates in the document and give the possible CUIs for each; in both cases, we selected the CUI that belongs to the UMLS semantic group *disorders*, as specified in the task description.

The first tool we explored is Metamap. For processing the documents, we use the following Metamap features: allow concept gap and word sense disambiguation.

After processing a document, the results were filtered, keeping only those tags that were mapped to a CUI that belongs to one of the following UMLS semantic types: *congenital abnormality, acquired abnormality, injury or poisoning, pathologic function, disease or syndrome, mental or behavioral dysfunction, cell or molecular dysfunction, experimental model of disease, anatomical abnormality, neoplastic process*, and *signs or symptoms*.

3.4 T-mapper

As an alternative to Metamap we experimented with T-mapper[3], an annotation tool developed at MindLab[4] that works in languages different than English and with any knowledge source (i.e. not only UMLS). The method implemented by T-mapper is inspired by the one in Metamap, with some modifications. The method works as follows:

1. Indexing and vocabulary generation: an inverted index and other data structures are built to perform fast lookups over the dictionary and the vocabulary list in C_g and C_s.

2. Sentence detection and tokenization: the input text is divided into sentences and then each sentence is divided into tokens using a whitespace as separator.

3. Spelling correction: to deal with noise and simple morphological variations, each token that does not match a word within the vocabulary is replaced by the most frequent word among the most similar words found above a threshold of 0.75. The similarity is computed using a normalized score based on the Levensthein distance.

[3] https://github.com/lariverosc/tmapper
[4] http://mindlaboratory.org/

4. Candidate generation and scoring: a subset that contains all the terms that match at least one of the words in the sentence is gene–rated, the terms contained in this set are called candidates. Once this subset is built, each of the candidate terms is scored using a simplified version of Metamap's scoring function (Aronson, 2001). In comparison, T-mapper's function uses only variation, coverage and cohesiveness as criteria, excluding centrality, since it is language dependant.

5. Candidate selection and disambiguation: the score computed in the previous step is used to choose the candidates that will be used as mappings. Ambiguity can occur because of two reasons: a tie in the scores or by overlapping over the sentence tokens. In the first case, the Lin's measure (Lin, 1998) is used as disambiguation criteria between the candidates and the previous detected concepts. In the second case, the most concrete term is chosen according to the UMLS hierarchy.

4 System Submissions

The team submitted three runs. The *run 0* was intended as a baseline; *run 1* used Metamap for UMLS concept mapping and *run 2* did this using T-mapper. Both *run 1* and *run 2* used the enhanced features for NER and applied the weirdness measure.

For *run 0*, the documents were processed with Metamap and those concepts mapped to a CUI belonging to one of the desired UMLS semantic types were chosen. Parallel to this, the do–cument was tagged using the default NER model. Finally, results were merged, preferring Metamap mapping outputs in the cases where a concept was mapped by both tools (in an ideal scenario, all terms mapped by Metamap would have also been mapped by the NER model).

Run 1 differs from *run 0* in two steps of the process: the NER model included the enhanced features described previously and its output was filtered, keeping only those concepts whose weirdness measure exceeds 0.7. For multiword concepts the weirdness of each word was aggregated.

Finally, *run 2* was equal to *run 1*, with the di–fference that T-mapper was used to map concepts to the UMLS meta thesaurus.

Rank	Run	Strict P	Strict R	Strict F
1	*best*	0.843	0.786	0.813
31	2	0.561	0.534	0.547
32	1	0.578	0.515	0.545
37	0	0.321	0.565	0.409

Table 1: Official results for task A obtained by the best system and our runs (ranked by exact acc.)

Rank	Run	Strict Accuracy
1	*best*	0.741
19	2	0.461
21	0	0.435
24	1	0.411

Table 2: Official results for task B obtained by the best system and our runs (ranked by exact acc.)

5 Results

For both task A and B, *run 2* produced the best performance among our systems. In Table 1 the results of the three runs are presented, together with the information of the system with the best performance among all participating teams (labeled as *best*). The position in the ranking is from a total of 43 submitted systems. Table 2 shows analogous results for Task B, where 37 systems were submitted.

Even though the official ranking is based on the strict accuracy, which only considers a tag to be correct if it matches exactly both the first and last characters, a relaxed accuracy is also provided by the organizers. This second scoring measure considers a tag to be correct if it has an overlap with the actual one. Tables 3 and 4 show these results.

In both tables 1 and 3, P stands for Precision, R for Recall, and F for F-score. The ranking is based on the F-score.

6 Discussion

The system that gave the best results for both tasks was the one based on T-mapper. Certain features

Rank	Run	Relax P	Relax R	Relax F
1	*best*	0.916	0.907	0.911
35	2	0.769	0.677	0.720
37	1	0.777	0.654	0.710
40	0	0.439	0.725	0.547

Table 3: Official results for task A obtained by the best system and our runs (ranked by relaxed acc.)

Rank	Run	Relaxed Accuracy
1	*best*	0.928
11	2	0.863
19	0	0.797
21	1	0.771

Table 4: Official results for task B obtained by the best system and our runs (ranked by relaxed acc.)

of this tool make this finding particularly inte–resting: it works for any language and ontology, and it is considerably faster than Metamap. While Metamap took 581 minutes to tag 133 documents, T-mapper only required 96 minutes (133 is the number of documents in the test set).

One aspect that might have damaged the performance of our system is the fact that, unlike most of the teams, we did not use the development data for training. However, there are still a number of changes that could be made, which would very likely improve the accuracy of our system. First, the tokenizer used for the NER model and for T-mapper were too simple. Separation was done based on blank spaces, therefore slashes, certain punctuation marks and hyphens might not be treated properly.

In addition to this, the spell checker used by T-mapper also needs to be improved. Currently, it gives a ranked list of options for each word that should be replaced, and automatically chooses the first one in the ranking. However, the best match is often the second or third in the list. Changing the criteria used to choose the replacement, taking into account word sense disambiguation, would enhance the accuracy of T-mapper.

The weirdness measure is also something that should be reconsidered, since it would be inte–resting to use a metric that responds better to unseen terms. And in case this was still the chosen measure, other training corpora could work better, since an ontology might lack words that are currently used in a medical context but do not have a CUI, and it also fails to give a notion of which words are more frequently used than others. It is not easy, however, to replace UMLS as corpus, since it is not easy to compete with its size and richness.

Finally, the OpenNLP NER system does not recognize discontinuous terms. Therefore, no CUI-less term with a gap can currently be identified by the system. For this reason, the NER method should be changed to one that allows this type of mentions to be present in texts.

For Task B, it is very interesting to see the di–fference between the strict and relaxed evaluation rankings. We go from being in position 19 to being in position 11. This might be partially explained by some of the flaws previously mentioned; in particular, the weak tokenizer and the incapability to identify CUI-less terms with gaps.

7 Conclusion

We participated with three runs in the Semeval 2014 task for analysis of clinical texts. Even though the performance of our runs indicates they still need to be enhanced in order to be com–petitive in this specific task, the performance of the run based on T-mapper compared to that of the ones that use Metamap proves that T-mapper is a viable alternative for mapping concepts to clinical terminologies. Moreover, T-mapper should also be considered for cases in which Metamap cannot be used: languages other than English and terminologies other than UMLS.

References

Khurshid Ahmad, Lee Gillam, Lena Tostevin, et al. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *TREC*.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304.

Sameer Pradhan, Noemie Elhadad, Brett R. South, David Martinez, Lee Chistensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova. 2013. Task 1: ShARe/CLEF eHealth evaluation lab 2013. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, Valencia, Spain, September.

Sameer Pradhan, Noemie Elhadad, Wendy W. Chapman, and Guergana Savova. 2014. Semeval-2014 task : Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland, August.