SZTE-NLP: Clinical Text Analysis with Named Entity Recognition

Melinda Katona and Richárd Farkas
Department of Informatics
University of Szeged
Árpád tér 2., Szeged, 6720, Hungary
{mkatona,rfarkas}@inf.u-szeged.hu

Abstract

This paper introduces our contribution to the SemEval-2014 Task 7 on ”Analysis of Clinical Text”. We implemented a system which combines MetaMap taggings and Illinois NER Tagger. MetaMap is developed to link the text of medical documents to the knowledge embedded in UMLS Metathesaurus. The UMLS contains a very rich lexicon while the promise of a NER system is to carry out context-sensitive tagging. Our system's performance was 0.345 F-measure in terms of strict evaluation and 0.551 F-measure in terms of relaxed evaluation.

1 Introduction

Clinical notes and discharge summaries from the patient's medical history contain a huge amount of useful information for medical researchers and also for hospitals. The automatic identification of these unstructured information is an important task for analysis of free-text electronic health records. Natural Language Processing (NLP) techniques provide a solution to process clinical documents and to help patients understand the contents of their clinical records (Tang et al., 2012; Lee et al., 2004).

In this paper we introduce an approach which discovers mentions of disorders in the free-text of discharge summaries. The system participated in the *SemEval-2014 Task 7: Analysis of Clinical Text, Task A.*

Task A aims at the identifying of mention concepts that belong to the UMLS (Bodenreider, 2004) semantic group ”disorders” and Task B is for mapping from each mention to a unique UMLS/SNOMED-CT CUI (Concept Unique Identifiers). Here are a few examples from the task description:

- The rhythm appears to be **atrial fibrillation**.

  „atrial fibrillation” is a mention of type disorders with CUI C0004238

- The **left atrium** is moderately **dilated**.

  „left atrium [...] dilated” is a mention of type disorders with CUI C0344720

- 53 year old man s/p **fall from ladder.**

  „fall from ladder” is a mention of type disorders with CUI C0337212

Many approaches have been published to solve these problems cf. (Skeppstedt et al., 2012; Pestian et al., 2007).

2 Approach

After a text-normalization step we run a Named Entity Recogniser (NER) on the documents. This NER model was trained on the training set of the shared task. It also employs a dictionary gathered from UMLS through MetaMap tagging. Our initial experiments revealed that MetaMap (Aronson and Lang, 2010) in its own gives a very poor precision hence we decided to investigate a NER approach which takes the context also into account.

2.1 Normalization

Clinical reports contain numerous special annotations, such as anonymized data (for example patient name), etc. We made the following steps to normalize texts:

- We removed the unnecessary characters, such as . , ! ? # : ; — = + * ˆ
- Then replaced the [****] anonymized tags with REPLACED_ANONYMOUS_DATA notation.

2.2 UMLS Dictionary

Our NER system constructs features from dictionaries as well. We created a dictionary from UMLS with the help of MetaMap for incorporating external knowledge into the NER. The use of a specialized dictionary is important because it contains phrases that occur in clinical texts.

MetaMap (Aronson and Lang, 2010) is developed to link the text of medical documents to the knowledge embedded in UMLS Metathesaurus. MetaMap employs natural language processing techniques working at the lexical/syntactic levels, for example handling acronyms/abbrevations, POS tagging, word sense disambiguation and so on.

Both the test and training datasets were used for creating our dictionary. We used MetaMap to collect disorders from raw texts. After that, we removed the redundant and most frequently used common words, based on a list of the 5000 most frequent English words according to the Google's n-gram corpus[1].

2.3 Named Entity Recognition

In the task "Analysis of Clinical Text", our task is to recognize mentions of concepts that belong to the UMLS semantic group "disorder", which can be viewed as a subclass of named entities, so NER approach is effective for this assignment.

For training, we used the Illinois Named Entity Recognition (Ratinov and Roth, 2009) system. By default, Illinois NER contains Wikipedia gazetters and categories, but in this task, we need one or more dictionary which contains disorders and other clinical text terminology.

NER is typically viewed as a sequence labeling problem. The typical models include HMM (Rabiner, 1989), CRF (Lafferty et al., 2001) and sequential application of Perceptron or Winnow (Collins, 2002). Illinois NER has several inference algorithms: Viterbi, beamsearch, greedy left-to-right decoding. In our approach, we used beamsearch. The beamsize was 3. Initially, we used bigger beamsize, but our empirical studies showed that applying a small beamsize is more effective.

Beside the decoding algorithm, an important question that has been studied extensively in the context of shallow parsing which was somewhat overlooked in the NER literature is the representation of text segments. Illinois NER contains several representation schemes such as BIO and BILOU - two of the most popular schemes. The BIO scheme is employed to train classifiers that identify **B**eginning, the **I**nside and the **O**utside of the text segment. The BILOU scheme is employed to train classifiers that identify the **B**eginning, the **I**nside and the **L**ast tokens of multi-token chunks as well as **U**nit-length chunks. We used the BILOU scheme.

The key intuition behind non-local features in NER has been that identical tokens should have identical label assignments. Ratinov and Roth (2009) consider three approaches proposed in the literature namely context aggregation, two-stage prediction aggregation and extended prediction history. The combination of these approaches is more stable and better than any approach taken alone.

In our experiments we used the combination of context aggregation and two-stage prediction aggregation. Context aggregation is the following approach in Illinois NER: for each token instance x_i we used the tokens in the window of size two around it as features: $c_i = x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}$. If the same token ($t$) appears in several locations in the text for each instance x_{i_j} ($x_{i_1}, x_{i_2}, \ldots, x_{i_N}$). We also aggregated the context across all instances within 200 tokens.

Context aggregation as done above can lead to an excessive number of features. Some instances of a token appear in easily-identifiable contexts. The resulting predictions were used as features at a second level of inference. This is a two-stage prediction aggregation.

3 Experimental Results

Our system was developed and trained only on the training set provided by the organizers and was evaluated on the test set. The performance was evaluated by Precision, Recall and F-measure in both "strict" and "relaxed" modes. "Strict" means that a concept is recognized correctly if the starting and ending offsets are the same as in gold standard and "relaxed" means that a disorder mention is correctly recognized as long as it overlaps with the gold standard disorder mention.

3.1 Dataset

For training and testing, we used the datasets provided by the shared task organisers. The train-

[1] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

	Strict			Relaxed		
	P	R	F	P	R	F
original NER	0.508	0.225	0.312	0.874	0.378	0.528
NER with normalization	0.509	0.229	0.316	0.875	0.383	0.528
NER with normalization and full dictionary	0.512	0.226	0.313	0.878	0.378	0.533
NER with normalization and filtered dictionary	0.516	0.232	0.320	0.890	0.390	0.542

Table 1: Evaluation results of our system on the training set (**P** - Precision, **R** - Recall, **F** - F-score).

ing dataset contains of 398 notes from different clinical documents including radiology reports, discharge summaries, and ECG/ECHO reports. For each note, disorder entities were annotated based on a pre-defined guideline and then mapped to SNOMED-CT concepts represented by UMLS CUIs. The reference UMLS version was 2012AB. If a disorder entity could not be found, it was marked as CUI-less, otherwise marked with CUI identifier.

The training set was used for system development, and we evaluated the system on the test set of 133 notes.

3.2 Results

We examined the contribution of our systems' steps. Table 1 summarizes the results where the first column contains result of named entity tagger without any modification. Normalization gave only a marginal improvement in accuracy. Next, we employed all MetaMap matches as a feature for the NER module. This decreased recall, because NER identified a lot of unnecessary expression. In our final and submitted system, we filtered this dictionary as described in the previous section.

Lastly, Table 2 shows our official evaluation results.

	Strict	Relaxed
Precision	0.547	0.884
Recall	0.252	0.401
F-score	0.345	0.551

Table 2: Results of our submission on the test set.

4 Error Analysis

In both strict and relaxed evaluation modes, precision is high but recall is low. We have found three important source of errors:

- multiple meaning words
- unknown disorders
- discontinuous phrases

A named entity tagger with context-aggregation mode does not monitor multiple meanings, so if a word has more occurrence, but in other meaning, it will be a bad tagging. For example

*"**Seizure-like** activity with clamped jaw and left lip twitching was then noted after several days of treatment. [...] Despite these therapies, she failed to recover, and began to show further signs of increasing intracranial pressure with increasing **seizure** activity and posturing [...]"*

Our sequence labeling approach cannot recognize discontinuous phrases. Even when every token was marked, we took only continuous sequences as named entity mentions. For example the sentence

*"The left **ventricular cavity** is moderately **dilated**."*

yields three errors in the strict evaluation scenario. We did not recognise the three token-long phrase while predicted two false positive mentions. We also note that this shortcoming of our approach is the reason for the huge difference between the achieved strict and relaxed scores.

The last error category is unrecognised disorders. For instance,

*"The PICC line was trimmed to the appropriate length and advanced over the 0.018 wire with the tip int the **axillary vein**"*

Named entity tagger identified hepatitis B, but hepatitis C not because dictionary does not contain it. Expansion of dictionary increase accuracy.

5 Conclusion

In this paper we examined a machine learning based disorder recognition system using MetaMap and Illinois Named Entity Recognition. Illinois NER uses different dictionaries for training. We created a new filtered in-domain dictionary and we showed that this dictionary is an important factor for accuracy. The results achieved on the training set and the test set show that the proposed clinical dictionary creation procedure is efficient.

Acknowledgements

Melinda Katona is supported by the European Union and co-funded by the European Social Fund. Project title: ”Telemedicine-focused research activities on the field of Matematics, Informatics and Medical sciences“ (project number: TÁMOP-4.2.2.A-11/1/KONV-2012-0073). Richárd Farkas was partially funded by the ”Hungarian National Excellence Program“ (TÁMOP 4.2.4.A/2-11-1-2012-0001), co-financed by the European Social Fund.

References

Alan R. Aronson and Franois-Michel Lang. 2010. An Overview of MetaMap: Historical Perspective and Recent Advances. *JAMIA*, 17:229–236.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32:267–270.

Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 1 – 8.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282 – 289.

Chih Lee, Wen-Juan Hou, and Hsin-Hsi Chen. 2004. Annotating Multiple Types of Biomedical Entities: A Single Word Classification Approach. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA ’04, pages 80–83.

John P. Pestian, Christopher Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A Shared Task Involving Multi-label Classification of Clinical Free Text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104.

Lawrence Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech recognition. *Proceedings of the IEEE*, 77:257–286.

L. Ratinov and D. Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *CoNLL*, 6.

Maria Skeppstedt, Maria Kvist, and Hercules Dalianis. 2012. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*.

Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2012. Clinical Entity Recognition Using Structural Support Vector Machines with Rich Features. In *Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics*, DTMBIO ’12, pages 13–20.