UTU: Disease Mention Recognition and Normalization with CRFs and Vector Space Representations

Suwisa Kaewphan[1,2,3]*, Kai Hakaka[1]*, Filip Ginter[1]
[1] Dept. of Information Technology, University of Turku, Finland
[2] Turku Centre for Computer Science (TUCS), Turku, Finland
[3] The University of Turku Graduate School (UTUGS), University of Turku, Finland
sukaew@utu.fi, kahaka@utu.fi, ginter@cs.utu.fi

Abstract

In this paper we present our system participating in the SemEval-2014 Task 7 in both subtasks A and B, aiming at recognizing and normalizing disease and symptom mentions from electronic medical records respectively. In subtask A, we used an existing NER system, NERsuite, with our own feature set tailored for this task. For subtask B, we combined word vector representations and supervised machine learning to map the recognized mentions to the corresponding UMLS concepts. Our system was placed 2nd and 5th out of 21 participants on subtasks A and B respectively showing competitive performance.

1 Introduction

The SemEval 2014 task 7 aims to advance the development of tools for analyzing clinical text. The task is organized by providing the researchers annotated clinical records to develop systems that can detect the mentions of diseases and symptoms in medical records. In particular, the SemEval task 7 comprises two subtasks, recognizing the mentions of diseases and symptoms (task A) and mapping the mentions to unique concept identifiers that belong to the semantic group of disorders in the Unified Medical Language System (UMLS).

Our team participated in both of these subtasks. In subtask A, we used an existing named entity recognition (NER) system, NERsuite, supplemented with UMLS dictionary and normalization similarity features. In subtask B, we combined compositional word vector representations with supervised machine learning to map the recognized mentions from task A to the UMLS concepts. Our best systems, evaluated on strict matching criteria, achieved F-score of 76.6% for the subtask A and accuracy of 60.1% for the subtask B, showing competitive performance in both tasks.

2 Task A: Named Entity Recognition with NERSuite

The ML approach based on conditional random fields (CRFs) has shown to have state-of-the-art performance in recognizing the biological entities. We thus performed task A by using NERsuite, an existing NER toolkit with competitive performance on biological entity recognition (Campos et al., 2013).

NERsuite is a NER system that is built on top of the CRFsuite (Okazaki, 2007). It consists of three language processing modules: a tokenizer, a modified version of the GENIA tagger and a named entity recognizer. NERsuite allows user-implemented features in addition to dictionary matching and features shown to benefit the systems such as raw token, lemma, part-of-speech (POS) and text chunk.

Prior to detecting the disease mentions by the recognizer module of NERsuite, the clinical text is split into sentences by using GENIA Sentence Splitter, a supervised ML system that is known to be well optimized for biomedical texts (Sætre et al., 2007). The sentences are subsequently tokenized and POS tagged.

To represent the positive entities, the “BIO” model was used in our system. The first tokens of positive mentions are labeled with “B” and the rest with “I”. Negative examples, non-entities, are thus labeled with “O”. This model was used for both contiguous and discontiguous entities.

The features include the normalization similarity (see Section 3.3), types of medical records (*discharge*, *echo*, *radiology* and *ecg*), and UMLS dic-

*These authors contributed equally.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 807–811,
Dublin, Ireland, August 23-24, 2014.

Trained Model	Precision	Recall	F-score
train + positive samples	77.3%	72.4%	74.8%
train + development	76.7 %	76.5%	76.6%

Table 1: The results of our different NERsuite models, announced by the organizers.

tionary matching in addition to NERsuite's own feature generation.

The UMLS dictionary is prepared by extracting the UMLS database for the semantic types demanded by the task. In addition to those 11 semantic types, "Finding" was also included in our dictionary since, according to its definition, the concept is also deemed relevant for the task. Due to the common use of acronyms, which are not extensively provided by UMLS, we also extended the coverage of our prepared UMLS dictionary by extracting medical acronyms from the UMLS database using regular expression.

We assessed the effect of dictionary matching by training the models with and without the compiled UMLS dictionary and evaluating against the development set. The model trained with dictionary features outperformed the one without. The best model was obtained by training the NERsuite with UMLS dictionary in case-number-symbol normalization mode. In this mode, all letters, numbers and symbols are converted to lower case, zero (0) and underscore (_) respectively.

The regularization parameter (C2) was selected by using development set to evaluate the best model. The default parameter (C2 = 1.0) gave the best performing system and thus was used throughout the work.

Finally, for the NER task, we submitted two models. The first model was trained with the original training data and duplicates of sentences with at least one entity mention. The second model was trained by using the combination of the first model's training data and development set.

2.1 Results and Discussions

Our NER system from both submissions benefited from the increased number of training examples while the more diverse training data set gave a better performance. The official results are shown in table 1.

The analysis of our best performing NER system is not possible since the gold standard of the test data is not publicly available. We thus simply analyze our second NER system based on the evaluation on the development data. The F-score of the system was 75.1% and 88.0% for the strict and relaxed evaluation criteria respectively. Among all the mistakes made by the system, the discontiguous entities were the most challenging ones for the NERsuite. In development data, the discontiguous entities contribute about 10% of all entities, however, only 2% were recognized correctly. On the contrary, the system did well for the other types as 73% were correctly recognized under strict criteria. This demonstrates that the "BIO" model has limitations in representing the discontiguous entities. Improving the model to better represent the discontiguous entities can possibly boost the performance of the NER system significantly.

3 Task B: Normalization with Compositional Vector Representations

Our normalization approach is based on continuous distributed word vector representations, namely the state-of-the-art method *word2vec* (Mikolov et al., 2013a). Our word2vec model was trained on a subset of abstracts and full articles from the PubMed and PubMed Central resources. This data was used as it was readily available to us from the EVEX resource (Van Landeghem et al., 2013). Before training, all non-alphanumeric characters were removed and all tokens were lower-cased. Even though a set of unannotated clinical reports was provided in the task to support unsupervised learning methods, our experiments on the development set showed better performance with the model trained with PubMed articles. This might be due to the size of the corpora, as the PubMed data included billions of tokens whereas the provided clinical reports totaled in over 200 million tokens.

The dimensionality of the word vectors was set to 300 and we used the continuous skip-gram approach. For other word2vec parameters default values were used.

One interesting feature demonstrated by Mikolov et al. (2013b; 2013c) is that the vectors conserve some of the semantic characteristics in element-wise addition and subtraction. In this task we used the same approach of simply summing the word-level vectors to create compositional vectors for multi-word entities and concepts, i.e. we looked up the vectors for every token appearing in a concept name or entity and summed them to form a vector to represent the whole phrase.

We then formed a lexicon including all preferred terms and synonyms of all the concepts in the subset of UMLS defined in the task guidelines. This lexicon is a mapping from the compositional vector representations of the concept names into the corresponding UMLS identifiers. To select the best concept for a recognized entity we calculated cosine similarity between the vector representation of the given entity and all the concept vectors in the lexicon and the concept with the highest similarity was chosen.

Word2vec is generally able to relate different forms of the same word to each other, but we noticed a small improvement in accuracy when possessive suffixes were removed and all tokens were lemmatized.

3.1 Detecting CUI-less Mentions

As some of the mentions in the training data do not have corresponding concepts in the semantic categories listed in the task guidelines, they are annotated as "CUI-less". However, our normalization approach will always find the nearest matching concept, thus getting penalized for wrong predictions in the official evaluation. To overcome this problem, we implemented three separate steps for detecting the "CUI-less" mentions. As the simplest approach we set a fixed cosine similarity threshold and if the maximal similarity falls below it, the mention is normalized to "CUI-less". The threshold value was selected using a grid search to optimize the performance on the official development set. Although this method resulted in decent performance, it is not capable of coping with cases where the mention has very high similarity or even exact match with a concept name. For instance our system normalized "aspiration" mentions into UMLS concept "Pulmonary aspiration" which has a synonym "Aspiration", thus resulting in an exact match. To resolve this kind of cases, we used similar approach as in the DNorm system (Leaman et al., 2013b), where the "CUI-less" mentions occurring several times in the training data were added to the concept lexicon with concept ID "CUI-less". As the final step we trained a binary SVM classifier to distinguish the "CUI-less" mentions. The classifier utilized bag-of-word features as well as the compositional vectors. The performance improvement provided by each of these steps is presented in table 2. This evaluation shows that each step increases the performance considerably, but the overall performance is still 3.9pp below perfect detection.

Method	Strict accuracy
B	43.6
T	48.4
T + L	53.5
T + L + C	55.4
O	59.3

Table 2: Evaluation of the different approaches to detect CUI-less entities on the official development set compared to a baseline without CUI-less detection and an oracle method with perfect detection. This evaluation was done with the entities recognized by our NER system instead of the gold standard entities. B = baseline without CUI-less detection, T = similarity threshold, L = Lexicon-based method, C = classifier, O = Oracle.

3.2 Acronym Resolution

Abbreviations, especially acronyms, form a considerable portion of the entity mentions in clinical reports. One of the problems in normalizing the acronyms is disambiguation as one acronym can be associated with multiple diseases. Previous normalization systems (Leaman et al., 2013b) handle this by selecting the matching concept with most occurrences in the training data. However, this approach does not resolve the problem of non-standard acronyms, i.e. acronyms that are not known in the UMLS vocabulary or in other medical acronym dictionaries. Our goal was to resolve both of these problems by looking at the other entities found in the same document instead of matching the acronym against the concept lexicon. With this approach for instance entity mention "CP" was on multiple occasions correctly normalized into the concept "Chest Pain", even though UMLS is not aware of this acronym for the given concept and in fact associates it with several other concepts such as "Chronic Pancreatitis" and "Cerebral Palsy". However, the overall gain in accuracy obtained from this method was only minor.

3.3 Normalization Feedback to Named Entity Recognition

While basic exact match dictionary features provide usually a large improvement in NER performance, they are prone to bias the system to high precision and low recall. As both noun and adjective forms of medical concepts, e.g. "atrium" and "atrial", are commonly used in clinical texts, the entities may not have exact dictionary matches. Moreover the different forms of medical terms may not share a common morphological root discovered by simple stemming methods, thus complicating approximate matching. In this task we tried to boost the recall of our entity recognition by feeding back the normalization similarity information as features. These features included the maximum similarity between the token and the UMLS concepts as a numerical value as well as a boolean feature describing whether the similarity exceeded a certain threshold.

In addition we experimented by calculating the similarities for bigrams and trigrams in a sliding window around the tokens, but these features did not provide any further performance improvements.

3.4 Other Directions Explored

The DNorm system utilizes TF-IDF vectors to represent the entities and concepts but instead of calculating cosine similarity, the system trains a ranking algorithm to measure the maximal similarity (Leaman et al., 2013a). Their evaluation, carried out on the NCBI disease corpus (Doğan et al., 2014), showed a notable improvement in performance compared to cosine similarity. In our analysis we noticed that in 39% of the false predictions made by our normalization system, the correct concept was in the top 10 most similar concepts. This strongly suggested that a similar ranking method might be beneficial with our system as well. To test this we trained a linear SVM to rerank the top 10 concepts with highest cosine similarity, but we were not able to increase the overall performance of the system. However, due to the strict time constraints of the task, we cannot conclude whether this approach is feasible or not.

As our compositional vectors are formed by summing the word vectors, each word has an equal weight in the sum. Due to this our system made various errors where the entity was a single word matching closely to several concepts with longer names. For instance entity "hypertensive" was falsely normalized to concept "Hypertensive cardiopathy" whereas the correct concept was "Hypertensive disorder". These mistakes could have been prevented to some extent if the more important words had had a larger weight in the sum, e.g. word "disorder" is of low significance when trying to distinguish different disorders. However, weighting the word vectors with their IDF values, document in this case being an UMLS concept, did not improve the performance.

Team	Strict accuracy	Relaxed accuracy
UTH_CCB	74.1	87.3
UWM	66.0	90.9
RelAgent	63.9	91.2
IxaMed	60.4	86.2
UTU	60.1	78.3

Table 3: Official evaluation results for the top 5 teams in the normalization task.

3.5 Results

The official results for the normalization task are shown in table 3. Our system achieved accuracy of 60.1% when evaluated with the official strict evaluation metric. This result suggests that compositional vector representations are a competitive approach for entity normalization. However, the best performing team surpassed our performance by 14.0pp, showing that there is plenty of room for other teams to improve. It is worth noting though that their recall in the NER task tops ours by 8.2pp thus drastically influencing the normalization results as well. To evaluate the normalization systems in isolation from the NER task, a separate evaluation set with gold standard entities should be provided.

4 Conclusions

Overall, our NER system can perform well with the same default settings of NERsuite for gene name recognition. The performance improves when relevant features, such as UMLS dictionary matching and word2vec similarity are added. We speculated that representing the nature of the data with more suitable model can improve the system performance further. As a part of a combined system, the improvement on NER system can result in the increased performance of normalization system.

Our normalization system showed competitive results as well, indicating that word2vec-based vector representations are a feasible way of solving the normalization task. As future work we would like to explore different methods for creating the compositional vectors and reassess the applicability of the reranking approach described in section 3.4.

Acknowledgements

Computational resources were provided by CSC — IT Center for Science Ltd, Espoo, Finland. This work was supported by the Academy of Finland.

References

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. Gimli: open source and high-performance biomedical name recognition. *BMC bioinformatics*, 14(1):54.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, Feb.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013a. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2013b. NCBI at 2013 ShARe/CLEF eHealth Shared Task: Disorder normalization in clinical notes with DNorm. In *Proceedings of the Conference and Labs of the Evaluation Forum*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR Workshop*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Y Matsubyashi, and Tomoko Ohta. 2007. AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In *Proceedings of the BioCreative II*, pages 209–212.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PLoS ONE*, 8(4):e55814.