Combining Conditional Random Fields and Word Embeddings for the CHEMDNER-patents task

Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez

Computer Science, University Carlos III of Madrid,
28911 Madrid, Spain
{isegura,vspaniag,pmf}@inf.uc3m.es
http://labda.inf.uc3m.es

Abstract. This paper describes our system developed for the BioCreative V CHEMDNER-patents task. The system is based on Conditional Random Field (CRF) trained with standard features used in current Named Entity Recognition (NER) systems as well as word clusters generated by the Word2Vec tool and a lexicon feature from the DINTO ontology.

Key words: Word Embedding, Conditional Random Field, Drug name recognition

1 Introduction

The goal of the CHEMDNER-patents task[1] is the extraction of chemical and biological data from medicinal chemistry patents. The previous edition, BioCreative IV CHEMDNER [2], was devoted to NER focusing on detecting chemical entity mentions from PubMed abstracts. Most participating systems used supervised machine learning techniques, being CRF the most used technique. The systems used different types of features: word level features, lexicon features and document features (for example, coocurrences of mentions).

Our work explores the effectiveness of new features for the CEMP (chemical entity mention in patents) task, in particular, word clusters generated using the Word2Vec tool [4], a word embedding model based on a neural network (NN). The essential assumption of word embedding is that semantically close words will have similar vectors. Word embeddings have shown promising results in NLP tasks, such as named entity recognition, sentiment analysis or parsing [8, 6, 7]. However, to the best of our knowledge, this technique has hardly ever been exploited in drug name recognition [3, 5]. In addition, our goal is also to study whether the DINTO ontology[2] [1] can provide valuable information for this task. The DINTO ontology contains a total of 8,786 drugs.

[1] http://www.biocreative.org/tasks/biocreative-v/track-2-chemdner/
[2] http://www.obofoundry.org/cgi-bin/detail.cgi?id=DINTO

2 Systems description and methods

Encouraged by the good results of the CRF-based methods in the previous edition of CHEMDNER [2], we propose a system based on CRF with the following feature set:

- The context window of three tokens to its right and to its left in the sentence. The context window also includes the current token.
- POS tags and lemmas in the context window are also considered.
- An orthography feature: upperInitial, allCaps, lowerCase and mixedCaps.
- A feature representing the type of token: word, number, symbol or punctuation.
- A binary feature that indicates whether the current token was found in the DINTO ontology.
- A feature representing the long word shape of the current token. This feature is defined by mapping any uppercase letter, lowercase letter, digit, and other characters to X x 0 and O respectively. For example, the long word shape of "C1-6alkyl" is X0O0xxxxx.
- A feature representing the brief word class of the current token. Consecutive uppercase letters, lowercase letters, digits, and other characters map to X, x, 0, and O, respectively. For example, the brief word shape of "C1-6alkyl" is X0O0x.

In addition, we used a word cluster as additional feature to represent the current token. The Word2vec tool includes a utility to compute word clusters using a k-means clustering algorithm. Thus, the word clusters were obtained from the word embeddings trained with the Word2Vec tool on the latest wikipedia dump[3] as well as on the 2013 release of MedLine. Word clusters represent words at a higher level abstraction that may help to recognize even those chemical compound and drug mentions that are not observed in the training set. We performed experiments for different values of k (100, 200, 300 and 400) in the k-means algorithm.

3 Discussion

The main hypothesis of this work is that incorporating word embeddings as features into a CRF model could help to recognize unseen or very rare drug mentions in the training set. However, the experiments on the development set showed that the use of word embeddings features does no seem to provide a significant improvement on the CRF-based system using only the standard features for the task of named entity recognition.

From the experiments on the development set, we can observe that the use of DINTO seems to achieve a very light increase in the number of true positives and a small descrease in the number of false negatives. However, these results

[3] http://dumps.wikimedia.org/

	TP	FP	FN	P	R	F1
baseline	25389	4436	6753	85%	79%	82%
CRF+ Dinto	25418	4437	6724	85%	79%	82%
W2VMD-100K	25687	4473	6455	85%	80%	82%
W2VMD-200K	25632	4414	6510	85%	80%	82%
W2VMD-300K	25615	4502	6527	85%	80%	82%
W2VMD-400K	25628	4501	6514	85%	80%	82%
W2VWIKI-100K	25680	4483	6462	85%	80%	82%
W2VWIKI-200K	25681	4546	6461	85%	80%	82%
W2VWIKI-300K	25646	4523	6496	85%	80%	82%
W2VWIKI-400K	25620	4580	6522	85%	80%	82%
DINTO-W2VMD-100K	25702	4428	6440	85%	80%	83%
DINTO-W2VMD-200K	25589	4441	6553	85%	80%	82%
DINTO-W2VMD-300K	25672	4479	6470	85%	80%	82%
DINTO-W2VMD-400K	25551	4488	6591	85%	79%	82%
DINTO-W2VWIKI-100K	25650	4445	6492	85%	80%	82%
DINTO-W2VWIKI-200K	25577	4458	6565	85%	80%	82%
DINTO-W2VWIKI-300K	25501	4533	6641	85%	79%	82%
DINTO-W2VWIKI-400K	25648	4529	6494	85%	80%	82%

Table 1. Experimental results on the CEMP development dataset. DINTO means that the system uses the DINTO ontology; WIKI means that the wikipedia corpus was used to train the word2vec models, while MD means that the corpus used was MedLine. Finally, k is the number of clusters for the K-means algorithm.

may be not statistically significant. Word2vec clusters from MedLine seem to achieve a light improvement of 1% in recall, but F1 is not increased. There are no significant difference between number of clusters used. The combination of DINTO plus word2vec clusters with k=100, it is the best model with a F1 of 83%. With k=400, the recall decreases a 1%. Both word2vec clusters from MedLine and Wikipedia achieve very close results. Based on the previous observations, we decided to use the following settings for the runs:

- run1: DINTO-W2VMD-100K (the word clusters were trained on MedLine).
- run2: W2VMD-100K (DINTO is not used in this run).
- run3: W2VWIKI-200K (DINTO is not used in this run).
- run4: DINTO-W2VWIKI-200K (the word clusters were trained on Wikipedia).
- run5: CRF+ DINTO (This run does not include the word clusters).

Table 2 shows the results obtained on the test dataset by all our runs and the results of the top participating system. Our best run (run 1) achieved a recall of 82.15% and precision of 86.3% (F1 of 84.17%). Our runs achieve results very close between them (see Table 2). In general, we obtain better performance on the test dataset than the development dataset (see Table 3).

As a result of the ranking, our 5 runs were placed in the positions 44 to 49 over a total of 93 submissions. Our runs and the top system achieve very close performance in terms of precision (only one point of difference), however the recall of the top system is almost 9% higher than ours.

	Precision	Recall	F-score
Top system	87.52%	91.29%	89.37%
Run 1	86.3%	**82.15%**	**84.17%**
Run 2	**86.32%**	81.95%	84.08%
Run 3	86.21%	81.99%	84.04%
Run 4	86.3%	81.85%	84.02%
Run 5	86.27%	82.02%	84.09%

Table 2. CEMP results on the test dataset.

Acknowledgments.

This work was supported by TrendMiner project [FP7-ICT287863] and by eGovernAbility-Access project (TIN2014-52665-C2-2-R).

References

1. Herrero-Zazo, M., Segura-Bedmar, I., Hastings, J., Martínez, P.: Dinto: Using owl ontologies and swrl rules to infer drug–drug interactions and their mechanisms. Journal of chemical information and modeling (2015)
2. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: Chemdner: The drugs and chemical names extraction challenge. J Cheminform 7(Suppl 1), S1 (2015)
3. Liu, S., Tang, B., Chen, Q., Wang, X., Fan, X.: Feature engineering for drug name recognition in biomedical texts: Feature conjunction and feature selection. Computational and mathematical methods in medicine 2015 (2015)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR 2013 Workshop Track (2013)
5. Segura-Bedmar, I., Suarez-Paniagua, V., Martinez, P.: Exploring word embedding for drug name recognition. In: Sixth Workshop on Health Text Mining and Information Analysis (2015)
6. Socher, R., Bauer, J., Manning, C.D., Ng, A.Y.: Parsing with compositional vector grammars. In: In Proceedings of the ACL conference. Citeseer (2013)
7. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP). vol. 1631, p. 1642. Citeseer (2013)
8. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 384–394. Association for Computational Linguistics (2010)