Combining the BANNER tool with the DINTO ontology for the CEMP task of BioCreative V.5

Cristóbal Colón-Ruiz, Isabel Segura-Bedmar, and Paloma Martínez

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

Abstract. This paper describes our system for the Chemical Entity Mention in Patents (CEMP) task of BioCreative V.5. The system consists of an adaptation of the BANNER tool, which is based on Conditional Random Fields (CRF) and has provided satisfactory results in the biomedical domain. In addition to the features provided by the tool for the recognition of entities in biomedical texts, a lexical feature is added using the DINTO ontology which will be combined with other ontologies such as ChEMBL and DrugBank.

Key words: BANNER, Conditional Random Fields, DINTO ontology, Chemical entity recognition

1 Introduction

Information related to drugs and chemical compounds constitutes an important pillar for research in the area of biology and biomedical sciences as well as for chemical experts. The enormous number of topics in which the chemical entities are present increases the interest for an efficient access to this information.

Natural language processing (NPL) and text mining technologies are one of the keys to improving access to this type of information from unstructured data such as patents, therefore, the CEMP task of BioCreative V.5 aims to detect chemical entities in medical chemistry patent abstracts automatically.

After an analysis with reference to previous editions of the task [3], it was verified that a large number of participants used supervised learning systems, being the conditional random fields one of the most representative techniques.

Our work seeks to verify the effectiveness of BANNER tool [1] for the recognition of chemical entities and consequently the effectiveness of the features provided by the tool. In addition, we analyze the contribution of the DINTO ontology [2], which contains a large number of chemical entities that can provide more information to the system.

2 Systems description and methods

Taking into account the results provided by CRF-based systems in previous editions, we propose the use of BANNER tool, an entity recognition system, based on conditional random fields and designed to increase domain independence.

Banner has a 3-stage pipeline, where the input is a sentence. During the first process, the sentences are tokenized. Then each token is represented by a series of features which are described below:

- The part of speech which the token forms in the sentence, that is, the process of assigning to each token its grammatical category within the text in relation to the adjacent words.
- The lemmatization of tokens, where each lemma represents the accepted form for all variations of words.
- Prefixes and suffixes up to 2, 3 and 4 characters for each token.
- A subsequence of tokens within the sentence, for this case, bigrams and trigrams are considered in the system.
- The normalization of tokens in word classes, replacing uppercase letters with 'A', lowercase ones with 'a', numbers with '0' and all other characters with 'x'. For example, the mention "C1-4alkyl" is a word class of type "A0x0aaaaa".

As a new feature, in addition to using the dictionaries provided by BANNER, we use the DINTO ontology to generate a dictionary of 13294 mentions. This feature indicates whether or not the token is found within the given dictionary. However, in addition to using the resources mentioned, other dictionaries are also used as those provided by the ChEMBL and DrugBank ontologies, as well as combinations of all of them.

For the labeling of the tokens, several IOB tagging schemes have been used (I = inside of a entity, O = outside of a entity, B = beginning of a entity), and finally, the CRF models are trained using the features of all tokens of the sentences provided by the training set.

3 Discussion

The main hypothesis of this work is that the incorporation of a dictionary provided by the DINTO ontology could help to identify and recognize mentions in the set of patents. Another of the main reasons is to adapt the tool of BANNER to the domain of the recognition of chemical entities to verify the independence of the tool based on the results obtained.

To test the tool, as well as the incorporation of features, we have performed a series of experiments with our development set (See Table 1). Because no development set was provided for the current CEMP task, we split the training set provided with 21000 instances into a new training set of 14000 instances and another set of 7000 instances for testing.

	IOB schema	dictionaries	TP	FP	FN	P	R	F1
base line	IOB		25811	5227	6329	0.8315	0.8030	0.8170
	IOBEW		25863	5236	6277	0.8316	0.8046	0.8179
	IO		25919	4986	6221	0.8386	0.8064	0.8222
BANNER+Dictionaries	IO	ChEMBL	25865	4913	6275	0.8403	0.8047	0.8221
	IO	DrugBank	25921	5079	6219	0.8361	0.8065	0.8210
	IO	ChEMBL+DrugBank	25906	5045	6234	0.8370	0.8060	0.8212
	IO	DINTO	25986	5064	6154	0.8369	0.8085	0.8224
	IO	DINTO+ChEMBL	25989	5017	6151	0.8381	0.8086	0.8231
	IOBEW	DINTO+ChEMBL	25879	5233	6261	0.8318	0.8051	0.8182
	IOB	DINTO+ChEMBL	25877	5256	6263	0.8311	0.8051	0.8179

Table 1. CEMP results on the development dataset. DINTO means that the dictionary used comes from the DINTO ontology; ChEMBL means that the dictionary used comes from the ChEMBL ontology; DrugBank means that the dictionary used comes from the DrugBank ontology

In the early experiments, which did not include dictionaries, we can see how BANNER provides, regardless of the scheme used, very similar results apparently without a significant statistical difference. However, the IO scheme seems to provide a slight improvement, moderately increasing the number of TP and decreasing the number of FP and FN.

Next, we train the models using the different dictionaries with the IO tagging scheme. As we can see in the table, the changes also do not show a significant variation, contributing a slight increase of the F1 when we used the combination of the dictionary of DINTO with the one of ChEMBL. Taking into account that this last configuration was done only with the IO tagging scheme, we proceed to check the results with the other schemas, giving as shown in the Table 1, results slightly lower than the initial configuration.

Based on the observed results, we decided to select the following configurations for the runs whose outputs were submited to the CEMP task:

- Run1: IO schema with DINTO and ChEMBL dictionaries.
- Run2: IO schema with ChEMBL and DrugBank dictionaries.
- Run3: IO schema with only DINTO dictionary.

Table 2 shows the results obtained in test dataset by the runs mentioned above. Our best run has achieved a precision of 88.42%, a recall of 82.64% and an F1 of 85.44%. As can be seen, in general the results achieved are very close between them, but we can also see how the BANNER tool is able to provide good results in the task of recognition of chemical entities, as well as the increase of the results achieved due to the use of two complementary dictionaries.

	IOB shema	dictionaries	P	R	F1
Run 1	IO	DINTO+ChEMBL	0.8842	0.8264	0.8544
Run 2	IO	ChEMBL+DrugBank	0.8831	0.8251	0.8531
Run 3	IO	DINTO	0.8799	0.8240	0.8510

Table 2. CEMP results on the test dataset. DINTO means that the dictionary used comes from the DINTO ontology; ChEMBL means that the dictionary used comes from the ChEMBL ontology; DrugBank means that the dictionary used comes from the DrugBank ontology

Acknowledgments. This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (eGovernAbility-Access project TIN2014-52665-C2-2-R).

References

1. Robert Leaman, Graciela Gonzalez: BANNER: An Executable Survey of Advances in Biomedical Named Entity Recognition. Pacific Symposium on Biocomputing 2008: 652-663
2. María Herrero-Zazo, Isabel Segura-Bedmar, Janna Hastings, Paloma Martínez: DINTO: Using OWL Ontologies and SWRL Rules to Infer Drug-Drug Interactions and Their Mechanisms. Journal of Chemical Information and Modeling 55(8): 1698-1707 (2015)
3. Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, Alfonso Valencia: CHEMDNER: The drugs and chemical names extraction challenge. J. Cheminformatics 7(S-1): S1 (2015)
4. Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez. 2015. Combining conditional random fields and word embeddings for the chemdner-patents task. In Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain. pages 9093.
5. Martin Prez Prez, Obdulia Rabal, Gael Prez Rodrguez, Miguel Vazquez, Florentino Fdez-Riverola, Julen Oyarzabal, Alfonso Valencia, Analia Lourenco and Martin Krallinger. Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks., p. 3-11