IBEnt: Chemical Entity Mentions in Patents using ChEBI

Andre Lamurias*, Luis F. Campos, and Francisco M. Couto

LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract. This article presents our approach to the CEMP task of BioCreative V.5, which consisted in using our system, IBEnt, to identify chemical entity mentions in patents through machine learning and semantic similarity techniques. The features used combine the results of a CRF classifier, two lexical matching methods (FiGO and MER) and semantic similarity measures on ChEBI ontology. We also tested the usage of MER by itself, without the machine learning approach. Combining these techniques, we submitted 5 runs for evaluation. We obtained better results using the machine learning approach with lexical and semantic similarity features. The best F-score obtained was 0.8541, while the MER system obtained 0.5967.

Key words: Named-Entity Recognition, Machine Learning, Semantic Similarity, Conditional Random Fields

1 Introduction

This paper presents our approach to the BioCreative V.5 CEMP task (Chemical Entity Mention in Patents) [13]. The objective of this task was to develop a system for detecting chemical entities in patent documents. A gold standard of 21,000 patents with chemical annotations was provided to the participants. For each chemical entity mentioned in a document, the start and end offsets were provided, along with the original text. We divided the gold standard into two partitions of equal size, which we refer to as training and development sets. The test set provided to the participants consisted of 9,000 patents. The participating systems were evaluated by the quality of their annotations in this test set.

Our approach used IBEnt [9], a framework to identify biomedical entities based on machine learning and semantic similarity techniques. We trained one classifier using Conditional Random Fields (CRF) and combined the results of that classifier with semantic techniques and a lexicon-based system to train a Random Forests classifier. The code used to generate our results is available at our GitHub repository[1]. The remainder of this article describes the features and resources used for this task, presents our results and discusses the performance of each approach.

* Corresponding author: alamurias@lasige.di.fc.ul.pt

[1] `https://github.com/lasigeBioTM/IBEnt`

2 Systems description and methods

We trained a CRF classifier with CRFsuite [11] on the training set annotations. The features used consisted in the linguistic, orthographic, morphological and contextual properties of the tokens, as well as domain-specific features. For most features, we considered a contextual window of size one, i.e., the value of the same feature for the previous and next token. Lemma and Part-of-Speech tags were obtained using Stanford CoreNLP [10]. Furthermore, we used three domain-specific features that have been also been used previously in similar tasks. These domain features checked if the token had a greek letter, a dash, or a periodic table element. A more detailed description of these features can be found in [9].

In the last decades the biomedical community has been developing and using ontologies to represent entities [2]. For example, in the case of chemical compounds we have Chemical Entities of Biological Interest (ChEBI) [6]. Each entity identified with the CRF classifier was matched to ChEBI, using a lexical similarity method, FiGO [4]. This method assigns a confidence score to each mapping, based on the information content of each word of the expression. Words that are more common have lower information content, contributing to lower confidence scores, while more informative words contribute to higher confidence scores. We refer to this confidence score as FiGO score.

We then computed the semantic similarity between every entity in the same sentence. Our assumption is that entities mentioned in a limited text window are more similar than entities mentioned across larger text windows. This assumption can be used to filter false positives made by the CRF classifier and FiGO[1, 9]. For example, if "2,3-bisphosphoglyceric acid", "cyclic 2,3-bisphospho-D-glyceric acid" and "2,3-bisphosphoglycerate" were recognized in the same sentence, and assuming that the first two are semantically related, and the latter entity has a low semantic similarity to the first two, then we would have less confidence on the latter entity being correctly recognized. The semantic similarity score of an entity consisted in the maximum similarity to other entities identified in the same sentence.

We used five semantic similarity measures (SSM) (Resnik [15], simUI [5], simGIC [14], h-simUI, h-simGIC [9]), therefore obtaining five semantic similarity scores for each entity.

We then used the CRF, lexical and semantic similarity scores as features for a Random Forests classifier. The Random Forests implementation used was from scikit-learn [12]. The objective of this classifier was to exclude false positives from the CRF results. As such, the training data consisted of one instance for each entity identified by the CRF classifier. Since the CRF classifier was trained on the training set, we trained the Random Forests classifier on the development set.

As a comparison to the machine learning approach, we also used a lexicon-based system - MER [3]. We constructed four lexicons using chemical entities

datasets freely available online (ChEBI[2], ChEMBL[3], DrugBank[4] and HMDB[5]). Using the training data, we tested which combination of these would give rise to the best performance. We found out that using a lexicon consisting of the terms included in ChEBI, ChEMBL and HMDB achieved the highest F-Score. One of the runs consisted in the results of using MER with this lexicon. We also constructed a lexicon consisting of all the terms annotated in the training set. We knew beforehand that using MER would return low performance scores, but we though that would be interesting to study how a fast and simple lexicon-based system as MER would compare with more complex systems that use machine learning.

Figure 1 provides an overview of our system. We obtained scores from CRF, FiGO and semantic similarity measures. These scores were used to train a Random Forests classifier. We also incorporated three features based on the three lexicons used with MER. An additional feature was added to each entity for each lexicon used, which had the value 1 if that entity was found on that lexicon, and 0 otherwise. After training the Random Forests classifier, we applied to the test set documents the same process that was applied to development set.

2.1 Runs

We combined the techniques previously described into 5 results submissions (runs) (Table 1). Our intention was to test a machine learning system (IBEnt), a rule-based system (MER) and a combination of both. Therefore, run 1 consisted in using IBEnt with the Random Forests classifier previously described. On run 2, we added features to this classifier based on the MER results. Run 3 consisted in using MER with a lexicon composed by the terms from ChEBI, ChEMBL and HDMB, while run 4 used MER with the terms found on the training set. Finally, run 5 combined the lexicons used in run 3 and 4 with a lexicons composed by the terms found on the test set by IBEnt (run 1). While run 2 represents how the results of a rule-based system can be used in the context of machine learning, the idea of run 5 was to show that machine learning can be used to generate lexicons that can then be used by more efficient lexicon-based systems.

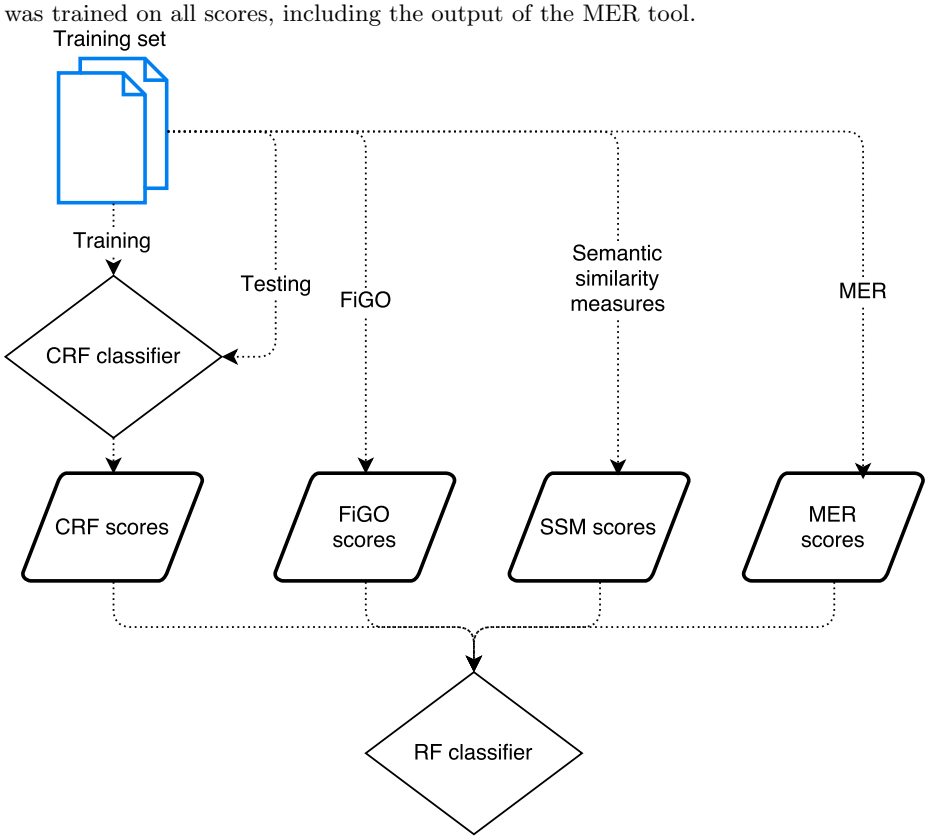
3 Results and Discussion

After processing the documents of the test set, we submitted the results of each run to the BeCalm platform. The precision, recall and F-score scores obtained are shown in Table 2.

The highest F-score obtained was with run 1, which used CRF, FiGO and semantic similarity features. Adding features based on lexicon matching (run 2) did not improve recall nor precision.

[2] `https://www.ebi.ac.uk/chebi/`
[3] `https://www.ebi.ac.uk/chembl/`
[4] `https://www.drugbank.ca/`
[5] `http://www.hmdb.ca/`

Fig. 1. Pipeline of the system used for our CEMP submission. Half of the training data was used to train a classifier and the other half was used to calculate FiGO and semantic similarity scores, independently. This Random Forests classifier was trained on all scores, including the output of the MER tool.



When comparing the runs that use a lexicon-based system (3, 4 and 5), the best F-score was 0.5967 (run 4). This run used only the entities found in the training data as lexicon. This lexicon performed better than combining the ChEBI, ChEMBL and HDMB vocabularies (run 3).

The script used to generate run 5 had an error that eliminated the entities detected on multiple lexicons. The minimum expected recall would be the same recall as run 4 (0.5747), since run 5 includes the lexicon used on run 4. However, when merging lexicons, repeated entities were accidentally excluded. Since we did not have access to the test set annotations, it is not possible to test again with the bugfix.

We measured the time necessary to process the test set using the Random Forests approach compared to the lexicon-based approach. While it took an average of 5.19 seconds to process each document using Random Forests, the

Table 1. Summary of techniques and resources used on each run

Run	Approach	Features	Lexicon
1	Random Forests	CRF, FiGO, SSM	-
2	Random Forests	CRF, FiGO, SSM, MER	ChEBI, ChEMBL, HDMB
3	MER	-	ChEBI, ChEMBL, HDMB
4	MER	-	Training set
5	MER	-	ChEBI, ChEMBL, HDMB, Training set, Run 1 output

Table 2. Results obtained on the test set during the competition.

Run	Precision	Recall	F-score
1	**0.8732**	**0.8338**	**0.8531**
2	0.6705	0.7452	0.7059
3	0.5367	0.3794	0.4445
4	0.6205	0.5747	0.5967
5	0.5476	0.3042	0.3911

lexicon-based approach took 0.37 seconds per document using the same hardware. This difference in processing time was the reason we tested this approach on some runs. Although we were not able to obtain F-scores as high as with machine learning, the lexicon-based approach was able to process the documents much faster.

On the previous edition of this task [7], the highest F-score obtained was 0.8937, which is 0.0406 higher than our best F-score on this edition. The highest precision was 0.8971, which is closer to the precision we obtained this year.

4 Conclusion

We present our open-source system, IBEnt, that participated in the CEMP task of BioCreative V.5. IBEnt is mainly based on machine learning and semantic similarity techniques. Semantic similarity is calculated using the ChEBI ontology. This system obtained an F-score of 0.8531, using CRF, FiGO and semantic similarity features. Furthermore, we combined this approach with MER, a lexicon-based system, to study how these two approaches can be used together.

Using a lexicon-based system, we obtained a best F-score of 0.5967. This type of approach can be used in cases where the response time is the priority, instead of the quality of the results. However the quality of the results obtained with this approach may be improved by incorporating a more comprehensive lexicon, by adding abbreviations, synonyms and other types or chemical descriptors. To accomplish this, we will have to carefully analyze all the descriptors of chemical compounds found by IBEnt and not found by MER.

In the future we could improve our results by testing different proportions between the training and development set. We divided the gold standard in two sets of the same size, one to train a CRF classifier and the other one to train a Random Forests classifier based on results from the CRF classifier. We may also apply our distant supervision methods to improve the classifiers[8]. However, a different partition of the gold standard could lead to better results, for example, using 70% of the documents to train CRF and the rest to train Random Forests.

Acknowledgments. This work was supported by FCT through funding of the PhD Grant ref. PD/BD/106083/2015 and of the LaSIGE Research Unit, ref. UID/CEC/00408/2013.

References

1. Alhelbawy, A., Gaizauskas, R.: Graph Ranking for Collective Named Entity Disambiguation. ACL pp. 75–80 (2014)
2. Barros, M., Couto, F., et al.: Knowledge representation and management: a linked data perspective. IMIA Yearbook pp. 178–183 (2016)
3. Couto, F.M., Campos, L., Lamurias, A.: MER: a minimal named-entity recognition tagger and annotation server. Proceedings of the BioCreative V.5 Challenge Evaluation Workshop (2017)
4. Couto, F.M., Silva, M.J., Coutinho, P.M.: Finding genomic ontology terms in text using evidence content. BMC bioinformatics 6 Suppl 1, S21 (2005)
5. Gentleman, R.: Visualizing and Distances Using GO (2005)
6. Hastings, J., De Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C.: The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. Nucleic Acids Research 41(D1), 456–463 (2013)
7. Krallinger, M., Rabal, O., Lourenço, A., Perez, M.P., Rodriguez, G.P., Vazquez, M., Leitner, F., Oyarzabal, J., Valencia, A.: Overview of the CHEMDNER patents task. Proceedings of the Fifth BioCreative Challenge Evaluation Workshop (Mcc), 63–75 (2015)
8. Lamurias, A., Clarke, L.A., Couto, F.M.: Extracting microRNA-gene relations from biomedical literature using distant supervision. PloS one 12(3), e0171929 (2017)
9. Lamurias, A., Ferreira, D.J.D., Couto, F.M.: Improving chemical entity recognition through h-index based semantic similarity. Journal of Cheminformatics 7(Suppl 1), S13 (2015)
10. Manning, C.D., Bauer, J., Finkel, J., Bethard, S.J., Surdeanu, M., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations pp. 55–60 (2014)

11. Okazaki, N.: CRFsuite: a fast implementation of conditional random fields (crfs) (2007), http://www.chokkan.org/software/crfsuite/
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
13. Perez, M.P., Rabal, O., Rodriguez, G.P., Vazquez, M., Fdez-Riverola, F., Oyarzabal, J., Valencia, A., Lourenco, A., Krallinger, M.: Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks. Proceedings of the BioCreative V.5 Challenge Evaluation Workshop pp. 3–11 (2017)
14. Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcão, A., Couto, F.: Metrics for GO based protein semantic similarity: a systematic evaluation. BMC bioinformatics 9(5), S4 (2008)
15. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of artificial intelligence research (1999)