LASIGE: using Conditional Random Fields and ChEBI ontology

Tiago Grego
Dep. de Informática
Faculdade de Ciências
Universidade de Lisboa
Portugal
tgrego@fc.ul.pt

Francisco Pinto
Dep. de Química e Bioquímica
Faculdade de Ciências
Universidade de Lisboa
Portugal
frpinto@fc.ul.pt

Francisco M. Couto
Dep. de Informática
Faculdade de Ciências
Universidade de Lisboa
Portugal
fcouto@di.fc.ul.pt

Abstract

For participating in the SemEval 2013 challenge of recognition and classification of drug names, we adapted our chemical entity recognition approach consisting in Conditional Random Fields for recognizing chemical terms and lexical similarity for entity resolution to the ChEBI ontology. We obtained promising results, with a best F-measure of 0.81 for the partial matching task when using post-processing. Using only Conditional Random Fields the results are slightly lower, achieving still a good result in terms of F-measure. Using the ChEBI ontology allowed a significant improvement in precision (best precision of 0.93 in partial matching task), which indicates that taking advantage of an ontology can be extremely useful for enhancing chemical entity recognition.

1 Introduction

Most chemical named entity recognition systems can be classified in two approaches: dictionary based and machine learning based approaches. Dictionary based approaches are usually easier to implement and maintain, but require a reference chemical term dictionary and are dependent on its completeness and quality. The availability of public chemical databases has been an issue until recently, when several publicly available databases such as PubChem (Wang et al., 2009), DrugBank (Wishart et al., 2006) and ChEBI (Degtyarenko et al., 2007) were released. An example of a popular system that uses this approach is Whatizit (Rebholz-Schuhmann et al., 2008). Machine learning based approaches are not limited to a terminology and are thus better suited for finding novel chemical terms that are yet to be inserted in reference databases. However this approach requires training data for a classifier to be able to successfully learn and perform the chemical entity recognition task. Some methods combine both approaches and thus are hybrid systems that aim to take the best out of both approaches (Jessop et al., 2011; Rocktäschel et al., 2012).

An annotated corpus of patent documents was released by ChEBI, and using such corpus as training data we developed an chemical entity recognition system (Grego et al., 2009) that uses a machine learning approach based on Conditional Random Fields (CRF) (Lafferty et al., 2001). We furthermore expanded our method to allow resolution of recognized entities to the ChEBI ontology (Grego et al., 2012).

This paper describes how our system (Grego et al., 2012) was adapted to perform the task of recognition and classification of drug names, and presents the results obtained in the task 9.1 of the 7th International Workshop on Semantic Evaluation (SemEval 2013).

2 Task and Dataset

The Task 9 of SemEval 2013 involved two sub-tasks: (9.1) recognition and classification of drug names, and (9.2) extraction of drug-drug interactions from Biomedical Texts (SemEval, 2013). The recognition and classification of drug names (Task 9.1) comprises two steps. First is chemical named entity recognition, that consists in finding in a sentence the offsets for the start and end of a chemical entity.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 660–666, Atlanta, Georgia, June 14-15, 2013.

An exact match is achieved by correctly identifying both the start and end offset, as curators manually provided them. If there is a mismatch in the offsets but there is some overlap with a manual annotation, then it is considered a partial match, otherwise it is a recognition error.

The second step consists in classifying each recognized entity in one of four possible entity types: i) Drug is any pharmaceutical product approved for human use; ii) Brand is a drug that was first developed by a pharmaceutical company; iii) Group refers to a class or group of drugs; iv) Drug_n is an active substance that has not been approved for human use. Thus, the evaluation takes into account not only entity recognition, but also the assigned type. Type matching assessment considers the entity type evaluation from partial matching entity recognition, while strict matching considers the entity type evaluation from exact matching.

For training, the DDI corpus dataset was provided (Segura-Bedmar et al., 2006). This dataset contains two sub-datasets. One that consists of MedLine abstracts, and other that contains DrugBank abstracts. An unannotated test dataset was provided for testing and evaluating the systems.

3 CRF entity recognition

Our method uses CRFs for building probabilistic models based on training datasets. We used the MALLET (McCallum, 2002) implementation of CRFs. MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text, which includes an implementation of linear chain CRFs.

A required first step in our method in the tokenization of the input text. For this task we have used a specifically adapted tokenizer for chemical text adapted from an open source project (Corbett et al., 2007).

Each token is then represented as a set of features. We kept using a set of features derived in our previous work (Grego et al., 2009), which includes for each token:

Stem: The stem of the token.

Prefix: The first three characters of the token.

Suffix: The last three characters of the token.

Number: Boolean that indicates if the token contains digits.

In addition to the set of features, each token is also given a label in accordance to the training data:

NO: A non-chemical token.

NE: A chemical entity represented by a single token.

S-NE: The first token of a multi-token chemical entity.

M-NE: A middle token of a multi-token chemical entity (only exists for entities composed by three or more tokens).

E-NE: The last token of a multi-token chemical entity.

The task of entity recognition will be the assignment of such labels to new, unannotated text, based on a model. The assigned label allows for named entities to be recognized and offsets provided.

For creating a model, it is required as input a set of annotated documents. Our method was initially developed using an annotated patent document corpus released to the public by the ChEBI team. This corpus can be found at [1], and we decided to keep using it as training data for a model. Together with this corpus, the DDI corpus training dataset provided for the task was used. The model produced by using this combination of training data, that we called All model, will be suited for general purpose chemical entity recognition.

We then prepared four datasets based on the DDI corpus dataset but containing only one type of annotated entities each. With that training data we prepared four more models, each trained only with one kind on entity type. Thus we have in total prepared five models:

All: A model trained with all entity types of the DDI corpus dataset, and the ChEBI released patent dataset.

[1] `http://chebi.cvs.sourceforge.net/viewvc/chebi/chapati/patentsGoldStandard/`

Drug: A model trained only with the entities of type drug in the DDI corpus dataset.

Brand: A model trained only with the entities of type brand in the DDI corpus dataset.

Group: A model trained only with the entities of type group in the DDI corpus dataset.

Drug_n: A model trained only with the entities of type drug_n in the DDI corpus dataset.

Using the type specific models it is possible to annotate text with only one entity type. Thus our method now has the capability of entity type classification in addition to named entity recognition, using these type specific models.

4 ChEBI resolution

After having recognized the named chemical entities, our method tries to perform their resolution to the ChEBI ontology. ChEBI (Chemical Entities of Biological Interest) is a freely available dictionary of small molecular entities. In addition to molecular entities, ChEBI contains groups (parts of molecular entities) and classes of entities, allowing for an ontological classification that specifies the relationships between molecular entities or classes of entities and their parents and/or children. The ontology structure provides an integrated overview of the relationships between chemical entities, both structural and functional.

The resolution method takes as input the string identified as being a chemical compound name and returns the most relevant ChEBI identifier along with a confidence score.

To perform the search for the most likely ChEBI term for a given entity an adaptation of FiGO, a lexical similarity method (Couto et al., 2005). Our adaptation compares the constituent words in the input string with the constituent words of each ChEBI term, to which different weights have been assigned according to its frequency in the ontology vocabulary (Grego et al., 2012). A resolution score between 0 and 1 is provided with the mapping, which corresponds to a maximum value in the case of a ChEBI term that has the exact name as the input string, and is lower otherwise.

5 Post-processing

To further improve the quality of the annotations provided by our method, some naïve rules were created and external resources used.

One of the rules implemented is derived from the resolution process, and corresponds in classifying an entity as type Group if its ChEBI name is plural. This is because ChEBI follows the convention of naming its terms always as a singular name, except for terms that represent classes of entities where a plural name can be used.

We have also used other resources in the post-processing besides ChEBI, namely a list of brand names extracted from DrugBank. This list of brand names was used to check if a given entity was part of that list, and if it was the entity should be of the type Brand.

A common English words list was also used as external resource in post-processing. If a recognized chemical entity was part of this list then it was a recognition error and should be filtered out and not be considered a chemical entity.

Some simple rules were also implemented in an effort to improve the quality of the annotations. For instance, if the recognized entity was found to be composed entirely by digits, then it should be filtered out because it is most certainly an annotation error. Also, if an entity starts or ends with a character such as "$\star$", "$-$", "$.$", "$,$" or "$'$", then those characters should be removed from the entity and the offsets corrected accordingly.

With such naïve but efficient rules it was expected that the performance of entity recognition would improve. An overview of the system architecture is provided in Figure 1.

6 Testing runs

Using different combinations of the described methods, three runs were submitted for evaluation and are now described.

Run 1: This run uses all of the described methods. Entity recognition is performed using all models, and the type classification is performed by using the type specific models in the following priority: if an entity was recognized using the Drug_n model, then type is Drug_n, else if it

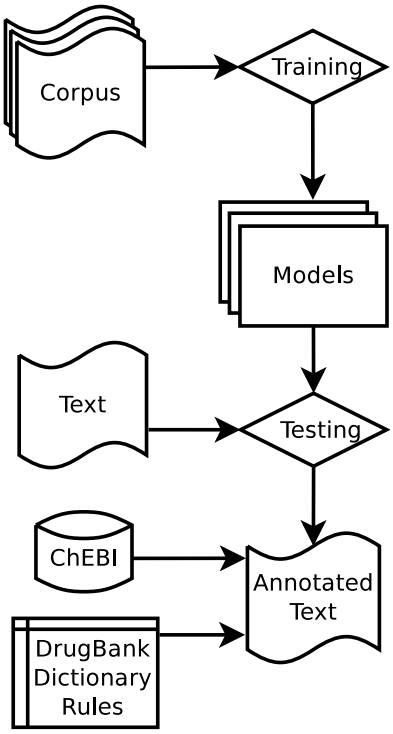

Figure 1: Overview of the system architecture. Based on annotated corpus, CRF models are created and used to annotate new documents.

was recognized using the Brand model, then type is Brand, else if it was recognized using the Group model, then type is Group, else and finally it is assigned the type Drug. Resolution to ChEBI is performed and all of the described post-processing rules applied.

Run 2: In this run only the classifiers are used. This means that the entity recognition is performed using all models, and the type classification is performed by using the type specific models as described in Run 1. However no extra processing is performed and the results are submitted as obtained directly from the classifiers.

Run 3: This run performs entity recognition in a similar way described in run 1, and performs entity recognition to the ChEBI ontology. However, only the entities successfully mapped to ChEBI, with a resolution score of at least 0.8, are considered. All the other entities are discarded in this phase. After resolution and the filtering of entities according to the resolution to ChEBI, all the described post-processing rules are applied in a similar way to Run 1.

7 Results and Discussion

The official evaluation results are presented in Table 1. We can observe that the obtained results are better for the DrugBank dataset than for the MedLine dataset. This may have happened because the DrugBank dataset is four times larger than the MedLine dataset, but also because while the DrugBank abstracts are quite focused in drug descriptions and use mostly systematic names, the MedLine ones are usually more generic and make more extensive use of trivial drug names. We obtained for the Run 1 a top F-measure of 0.81 in the full dataset for a partial matching assessment, and that value decreased to 0.78 when an exact matching assessment is considered. The values are very close, which means that our method is being able to efficiently find the correct offsets of the entities. However the F-measure decreases to 0.69 for partial matching and 0.66 for exact matching when the assignment of the entity type is considered. This means that there is room to improve in the task of classifying the chemical entities to the correct entity type.

Run 2 obtained results very similar to Run 1, only slightly less F-measure. The difference between those two runs was that Run 2 used only the classifiers, while Run 1 used rules and external resources in an effort to improve the results. We can thus conclude that the classifiers alone produce already good results and more sophisticated post-processing is required to obtain significant performance gains. Our post-processing was very simple as explained earlier, and can only slightly improve the results obtained with the CRF classifiers alone.

Run 3 obtained improved precision in all assessments. In this run only the entities that were successfully mapped to ChEBI were considered, and thus the precision of recognition was the best of our runs. This is because ChEBI contains high quality, manually validated chemical terms. If a recognized entity can be successfully mapped to this data source, then there is a good indication that it is, in fact, a valid chemical entity. However F-measure has decreased because of a loss in recall. ChEBI is still a young project containing slightly over 30,000 chem-

Assessment	Run	MedLine Dataset			DrugBank Dataset			Full Dataset		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Strict matching	1	0.6	0.54	**0.57**	0.82	0.72	**0.77**	0.7	0.62	**0.66**
	2	0.54	0.54	0.54	0.82	0.73	**0.77**	0.65	0.62	0.64
	3	**0.66**	0.48	0.56	**0.83**	0.58	0.68	**0.73**	0.52	0.61
Exact matching	1	0.78	0.7	**0.74**	0.89	0.78	**0.83**	0.83	0.74	**0.78**
	2	0.73	0.74	0.73	0.88	0.78	**0.83**	0.79	0.76	0.77
	3	**0.82**	0.6	0.69	**0.91**	0.63	0.74	**0.86**	0.61	0.72
Partial matching	1	0.81	0.73	**0.77**	0.91	0.8	**0.85**	0.86	0.76	**0.81**
	2	0.76	0.77	0.76	0.91	0.8	**0.85**	0.82	0.78	0.8
	3	**0.86**	0.63	0.72	**0.93**	0.65	0.76	**0.89**	0.64	0.74
Type matching	1	0.64	0.58	**0.61**	0.85	0.75	**0.8**	0.73	0.65	**0.69**
	2	0.57	0.58	0.58	0.85	0.75	**0.8**	0.69	0.66	0.67
	3	**0.71**	0.52	0.6	**0.87**	0.61	0.71	**0.78**	0.56	0.65

Table 1: Results obtained in Task 9.1 for the different assessments. Exact and Partial matching do not consider the entity type, while Strict and Type matching consider the entity type for Exact and Partial matching entity recognition respectively.

Entity Type	Run	MedLine Dataset			DrugBank Dataset			Full Dataset		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Drug	1	0.58	0.82	0.68	0.85	0.78	0.82	0.69	0.8	0.74
	2	0.51	0.82	0.63	0.83	0.81	0.82	0.64	0.82	0.72
	3	0.66	0.74	0.7	0.88	0.67	0.76	0.75	0.7	0.73
Brand	1	1	0.5	0.67	0.77	0.45	0.57	0.79	0.46	0.58
	2	0.67	0.33	0.44	0.91	0.4	0.55	0.88	0.39	0.54
	3	1	0.5	0.67	0.65	0.21	0.31	0.7	0.24	0.35
Group	1	0.7	0.54	0.61	0.82	0.85	0.83	0.76	0.67	0.71
	2	0.64	0.56	0.6	0.82	0.83	0.82	0.72	0.67	0.7
	3	0.7	0.47	0.56	0.83	0.69	0.76	0.76	0.56	0.65
Drug_n	1	0.48	0.11	0.18	0	0	0	0.42	0.11	0.17
	2	0.5	0.12	0.2	0	0	0	0.42	0.12	0.18
	3	0.48	0.1	0.17	0	0	0	0.41	0.1	0.16

Table 2: Results obtained in Task 9.1 for each entity type. In this evaluation only the entities of a specific type are considered at a time.

Run	MedLine Dataset			DrugBank Dataset			Full Dataset		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
1	0.69	0.50	0.58	0.61	0.52	0.56	0.67	0.51	0.58
2	0.58	0.46	0.51	0.64	0.51	0.57	0.67	0.50	0.57
3	0.71	0.45	0.55	0.59	0.39	0.47	0.66	0.4	0.5

Table 3: Macro-average measures obtained for each run.

ical entities, which is still a low amount of entities when compared with other chemical databases (for example, PubChem contains more than 10 times that amount). However ChEBI is growing at a steady pace and we believe its coverage will keep increasing while maintaining the high quality that allows for an excellent precision. Thus, as ChEBI evolves, our approach will mantain the high levels of precision but with a lower reduction in recall.

ChEBI is not only a chemical dictionary, but an ontology. This allows for a comparison recognized entities through semantic similarity measures that can be used to further enhance chemical entity recognition (Ferreira and Couto , 2010; Couto and Silva , 2011). This comparison can also be extremely useful in other task such as drug-drug interaction extraction. Moreover, even if with a relatively small ChEBI, it can be possible to increase coverage by integrating other available resources using Ontology Matching techniques (Faria et al., 2012).

In Table 2 we have the official results obtained for each entity type, and we can observe that our method is efficient in correctly classifying the Drug and Group types, where it achieves an F-measure of 0.74 and 0.71 correspondingly. However our method has some difficulties in correctly classifying entities of the Brand type, where an F-measure of 0.58 was obtained. The Drug_n entity type has proven to be a very challenging type to be correctly classified, and our system failed the correct classification of this type in most situations. This is possibly because the percentage of entities of this type is very limited, and also because the difference between this type and the Drug type is the fact that the later has been approved for human use, while the former has not. The feature set used cannot efficiently discriminate this information and external information about drug approval for human usage must be used for efficient detection of this type.

Overall, Run 1 has obtained the best results. However, the results from Run 2 have been very similar, which shows that the classifiers have been successful and the post-processing of Run 1 has been minimal. Run 3 was designed for high precision, because only the entities successfully mapped to the ChEBI ontology were considered. It does improve the obtained precision, but suffers a drop in recall. Table 3 presents the macro-average measures obtained for each run.

8 Conclusions

This paper presents our participation in the 7th International Workshop on Semantic Evaluation (SemEval 2013) using a CRF-based chemical entity recognition method and a lexical similarity based resolution method. We prepared type-specific CRF models to allow both recognition and type classification of the chemical entities. Mapping of the entities to the ChEBI ontology was performed using a lexical similarity based method, and several post-processing rules using external resources where implemented.

We submitted different runs on annotated test data using different combination of such methods, and obtained a best precision of 0.89 and a best F-measure of 0.81 in the entity recognition task. For the task of entity recognitions and classification we have obtained a best precision of 0.78 and a best F-measure of 0.69. We concluded that the classifiers provide already good results by their own, that can be slightly improved by using some naïve external resources and rules.

However, using ChEBI allows for a significant increase of precision, which is encouraging. We believe this result is a good indication that as ChEBI matures, the methods that take advantage of its ontology structure for entity recognition and classification will benefit more from its usage, increasing the F-measure obtained in the task.

9 Acknowledgments

The authors want to thank the Portuguese Fundação para a Ciência e Tecnologia through the financial support of the SPNet project (PTDC/EBB-EBI/113824/2009), the SOMER project (PTDC/EIA-EIA/119119/2010) and the PhD grant SFRH/BD/36015/2007 and through the LASIGE multi-annual support. The authors also wish to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807).

References

P. Corbett, C. Batchelor and S. Teufel. 2007. Annotation of chemical named entities. *Proceedings of the Work-*

shop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, 57–64.

F. M. Couto and M. J. Silva. 2011. Disjunctive shared information between ontology concepts: application to Gene Ontology. *Journal of Biomedical Semantics*, 2(5).

F. M. Couto, P. M. Coutinho and M. J. Silva. 2005. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6 (Suppl 1), S21.

K. Degtyarenko, P. Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj and M. Ashburner. 2007. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36, D344.

D. Faria, C. Pesquita, E. Santos, F. M. Couto, C. Stroe and I. F. Cruz 2012. Testing the AgreementMaker System in the Anatomy Task of OAEI 2012. *CoRR*, abs/1212.1625, `arXiv:1212.1625`.

J. D. Ferreira and F. M. Couto 2010. Semantic similarity for automatic classification of chemical compounds. *PLoS Computational Biology*, 6(9).

T. Grego, C. Pesquita, H. P. Bastos and F. M. Couto. 2012. Chemical Entity Recognition and Resolution to ChEBI. *ISRN Bioinformatics*, Article ID 619427.

T. Grego, P. Pezik, F. M. Couto and D. Rebholz-Schuhmann. 2009. Identification of Chemical Entities in Patent Documents. *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, volume 5518 of *Lecture Notes in Computer Science*, 934–941.

T. Grego, F. Pinto and F. M. Couto. 2012. Identifying Chemical Entities based on ChEBI. *Software Demonstration at the International Conference on Biomedical Ontologies (ICBO)*.

D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy and P. Murray-Rust 2011. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3(41).

J. Lafferty, A. McCallum and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*, 282–289.

A. K. McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch and A. Jimeno. 2008. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24(2):296–298.

T. Rocktäschel, M. Weidlich and U. Leser. 2012. ChemSpot: A Hybrid System for Chemical Named Entity Recognition. *Bioinformatics*, 28(12): 1633-1640.

I. Segura-Bedmar, P. Martínez and C. de Pablo-Sánchez 2006. Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, 44(5): 789–804.

Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant. 2009. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37, W623.

D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34, D668.

SemEval 2013. In Proceedings of the 7th International Workshop on Semantic Evaluation