Neji: Recognition of Chemical and Gene Mentions in Patent Texts

André Santos and Sérgio Matos*

DETI/IEETA, University of Aveiro,
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
{andre.jeronimo,aleixomatos}@ua.pt
http://bioinformatics.ua.pt/

Abstract. The BioCreative V.5 challenge focused on the recognition of chemicals and gene mentions in medicinal chemistry patents. For participation in the chemical entity (CEMP) and gene and protein (GPRO) recognition tasks, we used the concept recognition framework Neji and applied a machine-learning strategy using a optimized feature set. Our best submissions achieved an F-score of 86.6% for the identification of chemicals and 71.3% for the identification of gene names.

Key words: Text mining, Patents, Named entity recognition, Chemicals, Genes

1 Introduction

The BioCreative V.5 text mining challenge[1] focused on the development and evaluation of information extraction systems for recognition of chemical and gene entity mentions in chemistry patents [1]. Two offline sub-tasks were considered: recognition of chemical entity mentions (CEMP) and recognition of of gene and protein related objects (GPRO). For each sub-task, systems should identify all mentions of entities of the corresponding types in free-text and return the start and end indices of the text span.

We followed a machine-learning approach, using conditional random fields (CRF) models. We took advantage of the provided training and development corpora for performing feature selection and for identifying white and black lists to use in post-processing steps.

2 System description

We used Neji[2], a flexible and extensible concept recognition framework specially optimized for biomedical text [2]. Neji's architecture is illustrated in Figure 1,

* Corresponding author
[1] http://www.becalm.eu/pages/biocreative
[2] Available from https://github.com/BMDSoftware/neji

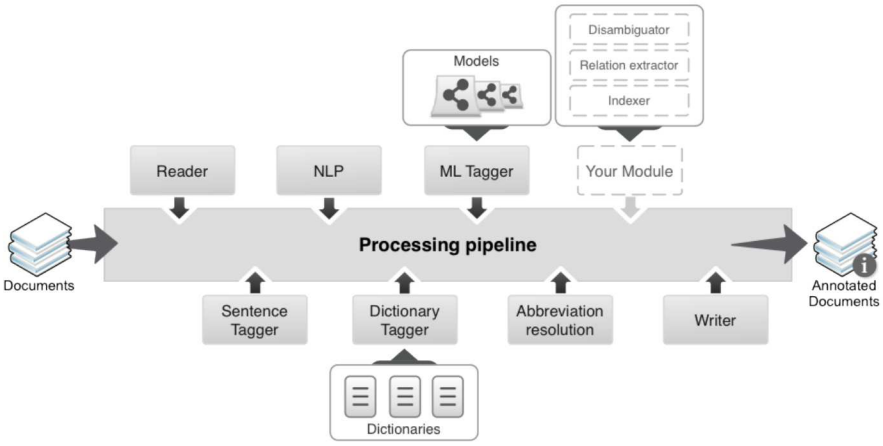

Fig. 1. Neji processing pipeline and architecture.

showing the various processing stages and input (Reader) and output (Writer) modules that provide support for a variety of formats, as well as allowing easy customization for new formats. The processing modules are managed through an efficient pipeline with multi-threading support. Neji includes natural language processing modules, based on GDep [14] and Apache OpenNLP [3], concept recognition modules based on dictionary matching and machine learning, and post-processing modules, including parentheses correction and abbreviation resolution. The machine learning component is based on Gimli [3], and makes use of the MALLET [12] implementation of Conditional Random Fields (CRFs) models [10]. This module provides simple methods for feature extraction and for training and applying CRF models for entity recognition.

We applied a machine learning (ML) approach, combined with dictionary-matching for obtaining lexicon features, as described in [4].

2.1 Corpora

The BioCreative V.5 corpus is composed of 21 thousand manually annotated patents (title and abstract), available for training the systems, and a further nine thousand patents used for evaluation. As with the previous BioCreative V CHEMDNER Patents task [8], chemical entities are annotated in seven classes: systematic, identifiers, formula, trivial abbreviation, family and multiple. We however grouped all mentions into a single class. Gene and protein related object (GPROs) annotations were divided into type 1, covering GPRO mentions that can be normalized to a database record; and type 2, including mentions that can not be normalized. We considered only identification of type 1 mentions.

[3] `https://opennlp.apache.org/`

The training set contains a total of 99634 chemical mentions and 17751 gene mentions (12422 of which of type 1).

In order to augment the available training data, we included the BioCreative IV CHEMDNER corpus, containing 10000 PubMed abstracts and a total of 84355 chemical entity mentions [9], which we used for training the recognition model for the CEMP sub-task, and the BioCreative II gene mention corpus, containing 20000 sentences from Pubmed abstracts with around 44500 gene mention annotations [16], which we used for training the GPRO task model.

Furthermore, to account for the expected differences between the patent and literature documents, as well as within corpus heterogeneity, we clustered the documents and created different recognition models for each cluster. For this, we applied bi-clustering, as provided in the scikit-learn machine learning library for python [13]. For the CEMP task, we combined 14 thousand patent documents from the BioCreative V.5 corpus, corresponding to the training and development sets of the BioCreative V task, with the 10 thousand documents from the BioCreative IV corpus, and obtained three clusters with 9032, 4622 and 10346 documents each. After the internal evaluation stage, we included the remaining 7000 documents from BioCreative V.5 and re-created the clusters, obtaining 10758, 6670 and 13572 for each cluster. For the GPRO task, we combined 14 thousand patents with the 20000 sentence from BioCreative II and generated clusters containing 18616, 4839 and 10545 documents. Similarly, after the internal evaluation phase we included the remaining documents and obtained clusters with 22165, 5771 and 13064 documents. Table 1 shows the distribution of patent and literature documents across the clusters.

Table 1. Composition of the clusters.

	CEMP				GPRO			
	test set removed		test set included		test set removed		test set included	
Cluster	patents	literature	patents	literature	patents	literature	patents	literature
1	1776	7256	5021	5737	6479	12137	9493	12672
2	4516	106	6540	130	2715	2124	4366	1405
3	7708	2638	9439	4133	4806	5739	7141	5923

2.2 Pre-processing

Sentence splitting was performed with Lingpipe[4]. Tokenization, lemmatization, part-of-speech (POS) tagging, chunking and dependency parsing were performed using Neji's custom version of GDep [3]. The BIO scheme was used for encoding the annotations.

[4] `http://alias-i.com/lingpipe`

2.3 Model and feature selection

We performed recursive feature elimination by training on the 7000 documents that compose the training set of BioCreative V, and testing on the development set. Together with this feature selection step, we tested CRF models with orders 1 and 2 and with forward (from left to right) and backward (from right to left) parsing directions.

2.4 Post-processing

Exclusion and inclusion lists were generated by analyzing the false-positive and false-negative mentions, respectively, obtained on the 7000 documents used for internal evaluation.

2.5 Ranking

To score and rank the annotations, we used the confidence scores provided by the CRF models, which is a value between 0 and 1 that reflects the certainty of the model generating each annotation.

3 Results and Discussion

Table 2 shows the feature sets that originated the best results for each task, based on cross-validation over the 14 thousand documents used for development. Interestingly, the results for the CEMP task improved with the inclusion of gene lexicon features, in addition to the chemical lexicon. The reverse was also true for the GPRO task. The best results were obtained using a first order CRF with backward parsing for the CEMP task, and a second order CRF with forward parsing for the GPRO task.

Tables 3 and 4 describe the submitted runs for participation in the CEMP and GPRO tasks, respectively, the internal evaluation results obtained on the BioCreative V CHEMDNER Patents test set, and official results on the BioCreative V.5 test set.

The results show that the post-processing stage improved considerably the results during internal evaluation, but this improvement was not replicated on the final test set. The use of distinct models trained on clustered documents originated slight improvements in recall in the CEMP task, and a large improvement also in recall in the GPRO task. These improvements were however balanced by significant reductions in precision, leading to worst overall results for CEMP and slightly better result for the GPRO task.

Following the challenge, we performed feature selection separately for each cluster of the CEMP task. For this, we divided the documents in each cluster in two groups and evaluated the impact of each feature by cross-testing with these two groups and taking the average f-score. With this strategy, we obtained three models with different feature sets which were then applied to the 7000

Table 2. Features used for training recognition models for each task.

Group	Feature	CEMP	GPRO
NLP	Token	x	
	Stem	x	
	Lemma	x	x
	POS	x	x
	Chunk tags	x	x
	Dependency parsing	x	x
Orthographic	Capitalization (e.g., "InitCap", "AllCaps")	x	x
	Digits and capitalized characters counting (e.g., "TwoDigit", "TwoCap")	x	x
	Symbols (e.g., "Dash", "Dot")		
	Greek letters (e.g., "α")	x	x
	Roman digits	x	
Morphological	Prefixes and suffixes	x	x
	Character n-grams		x
	Word shape features to reflect how letters, digits and symbols are organized in the token (e.g., the structure of "Abc:1234" is expressed as "Aaa#1111")	x	x
Lexicons	Chemical entity names from Jochem [7], ChEBI [6] and CTD [5]	x	x
	Gene names from BioThesaurus [11]	x	x
	Trigger words from BioLexicon [15]	x	x
Local context	Conjunctions of lemma and POS features of the windows {-1, 0}, {-2, -1}, {0, 1}, {-1, 1} and {-3, -1}	x	x

Table 3. Runs submitted to the CEMP sub-task.

		BC V test set			BC V.5 test set		
run	Description	P	R	F1	P	R	F1
1	1 model trained with 21k BioCreative V.5 training data	86.2	86.2	86.2	89.0	84.3	**86.6**
2	as above, plus FP filtering	**89.2**	86.2	**87.7**	**89.3**	84.0	86.6
3	clusters 1, 2, 3	83.4	79.6	81.4	84.6	**86.7**	85.6
4	BC V + BC IV	82.4	**89.2**	85.7	85.1	86.6	85.8
5	clusters 2, 3	81.9	88.7	85.2	85.7	85.8	85.8

Table 4. Runs submitted to the GPRO sub-task.

		BC V test set			BC V.5 test set		
run	Description	P	R	F1	P	R	F1
1	1 model trained with 21k BioCreative V.5 training data	71.2	62.9	66.8	**78.8**	62.0	69.4
2	as above, plus FN/FP filtering	**77.3**	**77.0**	**77.1**	77.4	61.9	68.9
3	clusters 1, 2, 3	47.2	73.3	57.4	68.8	65.8	67.3
4	BC V + BC II	34.1	76.6	47.2	61.6	71.0	66.0
5	clusters 1, 3	50.6	73.0	59.8	71.5	**71.1**	**71.3**

documents in the BioCreative V CHEMDNER Patents test set. Using intersection, we achieved an f-score of 86.1, a precision of 77.0 and a recall of 97.8. This was improved to an f-score of 86.9, precision of 78.8 and recall of 96.9 by using false-positives filtering as in the submitted run 2.

Acknowledgments This work was supported by Portuguese National Funds through FCT - Foundation for Science and Technology, in the context of the project IF/01694/2013. Sérgio Matos is funded under the FCT Investigator programme.

References

1. Pérez-Pérez, M., Rabal, O., Pérez-Rodríguez, G., et al.: 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., p. 3-11 (2017)
2. Campos, D., Matos, S., Oliveira, J.L.: A modular framework for biomedical concept recognition. BMC bioinformatics 14(281) (2013)
3. Campos, D., Matos, S., Oliveira, J.: Gimli: open source and high-performance biomedical name recognition. BMC bioinformatics 14(1), 54 (2013)
4. Campos, D., Matos, S., Oliveira, J.L.: A document processing pipeline for annotating chemical entities in scientific documents. J Cheminform 7(Suppl 1), S7 (2014)
5. Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., Rosenstein, M.C., Wiegers, T.C., Mattingly, C.J.: Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic acids research 37(Database issue), D786–92 (Jan 2009)
6. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic acids research 36(suppl 1), D344–D350 (2008)
7. Hettne, K.M., Stierum, R.H., Schuemie, M.J., Hendriksen, P.J.M., Schijvenaars, B.J.a., Mulligen, E.M.v., Kleinjans, J., Kors, J.a.: A dictionary to identify small molecules and drugs in free text. Bioinformatics (Oxford, England) 25(22), 2983–2991 (Nov 2009)
8. Krallinger, M., Rabal, O., Lourenço, A., Perez-Perez, M., Rodriguez, G.P., Vazquez, M., Leitner, F., Oyarzabal, F., Valencia, A.: Overview of the CHEMDNER patents task. In: Fifth BioCreative Challenge Evaluation Workshop. pp. 63-75. Seville, Spain (2015).
9. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., et al.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of cheminformatics 7 (1), S2 (2015)
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
11. Liu, H., Hu, Z.Z., Zhang, J., Wu, C.H.: BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics 22, 103–105 (2006)
12. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit (2002)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

14. Sagae, K.: Dependency parsing and domain adaptation with LR models and parser ensembles. In: Eleventh Conference on Computational Natural Language Learning. pp. 1044–1050. Association for Computational Linguistics, Prague, Czech Republic (2007)
15. Sasaki, Y., Montemagni, S., Pezik, P., Rebholz-Schuhmann, D., McNaught, J., Ananiadou, S.: Biolexicon: A lexical resource for the biology domain. In: Third International Symposium on Semantic Mining in Biomedicine, pp. 109–116. Jena, Germany (2008)
16. Smith, L., Tanabe, L. K., Ando, R. J. nee, Kuo, C.-J., Chung, I.-F., Hsu, C.-N., et al.: Overview of BioCreative II gene mention recognition. Genome Biology 9 (Suppl 2), S2 (2008)