Identification of chemical and gene mentions in patent texts using feature-rich conditional random fields

Sérgio Matos, José Sequeira, David Campos, and José Luís Oliveira

IEETA/DETI, University of Aveiro,
Campus Universitrio de Santiago, 3810-193 Aveiro, Portugal
{aleixomatos,sequeira,jlo}@ua.pt
BMD Software, 3810-074 Aveiro, Portugal
david.campos@bmd-software.com

Abstract. This article describes the application of Neji, a text-processing and concept recognition framework, to the automatic recognition of chemicals and gene mentions in medicinal chemistry patents. We used conditional random fields models trained with a otimized set of features including linguistic, orthographic, morphological, dictionary matching and local context features, dictionary-matching, and post-processing based on exclusion lists. Using cross-validation on the merged training and development sets of the BioCreative V CHEMDNER Patents task, we obtained an average F-score of 82.8% for the identification of chemicals and 42.4% for the identification of gene names.

Key words: Chemicals, Patents, Named Entity Recognition, Machine Learning

1 Introduction

The BioCreative V CHEMDNER Patents task[1] aimed at evaluating and promoting the application of automatic text mining methods for the identification of chemical and biological data mentioned on medicinal chemistry patents. Three different sub-tasks were considered:

- Chemical entity mention in patents (CEMP): the main named-entity recognition (NER) task, aimed at the detection of chemical named entity mentions;
- Chemical passage detection (CPD): a text classification task to identify patent titles and abstracts that mention chemical compounds;
- Gene and protein related object task (GPRO): a NER task aimed at identifying mentions of gene and protein related objects mentioned in patent tiles and abstracts.

[1] http://www.biocreative.org/tasks/biocreative-v/track-2-chemdner/

Our approach for each of the tasks was based on conditional random fields (CRF) models and dictionary-matching. We used the provided training and development corpora to identify the best feature set for training the models, and compared different combinations of model order, parsing direction, and exclusion lists.

2 Materials and methods

We used the concept recognition framework Neji [3] and applied a combined machine learning (ML) and dictionary-matching approach, similar to that described in [2]. Neji integrates a machine learning component based on Gimli [1], which is used for feature extraction and for training the CRF models. Additionally, Neji includes modules for text processing and natural language processing (NLP), providing from sentence-splitting to dependency parsing, for dictionary-matching, and also for post-processing including parentheses correction and abbreviation resolution. Figure 1 illustrates the overall architecture of our solution.

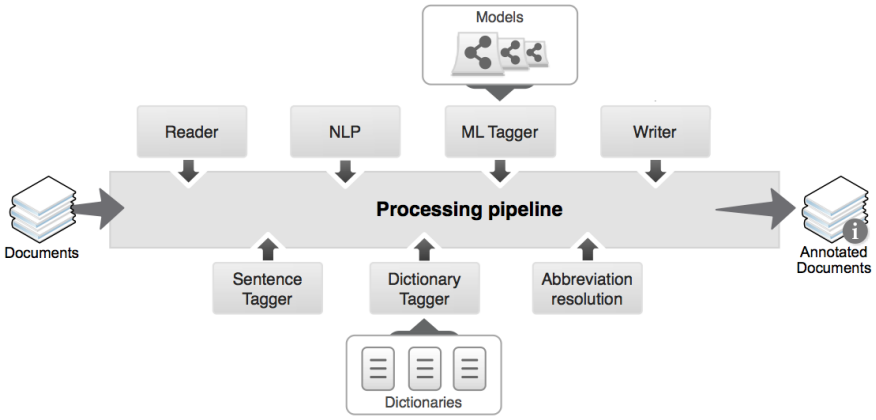

Fig. 1. Architecture of the Neji framework. Dictionary-matching and conditional random fields (CRF) models were used to annotate chemical and gene mentions.

2.1 Corpus

The organizers of the BioCreative V CHEMDNER Patents task provided a corpus divided in training, development and test set, each comprising a total of 7000 manually annotated documents (title and abstract). The patents were selected from several international patent organizations, and patents with an associated publication date between 2005 to 2014 and with titles and abstracts written in English were selected. Annotations for chemical entities are provided in seven

classes: systematic, identifiers, formula, trivial abbreviation, family and multiple. However, since descrimination between the classes was not an objective of the task, we grouped all classes into a single one. In the case of gene and protein related objects (GPROs), the annotations were divided in two groups: type 1, covering GPRO mentions that can be normalized to a database record; and type 2, including mentions that can not be normalized. Only the automatic identification of type 1 mentions was evaluated in this challenge. The training set contains a total of 33543 chemical mentions and 6876 gene mentions (4395 of which of type 1), and the development set contains 32142 chemical and 6263 gene mentions (3934 of type 1).

2.2 Pre-processing

Lingpipe[2] was applied to perform sentence splitting, using a model trained on biomedical corpora. NLP tasks (tokenization, lemmatization, part-of-speech (POS) tagging, chunking and dependency parsing) were performed using a custom version of GDep [1, 9]. The BIO scheme was used to encode the annotations.

2.3 Feature set

In order to select an optimized feature set, we performed recursive feature elimination using five folds of the merged training and development sets. The following feature set originated the best results for the CEMP task:

- NLP features:
  - Token, lemma, POS, chunk tags and dependency parsing features.
- Orthographic features:
  - Capitalization (e.g., "InitCap" and "AllCaps");
  - Digits and capitalized characters counting (e.g., "TwoDigit" and "TwoCap");
  - Symbols (e.g., "Dash", "Dot" and "Comma");
  - Greek letters (e.g., features for "alpha" and "α").
- Morphological features:
  - Prefixes;
  - 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").
- Domain knowledge:
  - Dictionary matching using a combined dictionary with terms from Jochem [6], ChEBI [5] and CTD [4].
- Local context:
  - Conjunctions of lemma and POS features of the windows {-1, 0}, {-2, -1}, {0, 1}, {-1, 1} and {-3, -1}.

Other features, such as windows and char n-grams, were tested but did not improve the cross validation results. We followed the same strategy for the GPRO task.

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

2.4 Model

We followed a supervised machine learning approach, using the implementation of Conditional Random Fields (CRFs) [7] provided by MALLET [8]. CRF models with different orders and with different parsing directions, that is forward (from left to right) and backward (from right to left), were tested. To harmonize overlapping annotations from different models, we apply a simple algorithm that considers the confidence scores provided by each CRF model and selects the one with the highest scores.

2.5 Dictionary matching

Dictionary matching was used both for obtaining features for training the CRF models and, in the case of chemical entity mentions, also for adding extra annotations that were missed by the machine-learning component, despite the dictionary-based features. For this, we filtered the combined Jochem, ChEBI and CTD dictionary, in order to minimized the number of false positives. To select the exclusion list to filter the dictionary, we obtained from the cross-validation runs the number of times a dictionary term was found as a false positive (FP) and as a true positive (TP). We set a threshold of 0.6 to the ratio $FP/(FP+TP)$, removing the terms that were above this threshold.

2.6 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. Annotations obtained through dictionary-matching were assigned a score of 1. For the CPD task, we calculated the average score of up to ten annotations with highest score and set that as the document score. This calculation aims to give higher score to documents containing more annotations, as these would, in principle, include more relevant information.

3 Results and discussion

CRF models with orders 1 and 2 and with forward and backward parsing were considered. Based on cross-validation results on the merged training and development set, models trained with backward parsing obtained consistently better results than forward parsing models on both CEMP and GPRO tasks. We also tested the combination of models of orders 1 and 2, and the combination of machine-learning models with dictionaries.

Table 1 presents the best results obtained on the merged set for the CEMP task, based on the following configurations:

1. Backward parsing CRF model of order 2;
2. Backward parsing CRF model of order 1, plus filtered dictionaries;

3. Backward parsing CRF model of order 2, plus filtered dictionaries;
4. Harmonized annotations of backward parsing CRF models of orders 1 and 2;
5. Harmonized annotations of backward parsing CRF models of orders 1 and 2, plus filtered dictionaries.

Table 1. Micro-averaged 5-fold cross-validation results in the merged training and development set for the CEMP task.

	Configuration	**Precision**	**Recall**	**F-score**
CEMP	1	82.12%	79.07%	80.50%
	2	81.42%	81.57%	81.51%
	3	81.49%	84.14%	82.80%
	4	80.90%	82.33%	81.60%
	5	81.41%	84.14%	82.80%

On the GPRO task, likely due to the small number of annotations, the second order model did not produce improved results when compared to the first order model. Moreover, the feature selection step improved the recognition performance of the model slightly, which may also be a result of the reduced amount of training data available. We submitted runs with the following configurations for the test phase of the task:

1. Backward parsing CRF model of order 1, trained with all annotations and using the GPRO features;
2. Backward parsing CRF model of order 1, trained with all annotations and using the CEMP features;
3. Backward parsing CRF model of order 1, trained on type 1 annotations only, using the GPRO features;
4. Backward parsing CRF model of order 1, trained on type 1 annotations only, using the CEMP features.
5. Harmonized annotations of backward parsing CRF models of order 1 trained on type 1 annotations only, using the GPRO and CEMP features.

4 Conclusion

This article presented a combined dictionary and CRF-based solution for chemical and gene mention recognition in patent documents. It uses a rich feature set including linguistic, orthographic, morphological, domain knowledge and local context (conjunctions) features. We tested both forward and backward parsing models, and well as the harmonization of CRF models of different orders. The best performance results achieved by cross-validation on the merged BioCreative V CHEMDNER training and development set were an F-scores of 82.8% on the CEMP, 86.3% on the CPD, and 42.4% on the GPRO task.

Acknowledgments. This work was supported by national funds through FCT - Foundation for Science and Technology, in the context of the projects UID/CEC/00127/2013 and Incentivo/EEI/UI0127/2014. DC has received support from the HemoSpec European project (EC contract number 611682). SM is funded by FCT under the FCT Investigator programme.

References

1. Campos, D., Matos, S., Oliveira, J.: Gimli: open source and high-performance biomedical name recognition. BMC bioinformatics 14(1), 54 (2013)
2. 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)
3. Campos, D., Matos, S., Oliveira, J.L.: A modular framework for biomedical concept recognition. BMC bioinformatics 14(281) (2013)
4. 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)
5. 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)
6. 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)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
8. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit (2002)
9. 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)