Chemical name recognition with harmonized feature-rich conditional random fields

David Campos, Sérgio Matos, and José Luís Oliveira

IEETA/DETI, University of Aveiro,
Campus Universitrio de Santiago, 3810-193 Aveiro, Portugal
{david.campos,aleixomatos,jlo}@ua.pt

Abstract. This article presents a machine learning-based solution for automatic chemical and drug name recognition on scientific documents, which was applied in the BioCreative IV CHEMDNER task, namely in the chemical entity mention recognition (CEM) and the chemical document indexing (CDI) sub-tasks. The proposed approach applies conditional random fields with a rich feature set, including linguistic, orthographic, morphological, dictionary matching and local context (i.e., conjunctions) features. Post-processing modules are also integrated, performing parentheses correction and abbreviation resolution. In the end, heterogeneous CRF models are harmonized to generate improved annotations. The achieved performance results in the development set are encouraging, with F-scores of 83.71% on CEM and 82.05% on CDI.

Key words: Chemicals, Named Entity Recognition, Machine Learning

1 Introduction

The BioCreative IV CHEMDNER challenge intends to promote the development of solutions to perform automatic recognition of mentions of chemical compounds and drugs on scientific documents, which is a challenging and complex task with increasing research interest [11]. Two different sub-tasks were organized:

- Chemical Entity Mention recognition (CEM): for a given document, provide the start and end indices corresponding to all mentioned chemical entities;
- Chemical Document Indexing (CDI): for a given document, provide a ranked list of mentioned chemical entities.

In order to participate in the CEM and CDI sub-tasks, we developed a machine learning-based solution taking advantage of the provided annotated corpus.

2 Materials and methods

The approach described in this document was developed on top of two frameworks: Gimli [2] and Neji [3]. Gimli is used for feature extraction and to train the machine learning (ML) models, and Neji is used to pre- and post-process the corpus and to apply multi-threaded document annotation. Figure 1 illustrates the overall architecture and required steps.

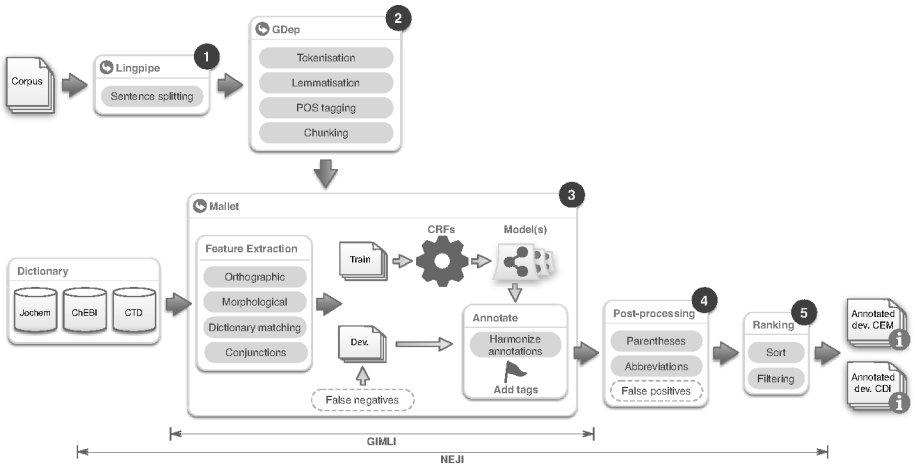

Fig. 1. Overall architecture of the described solution, presenting the pipeline of required steps, tools and external resources. Boxes with dotted lines indicate optional processing modules.

2.1 Corpus

The corpus provided by the challenge organizers[1] is divided in two sets: train and development. The train set contains 3500 abstracts annotated with 29478 chemical annotations, and the development set contains 3500 abstracts with 29526 chemical annotations. Annotations are provided in seven classes: systematic, identifiers, formula, trivial abbreviation, family and multiple. However, we grouped all classes into a single "master" class.

2.2 Pre-processing

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

2.3 Feature set

A rich feature set is applied to properly represent chemical names' characteristics:

- NLP features:
  - Token, lemma, POS and chunk tags.

[1] `http://www.biocreative.org/tasks/biocreative-iv/chemdner`
[2] `http://alias-i.com/lingpipe`

- Orthographic features:
  - 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:
  - Suffixes, prefixes and char n-grams of 2, 3 and 4 characters;
  - 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, capitalization and dependency parsing were tested but did not provided positive outcomes in the development set.

2.4 Model

A supervised machine learning approach is followed, through the application of Conditional Random Fields (CRFs) [7] provided by MALLET [8]. In order to obtain models with heterogeneous characteristics and achieve improved results, CRF models with different orders were considered. Additionally, we also trained models with different parsing directions: forward (from left to right) and backward (from right to left).

2.5 Post-processing

In order to solve some errors generated by the CRF model, our solution integrates two mandatory post-processing modules, implementing parentheses correction and abbreviation resolution. To perform parentheses correction, the number of parentheses (round, square and curly) on each annotation is verified and the annotation is removed if this is an odd number, since it clearly indicates a mistake by the ML model. Regarding abbreviation resolution, we adapt a simple but effective abbreviation definition recognizer [10], which is based on a set of pattern-matching rules to identify abbreviations and their full forms. Thus, if one of the forms is annotated as an entity name, the other one is added as a new annotation. Additionally, if one of the forms is not completely annotated, we expand the annotation boundaries using the result from the abbreviation extraction tool.

A third and optional post-processing module is added, in order to remove annotation mistakes and add non-learned annotations. After training a CRF model in the training set, we annotated the development set and analyzed the mistakes, collecting false positives and false negatives. That way, false negative annotations are added through dictionary matching and false positive annotations are discarded.

2.6 Harmonization

Since most recent results on biomedical NER indicate that better performance results can be achieved by combining annotations from systems with different characteristics [1], we apply a simple algorithm to harmonize annotations provided by CRF models with different orders. Thus, the harmonization algorithm considers the confidence scores provided by each CRF model and selects the intersecting and repeated annotations with the highest scores. If an annotation does not intersect with others, it is added to the final list of annotations.

2.7 Ranking

Ranking is provided based on 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 added through the dictionary of false negatives have a confidence score of one. In that way, raking simply orders the annotations in descending order of scores. In the case of the CDI task, an additional filtering step is applied, removing repeated annotations with the same case-insensitive text. In the end, a list of unique text annotations is obtained.

3 Results and discussion

CRF models with orders 1, 2, 3 and 4 were considered. Due to the long training times of higher order models (order 4 takes almost 24 hours), and the negative performance results obtained in the development set, we discarded models with orders 3 and 4. Afterwards, we tested CRF models with forward and backward parsing. However, backward parsing models did not provided positive outcomes in the development set. In the end, the following runs were submitted:

- Run 1: harmonized annotations of forward parsing order 1 and order 2 CRF models;
- Run 2: harmonized annotations of forward parsing order 1 and order 2 CRF models with false negatives annotation and false positives filtering;
- Run 3: forward parsing order 1 CRF model;
- Run 4: forward parsing order 1 CRF model with false negatives annotation and false positives filtering;
- Run 5: forward parsing order 2 CRF model.

Table 1 presents the final results achieved by each run on the development set of CEM and CDI tasks. Since runs 2 and 4 use false positives and false negatives collected from the development set, the presented results are overly optimistic. The real impact of such step will be evaluated using the test set.

Discarding overly optimistic runs, the best results are achieved by the harmonized solution, with F-scores of 83.71% on CEM and 82.05% on CDI. The harmonization approach outperforms single models due to significant recall improvements. The same is verified even when false negatives and false positives

are used. On the other hand, the order 1 model provides better results than the order 2 model on CEM and CDI tasks. Moreover, the order 1 model is the solution the achieves the best precision results on both tasks.

Table 1. Micro-averaged results in the development set of CEM and CDI tasks. Bold values indicate the best results discarding overly optimistic runs.

	Run	Precision	Recall	F-score	FAP-score
CEM	1	83.74%	**83.68%**	**83.71%**	**76.28%**
	2*	89.03%	89.30%	89.16%	82.58%
	3	**84.79%**	81.90%	83.32%	75.72%
	4*	89.11%	86.95%	88.02%	81.02%
	5	84.41%	80.41%	82.36%	74.58%
CDI	1	83.53%	**80.63%**	**82.05%**	**74.63%**
	2*	90.36%	87.45%	88.88%	82.14%
	3	**85.08%**	78.85%	81.85%	74.27%
	4*	90.38%	85.01%	87.61%	80.47%
	5	84.38%	79.40%	81.81%	74.45%

*overly optimistic run due to the usage of false positive and false negatives collected from the development set.

4 Conclusion

This article presented a CRF-based solution for automatic chemical and drug name recognition. It takes advantage of a rich feature set, namely linguistic, orthographic, morphological, domain knowledge (i.e., dictionary matching) and local context (i.e., conjunctions) features. Various post-processing modules are also integrated, performing parentheses correction and abbreviation resolution. In the end, CRF models with different orders are harmonized to obtain improved annotations. The final performance results achieved in the BioCreative IV CHEMDNER development set are encouraging, with F-scores of 83.71% on CEM and 82.05% on CDI. Further improvements may include using more and better domain knowledge, apply techniques for better context definition, and take advantage of an improved raking strategy.

Acknowledgments. This work was supported by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology, in the context of the projects FCOMP-01-0124-FEDER-010029 (FCT reference PTDC/EIA-CCO/100541/2008), FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011) and Incentivo/EEI/UI0127/2013. S. Matos is funded by FCT under the Ciência2007 programme.

References

1. Campos, D., Matos, S., Lewin, I., Oliveira, J.L., Rebholz-Schuhmann, D.: Harmonization of gene/protein annotations: towards a gold standard MEDLINE. Bioinformatics (Oxford, England) 28(9), 1253–1261 (May 2012)
2. Campos, D., Matos, S., Oliveira, J.: Gimli: open source and high-performance biomedical name recognition. BMC bioinformatics 14(1), 54 (2013)
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)
10. Schwartz, A.S., Hearst, M.A.: A simple algorithm for identifying abbreviation definitions in biomedical text. In: Pacific Symposium on Biocomputing. pp. 451–462. Computer Science Division, University of California, Berkeley, Berkeley, CA 94720, USA, Hawaii, HI, USA (2003)
11. Vazquez, M., Krallinger, M., Leitner, F.: Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. Molecular Informatics 30(6-7), 506–519 (2011)