Chemistry-specific Features and Heuristics for Developing a CRF-based Chemical Named Entity Recogniser

Riza Theresa Batista-Navarro, Rafal Rak, and Sophia Ananiadou

National Centre for Text Mining
Manchester Institute of Biotechnology
131 Princess St., Manchester
United Kingdom M1 7DN
riza.batista-navarro@cs.man.ac.uk
rafal.rak@manchester.ac.uk
sophia.ananiadou@manchester.ac.uk

Abstract. We describe and compare methods developed for the BioCreative IV chemical compound and drug name recognition (CHEMDNER) task. The presented conditional random fields (CRF)-based named entity recogniser employs a statistical model trained on domain-specific features, in addition to those typically used in biomedical NERs. In order to increase recall, two heuristics-based post-processing steps were introduced, namely, abbreviation recognition and re-labelling based on a token's chemical segment composition. The chemical NER was used to generate predictions for both the Chemical Entity Mention recognition (CEM) and Chemical Document Indexing (CDI) subtasks of the challenge. Results obtained from training a model on the provided training set and testing on the development set show that employing chemistry-specific features and heuristics leads to an increase in performance in both subtasks.

Key words: Chemical named entity recognition, Sequence labelling, Conditional random fields, Feature engineering

1 Introduction

Whilst most of the biomedical text mining efforts in the last decade have focussed on the identification of genes, their products and the interactions between them, there has been a recent surge in interest in extracting chemical information from the literature. Comprehensively capturing mentions of chemical molecules is not straightforward due to the various conventions by which they are referred to in text (trivial names, brand names, abbreviations, structures, etc.) [5, 8, 18]. Also, the development of publicly available, gold-standard, benchmark corpora for chemical named entity recognition has received relatively less attention [7] compared to that for gene or protein recognition.

Table 1. Feature set. Unigrams and bigrams are within left and right context windows of size 2. Newly introduced features are listed in the latter part of the table.

Feature	Description
Character	2,3,4-grams
Token	Unigrams and bigrams
Lemma	Unigrams and bigrams
POS tag	Unigrams and bigrams
Lemma and POS tag	Unigrams and bigrams
Chunk	Chunk tag, last word of chunk, presence of "the" in chunk
Ortography	Features used by Lee et al. [13]
Word shape	Full, collapsed, full with only numbers normalised
Dictionaries	Number of containing dictionaries, unigrams and bigrams
Affixes	Chemical prefix and suffix matches of sizes 2,3 and 4
Symbols	Matches with chemical element symbols
Basic segments	Number of chemical basic segments [4]

To close this gap, the organisers of the BioCreative IV CHEMDNER track released a corpus annotated with names of chemical compounds and drugs to encourage NLP researchers to implement tools capable of their automatic extraction. Two tasks were defined, namely, chemical document indexing (CDI) and chemical entity mention recognition (CEM). For the former, participants are asked to return a ranked list of unique chemical entities described in a document. The latter, in contrast, requires participants to provide the locations of all chemical name instances found. The corpus was split into training, development and test sets.

2 Systems description and methods

We cast the problem as a sequence labelling task and built a chemical NER based on NERsuite [3], an implementation of the CRF algorithm [12].

For both subtasks, we applied the following pre-processing steps on each document: sentence splitting using LingPipe MEDLINE Sentence Model [2], tokenisation with the OSCAR4 tokeniser [10], and part-of-speech and chunk tagging with GENIA tagger [17]. Machine learning features for each token are listed in Table 1. In generating the dictionary features, the following chemical resources were used: Chemical Entities of Biological Interest (ChEBI) [14], DrugBank [11], Joint Chemical Dictionary (Jochem) [9] and PubChem Compound [6]. We adopted the simple begin-inside-outside (BIO) label set for tagging tokens.

The output of tagging using the model (i.e., labels and marginal probabilities) were subjected to two heuristics-based post-processing steps. Abbreviation recognition is first performed in cases where a chemical named entity returned by the CRF model precedes a token enclosed by parentheses. Based on a simple algorithm [16], we verify whether the token qualifies as an abbreviation of the named entity. The second step involved re-labelling those "outside" ("O") tokens

Table 2. Evaluation on the development set for CEM

	Macro			Micro		
	P	R	F1	P	R	F1
NERsuite (Baseline)	86.66	79.01	80.89	88.55	76.82	82.27
NERsuite+Chem	**88.26**	81.11	82.86	**89.87**	78.98	84.07
NERsuite+Chem+H	87.71	**81.91**	**83.06**	89.28	**79.90**	**84.33**

Table 3. Evaluation on the development set for CDI. AvrgP = Average Precision; FAP-s = F-measured Average Precision score.

	Macro			Micro		
	F1	AvrgP	FAP-s	F1	AvrgP	FAP-s
NERsuite (Baseline)	82.59	75.53	78.90	83.33	73.68	78.21
NERsuite+Chem	84.93	77.97	81.30	**85.73**	76.80	81.02
NERsuite+Chem+H	**84.95**	**78.43**	**81.56**	85.72	**77.10**	**81.18**

whose marginal probabilities were less than a chosen threshold. The purpose of the re-labelling was to alleviate the problem of partial recognition of chemical names. Having extracted the chemical basic segments [4] in each such token, we computed the ratio of the number of characters making up the basic segments to the total number of characters in the token. If the ratio was greater than a chosen threshold, the token was re-labelled as part of a chemical name.

In order to rank chemical names found in a document, we used marginal probabilities given by the NER. For each chemical name, we computed a confidence score by taking the average of the marginal probabilities of the tokens composing the chemical name. Chemical names were then ranked in decreasing order of their confidence scores.

3 Results and Discussion

Tables 2 and 3 show the results of evaluating our NER on the development set using models trained on the training set. The addition of chemistry-specific features (NERsuite+Chem) in the CEM task (Table 2) increased the precision. By adding both chemistry-specific features and post-processing heuristics (NERsuite+Chem+H), we obtained optimal recall and F1-score.

Ranking-based measures for the CDI task (shown in Table 3) exhibit similar behaviour, i.e., the performance of the NER with chemistry-specific features and post-processing heuristics is superior to the remaining methods. The difference between CEM and CDI results in terms of F1 is the result of having a unique list of chemical names for each document for the CDI task, which is not the case for the CEM task.

In producing the predictions for the CHEMDNER test set, we trained a final model on both the training and development sets. The five runs submitted correspond to the variants described in Table 4.

We are planning to make the tool public by including the entire processing pipeline in Argo, a web-based, text-mining platform [1, 15].

Table 4. Variants of NERsuite+Chem used for the five runs

	Description
NERsuite+Chem	without heuristics
NERsuite+Chem+H_a	with only abbreviation recognition
NERsuite+Chem+H_{t1}	with only token relabelling, thresholds optimised for precision
NERsuite+Chem+H_1	with both heuristics, thresholds optimised for precision
NERsuite+Chem+H_2	with both heuristics, thresholds optimised for recall

Acknowledgements This work was partially supported by Europe PubMed Central funders (led by Wellcome Trust).

References

1. Argo: A Web-based Text Mining Workbench. `http://argo.nactem.ac.uk`
2. LingPipe 4.1.0. `http://alias-i.com/lingpipe`
3. NERsuite: A Named Entity Recognition toolkit. `http://nersuite.nlplab.org`
4. American Chemical Society: Registry file basic name segment dictionary. Tech. rep. (1993)
5. Banville, D.L.: Mining chemical structural information from the drug literature REVIEWS 11(1), 35–42 (2006)
6. Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H.: PubChem: Integrated Platform of Small Molecules and Biological Activities. Annual Reports in Computational Chemistry 4 (2008)
7. Grego, T., Pesquita, C., Bastos, H.P., Couto, F.M.: Chemical Entity Recognition and Resolution to ChEBI. ISRN Bioinformatics 2012, 9 (2012)
8. Gurulingappa, H., Mudi, A., Toldo, L., Hofmann-Apitius, M., Bhate, J.: Challenges in mining the literature for chemical information. RSC Adv. 3, 16194–16211 (2013)
9. 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 25(22), 2983–2991 (2009)
10. Jessop, D., Adams, S., Willighagen, E., Hawizy, L., Murray-Rust, P.: OSCAR4: a flexible architecture for chemical text-mining. Journal of Cheminformatics 3(1), 41 (2011)
11. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C., Wishart, D.S.: Drugbank 3.0: a comprehensive resource for omics research on drugs. Nucleic Acids Research 39(suppl 1), D1035–D1041 (2011)

12. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
13. Lee, K.J., Hwang, Y.S., Kim, S., Rim, H.C.: Biomedical named entity recognition using two-phase model based on SVMs. Journal of Biomedical Informatics 37(6), 436 – 447 (2004)
14. de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C.: Chemical entities of biological interest: an update. Nucleic Acids Research 38(suppl 1), D249–D254 (2010)
15. Rak, R., Rowley, A., Black, W., Ananiadou, S.: Argo: an integrative, interactive, text mining-based workbench supporting curation. Database : The Journal of Biological Databases and Curation p. bas010 (2012)
16. Schwartz, A.S., Hearst, M.A.: A simple algorithm for identifying abbreviation definitions in biomedical text. In: Pacific Symposium on Biocomputing. pp. 451–462 (2003)
17. Tsuruoka, Y., Tateisi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. In: Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS, vol. 3746, pp. 382–392. Springer-Verlag, Volos, Greece (November 2005)
18. Vazquez, M., Krallinger, M., Leitner, F., Valencia, A.: Text mining for drugs and chemical compounds: Methods, tools and applications. Molecular Informatics 30(6-7), 506–519 (2011)