Adapting ChER for the recognition of chemical mentions in patents

Riza Batista-Navarro and Sophia Ananiadou

National Centre for Text Mining, School of Computer Science
University of Manchester
{riza.batista,sophia.ananiadou}@manchester.ac.uk

Abstract. ChER (Chemical Entity Recogniser) is a pipeline of natural language processing tools optimised for the recognition of chemical names in scientific abstracts. It formed the basis of our submissions to the previous edition of the CHEMDNER track in BioCreative IV, and was one of the top-performing systems both for the chemical document indexing (CDI) and chemical entity mention recognition (CEM) subtasks. As part of our contribution to the new chemical patents-focussed CHEMDNER track in BioCreative V, we adapted ChER for the recognition of chemical names in medicinal chemistry patent documents. As our previous study showed that the incorporation of chemical dictionary features brings about the most significant boost in performance, we focussed our current efforts on investigating the contribution of features drawn from various semantic resources, including RxNorm, US FDA's Orange Book, Health Canada's Drug Product Database and BioSemantics' Chemical Patent Corpus. Results of our methods on the CHEMDNER test corpus demonstrate that the addition of features based on patent-specific dictionaries gives the most optimal performance on the Chemical Entity Mention in Patents (CEMP) task.

Key words: Chemical name recognition, Conditional random fields, Feature engineering, Text mining, Patent mining

1 Introduction

Extraction of chemical information from text has gained significant attention from the text mining community recently. Contributions of various text mining groups to the CHEMDNER track of BioCreative IV, for example, have advanced the state of the art for chemical document indexing and chemical name recognition based on scientific abstracts [15]. Another rich source of chemical information, however, are patent documents which are characterised by a specific subdomain language. Thus far, only a handful of information extraction tools have attempted to address text mining from chemical patents [10, 14, 19]. The CHEMDNER track of BioCreative V was thus organised to foster the development of more methods for chemical patent mining. It consists of three subtasks: (1) the Chemical Entity Mention in Patents (CEMP) task, focussed on the demarcation of chemical mentions; (2) the Chemical Passage Detection (CPD) task

for detecting whether a patent contains chemical mentions or not; and (3) the Gene and Protein-Related Object (GRPO) task, which required the recognition of gene/protein mentions. As part of our contribution to the CEMP task, we adapted our pipeline of natural language processing tools, henceforth referred to as ChER [7], for the recognition of chemical mentions in patents. Drawing inspiration from our previous experience in BioCreative IV which showed that the incorporation of dictionary features brought about the largest boost in performance, we investigated the further semantic enrichment of our features by considering patent-specific resources. Results of the evaluation of our methods on the CHEMDNER test corpus show that the inclusion of features from all dictionaries we have at hand leads to optimal performance.

2 Systems description and methods

In this section, we first provide an overview of ChER and then proceed to describing the features from chemical resources that we have incorporated into it, as a means for adapting it for chemical mention recognition in patents.

2.1 Overview of the Chemical Entity Recogniser (ChER)

ChER is a text mining pipeline which consists of several natural language processing tools, primarily based on the conditional random fields (CRF) algorithm [16]. Its optimised performance on chemical mention recognition in PubMed abstracts, as well as availability as a Web-based workflow[1], was facilitated by our text mining workbench Argo [18]. Below are its various components.

Cafetiere Sentence Splitter. Each input document is processed by the rule-based Cafetiere Sentence Splitter [2], which has been specifically designed to handle biomedical text.

OSCAR4 Tokeniser. Sentences are segmented into tokens by the tokeniser that comes with OSCAR4 [13] which was optimised for text containing chemical mentions. It is capable of keeping intact long systematic names, even if they contain punctuation.

GENIA Tagger. Tokens are enriched with their lemmatised forms, part-of-speech (POS) and chunk tags by the GENIA Tagger [20], which comes with a model trained on biomedical text.

NERsuite Tagger. Our chosen implementation of CRFs is the NERsuite package, which allows for the training and application of statistical models for sequence labelling [4]. The task was ultimately defined as the assignment of begin-inside-outside (BIO) labels to tokens. By default, NERsuite represents each token as a

[1] http://argo.nactem.ac.uk

set of lexical, orthographic and syntactic features, details of which can be found in our previous publication [7]. We enriched this feature set by incorporating semantic features such as chemical prefixes/suffixes as well as token matches with entries in dictionaries such as ChEBI [11], DrugBank [17], the Comparative Toxicogenomics Database (CTD) [9], PubChem Compound [8] and the Joint Chemical Dictionary (Jochem) [12]. As we shall describe later, however, we explored additional resources containing patent-specific information in order to adapt ChER for the current task.

Post-processing heuristics. As a means for increasing recall, we integrated into ChER two post-processing methods: one for abbreviation resolution and the other for calculating the chemical segment make-up of tokens. We refer the reader to our previous work [7] for a detailed description of these methods.

2.2 Features from patent-specific resources

In our aim to adapt ChER for the recognition of chemical mentions in patents, we investigated the enrichment of our semantic features using resources that contain patent-specific information. We describe each of these additional resources below.

RxNorm. Integrating several terminological resources, RxNorm contains the generic and brand names of thousands of clinical drugs and drug packs [5]. It includes information from databases that store industry-used names and identifiers, such as the Medi-Span Master Drug Data Base and the US Food and Drug Administration's Structured Product Labels.

Orange Book. The US Food and Drug Administration (FDA) publishes a list of approved drug products commonly known as the Orange Book [1]. It contains patent-relevant information such as active ingredient, proprietary name, applicant name and application number.

Drug Product Database. Similar to the US FDA's Orange Book, the Drug Product Database (DPD) published by Health Canada [3] contains information on drugs approved for use but in Canada. It also includes patent application information.

Chemical Patent Corpus. The BioSemantics group of the Erasmus Medical Center in Rotterdam (the Netherlands) published a corpus of 200 full patents in which mentions of chemical entities have been annotated [6]. We compiled a list of drug mentions based on the annotated text spans in this corpus. These include IUPAC names, SMILES and InChi strings, generic and brand names, abbreviations, formulas and registry numbers.

The richest version of our feature set includes token matches with all nine dictionaries (i.e., the same five dictionaries used in the original version of ChER and the four additional ones just described).

Table 1: Evaluation of the patents-adapted version of ChER on the CHEMDNER test corpus

	Dictionaries used	Precision	Recall	F-score
Run 1	All	87.878	**84.282**	**86.042**
Run 2	Jochem	88.020	83.060	85.468
Run 3	CTD	87.915	83.269	85.529
Run 4	RxNorm	**88.073**	83.195	85.565
Run 5	CTD+Regex	87.327	83.298	85.265

3 Results and Discussion

Using NERsuite, we trained CRF models with features extracted over the combined CHEMDNER training and development corpora, containing a total of 14,000 patent documents. Five slightly different versions of our method were officially evaluated on the CHEMDNER test corpus of 7,000 patents, the results of which are shown in Table 1. In Run 1, all nine resources were utilised in the generation of dictionary features. Each of Runs 2, 3 and 4, meanwhile, exploited only one dictionary: Jochem, CTD and RxNorm, respectively. These were chosen on the basis of them having the most number of manually reviewed chemical names. Run 5 made use of CTD in conjunction with a regular expression that matches chemical formulas. Results show that the versions of our approach utilising patent-specific dictionaries yielded the best performance. Run 4, which exploited RxNorm as its sole dictionary, obtained the best precision (88.073%) most likely due to it being the version that took advantage of only a patent-specific resource. Run 1, which leveraged all dictionaries, including the four patent-specific ones, obtained optimal recall (84.282%) and F-score (86.042%).

References

1. Approved Drug Products with Therapeutic Equivalence Evaluations (Orange Book) http://www.fda.gov/Drugs/InformationOnDrugs/ucm129662.htm, accessed: July 2015
2. Cafetiere English Sentence Detector http://metashare.metanet4u.eu/repository/browse/u-compare-cafetiere-english-sentence-detector/aff1ddc0bc8911e1a404080027e73ea259aeca28412944ea97f7b2580a41caec/#, accessed: July 2015
3. Drug Product Database http://www.hc-sc.gc.ca/dhp-mps/prodpharma/databasdon/index-eng.php, accessed: July 2015
4. NERsuite: A Named Entity Recognition toolkit. http://nersuite.nlplab.org, accessed: July 2015
5. RxNorm Overview http://www.nlm.nih.gov/research/umls/rxnorm/overview.html, accessed: July 2015
6. Akhondi, S.A., Klenner, A.G., Tyrchan, C., Manchala, A.K., Boppana, K., Lowe, D., Zimmermann, M., Jagarlapudi, S.A.R.P., Sayle, R., Kors, J.A., Muresan, S.:

Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. PLoS ONE 9(9), e107477 (09 2014)
7. Batista-Navarro, R., Rak, R., Ananiadou, S.: Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. Journal of Cheminformatics 7(Suppl 1), S6 (2015)
8. Bolton, E., Wang, Y., Thiessen, P., Bryant, S.: PubChem: Integrated Platform of Small Molecules and Biological Activities. Annual Reports in Computational Chemistry 4 (2008)
9. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Wiegers, T.C., Mattingly, C.J.: The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Research (2014)
10. Grego, T., Pezik, P., Couto, F., Rebholz-Schuhmann, D.: Identification of chemical entities in patent documents. Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part II: Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living. IWANN '09 pp. 942–949 (2009)
11. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C.: The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Research (2012)
12. Hettne, K., Stierum, R., Schuemie, M., Hendriksen, P., Schijvenaars, B., van Mulligen, E., Kleinjans, J., Kors, J.: A dictionary to identify small molecules and drugs in free text. Bioinformatics 25(22), 2983–2991 (2009)
13. 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)
14. Jessop, D., Adams, S., Murray-Rust, P.: Mining chemical information from open patents. Journal of Cheminformatics 3(1), 40 (2011)
15. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: CHEMDNER: The drugs and chemical names extraction challenge. Journal of Cheminformatics 7(Suppl 1), S1 (2015)
16. 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)
17. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B., Zhou, Y., Wishart, D.S.: DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Research 42(D1), D1091–D1097 (2014)
18. 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 2012, bas010 (2012)
19. Sayle, R., Xie, P.H., Muresan, S.: Improved Chemical Text Mining of Patents with Infinite Dictionaries and Automatic Spelling Correction. Journal of Chemical Information and Modeling 52(1), 51–62 (2012)
20. 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, Lecture Notes in Computer Science, vol. 3746, pp. 382–392. Springer-Verlag, Volos, Greece (November 2005)