Recognition of Chemical Entity Mention in Patents using CRF, Domain Specific Dictionaries and Features

Venkata Ravindra Nittala[1,2*], Srinivas Jonnalagadda[2], and Manish Shrivastava[1]

[1] Language Technologies Research Center, International Institute of Information Technology, Gachibowli, Hyderabad, 500 032, India
[2] Ojus Software Labs Private Limited, 16-11-310/12/A/1/1, Saleem Nagar Colony 2, Malakpet, Hyderabad, 500 036, India
ravindra.n@students.iiit.ac.in, js@ojuslabs.com, m.shrivastava@iiit.ac.in

Abstract. We present a system employing domain specific dictionaries and features to recognize chemical entities. The system utilizes sentence segmentation, tokenization, feature generation, Conditional Random Field (CRF) training and one post-processing step. The dictionaries were compiled from PubChem, Wikipedia, ChEMBL, DrugBank, word2vec clusters from US patents belonging to A61K class. We report the evaluation results of the run where development set was not included as part of the training set. The best performing model for CEMP task has the micro average precision, recall and F-score values of **87.26%**, **79.98%** and **83.46%**, respectively.

Key words: Chemical Named Entity Recognition, NER, CRF, Domain-dependent, Dictionary

1 Introduction

Chemical patents are a significant source of competitive intelligence in both chemical and pharmaceutical industries. These are filed as applications months, or even years, ahead of disclosure in peer reviewed journal articles. However, analyzing patents manually is a labour-intensive task, as they include numerous examples, assay data points and claims expressed as Markush structures. It can be very time consuming to extract useful information from patents manually, even for experts. Hence, Natural Language Processing (NLP) and text mining technologies are key to automating the extraction of information from patent text. Recognizing chemical named entities is a useful first step in this direction. In this paper we report our participation in BioCreative V CHEMDNER CEMP (chemical entity mention in patents) main task.

2 Methods

2.1 Overview

Overview of the chemical compound and drug name recognition (CHEMDNER) community challenge was reviewed by Krallinger et al [1]. CEMP task extends the challenge to text from medicinal chemistry patents.

2.2 CRF

Conditional random field is a probabilistic framework for labeling and segmenting data. It is a form of undirected graphical model that defines a single log-linear distribution over label sequences given a particular observation sequence. We treated CEMP as a sequence-labelling task, hence selected CRF as the learning algorithm. CRFs offer several advantages over hidden Markov models and stochastic grammars for sequence-labelling task, with the ability to relax strong independence assumptions and also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models, which can be biased towards states with few successor states [2].

2.3 Experiment Setup

CRF++ software [3] was utilized for training CRF models. Five models were trained for submission and all models are based on CRF, domain specific dictionaries, features and utilize IOB label set. Models 1, 3, 4 and 5 were trained on CEMP training set and evaluated on CEMP development set. Model 2 was trained on CEMP development set and evaluated on CEMP training set. In training CRF models, context words occurring within the window of four words before and after the current word were selected. Experiments were performed by varying the cut-off threshold of features (f) and hyper-parameter (c) in CRF++. In models 1, 2, 3, 4 and 5, parameter f was set to 4, 4, 3, 3 and 2 respectively. The parameter c was set to 3.0 in all models except model 4 where it was set to 4.0. All other parameters were set to their default values.
Two custom programs were written in Java version 8 to generate input for CRF learning and to decode the CRF models output. Java's *java.text.BreakIterator* [4] was utilized for sentence segmentation. Word tokenization was performed on sentences with space as delimiter; in addition letters, numbers and punctuation characters were treated as separate tokens. No distinction was made between lowercase and uppercase letters. e.g. '(CH2)pCOOC 1-6alkyl' would be tokenized as '(CH 2) pCOOC 1 - 6 alkyl'. Results were evaluated with Official BioCreative II.5 evaluation script version 3.2 using 'exact mention' strategy.

2.4 Features

Systematic names (e.g. 6,7-methylenedioxy-4-(3'-methylanilino)quinazoline) and trivial names (e.g. Codeine or 3-methylmorphine) include many digits, punctuation, non-alphanumeric characters and semantically rich morphemes. In considering the above fact, the following features were extracted from the training data and utilized in training CRF models.

Word, Lemma and POS tags: The token itself (in lowercase) was added as a feature. Lemma and POS tags were generated using Hepple Tagger available in Dragon Toolkit [5].

Space: Binary features for space occurring before and after token in original text.

Character Counts: Character counts of digits, uppercase and lowercase letters in each token.

Punctuation Characters: Presence of punctuation characters, dash, mathematical equals, mathematical identical.

Chemical Prefix and Suffix: Most frequently occurring prefixes, suffixes of length 3 and 4 were extracted from the systematic and trivial name dictionaries. e.g. Prefixes: meth, etha, prop, etc. Suffixes: ane, ene, yne, etc.

Case Pattern Features: Feature created by replacing all uppercase letters with 'A', all lowercase letters with 'a' and all digits with '0' upto length of 8 characters [6].

Character n-grams of Token: Character n-grams extracted from token of length 1 to 4 were added as features. e.g. acetyl would have: a, ac, ace, acet, l, yl, tyl and etyl.

Presence in Dictionary: A binary feature for presence of Chemical element names, Chemical element symbols, Amino acid names, Amino acid codes of length 1 and 3, Systematic names, Trivial names, Family names, Greek letters and Greek symbols.

Molecular Formula Parser: A method to assign a numeric score to a string on scale of 0 to 1. Where a score of 1 indicates the string is 100% molecular formula and a score of 0 indicates this is not a molecular formula. e.g. NaCl and PhNMe would have a score of 1.0000.

2.5 Dictionaries

Systematic names, trivial names, family names were assembled as dictionaries, after pre-processing to remove any punctuation characters, digits, alphabetic strings of length less than 4 characters, and converting all the strings to lowercase. Any words found in English dictionary which are not related to chemicals were removed.

Systematic and Trivial Names: PubChem Compound SDF files [7], ChEMBL Database [8], DrugBank entries [9–12], Wikipedia article titles [13].

Family Names: Titles and abstracts were collected from 62,762 US patents [14] belonging to A61K class filed from year 2005 to 2015. After pre-processing, word clusters were generated using word2vec software. Clusters having words like 'alkyl', 'aryl', 'alkenyl', etc., were merged and used as dictionary of family names. This corpus of 62,762 patents were also utilized to check the consistency of sentence segmentation.

2.6 Post Processing

One post-processing step was done on recognized mentions. An attempt was made to balance each mention of parenthesis, square brackets and curly brackets, by adding or removing one character to the right or left of the mention. If this did not balance parenthesis, that mention was not included in final results [6].

3 Experimental Results

Table 1 show the results of five models trained on official CEMP corpus. It shows that CRF is able to achieve very high precision of about 87-88%. Though, the recall value has been consistently low at around 79.9%. On investigation of errors, it was concluded that most of the errors are due to tokenization, character set conversion, missing space and other typographical errors in the patent text corpus.

Table 1. Results on training and development set

	Micro Average			**Macro Average**		
Model	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
1	0.8721	0.7998	0.8344	0.8309	0.7777	0.7867
2	0.8869	0.7924	0.8370	0.8388	0.7618	0.7823
3	0.8715	0.7999	0.8342	0.8299	0.7764	0.7855
4	0.8712	0.7994	0.8337	0.8288	0.7768	0.7852
5	**0.8726**	**0.7998**	**0.8346**	0.8307	0.7773	0.7865

4 Conclusion

In this paper we report about our submission to BioCreative V CHEMDNER Track-2 CEMP task. It was a enriching experience participating in CEMP task. The use of domain specific dictionaries, features and CRF as machine learning algorithm, led to achieving very high precision of the system. There is further scope for improvement, in increasing recall of the system. We propose to investigate new features and unsupervised learning methods relevant to achieving high recall.

Acknowledgments. We would like to acknowledge the financial support received from Ojus Software Labs Pvt. Ltd.

References

1. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2013, October). Overview of the chemical compound and drug name recognition (CHEMDNER) task. In BioCreative Challenge Evaluation Workshop (Vol. 2, p. 2).
2. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ICML. pp. 282289 (2001)
3. CRF++: Yet Another CRF toolkit, `https://taku910.github.io/crfpp/`
4. Java Platform Standard Edition 8 Documentation, `https://docs.oracle.com/javase/8/docs/`
5. Zhou, X., Zhang, X., and Hu, X., "Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining," In proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), October 29-31, 2007, Patras, Greece
6. Leaman R, Wei C-H, Lu Z (2015) tmChem: a high performance tool for chemical named entity recognition and normalization, Journal of Cheminformatics, 7(Suppl 1): S3
7. Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 IN Wheeler RA and Spellmeyer DC, eds. Annual Reports in Computational Chemistry, Volume 4. Oxford, UK: Elsevier, 2008, pp. 217-241. doi:10.1016/S1574-1400(08)00012-1.
8. A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Kruger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos and J.P. Overington (2014) 'The ChEMBL bioactivity database: an update.' Nucleic Acids Res., 42 1083-1090.
9. DrugBank 4.0: shedding new light on drug metabolism.Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS.Nucleic Acids Res. 2014 Jan 1;42(1):D1091-7.
10. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. 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 AC, Wishart DS.Nucleic Acids Res. 2011 Jan;39(Database issue):D1035-41.
11. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M.Nucleic Acids Res. 2008 Jan;36(Database issue):D901-6.
12. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J.Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72.
13. Wikipedia: The Free Encyclopedia. Wikimedia Foundation Inc. `https://en.wikipedia.org/wiki/Main_Page`
14. United States Patent and Trademark Office Bulk Downloads, `https://www.google.com/googlebooks/uspto.html`
15. word2vec Tool for computing continuous distributed representations of words. `https://code.google.com/p/word2vec/`