ChemGrab: Identification of Chemical Names Using a Combined Negative-Dictionary and Rule-Based Approach

[1]Vivekanand Sharma, PhD and *[2]Indra Neil Sarkar, PhD, MLIS

Center for Biomedical Informatics, Brown University, Providence, RI USA

[1]vivekanand_sharma@brown.edu;
*[2]neil_sarkar@brown.edu

Abstract. The growing volume of electronically available text provides the opportunity to extract potentially relevant information that may offer valuable insights. To this end, the cataloguing of documents based on named entity mentions is an essential task. Development of text mining approaches for extracting entities and relationships may enable efficient management and retrieval of relevant information within specific contexts. The system described here, ChemGrab, focused on the BioCreative V.5 CEMP Challenge that aims to identify mentions of chemical entities from within patent text. The approach used in this study to identify chemical mentions used a combination of a negative-dictionary and rules based on word-level features. The system performance on the test set achieved a micro precision, recall, and F-score of 0.53, 0.67, and 0.59, respectively.

Keywords. Named Entity Recognition; Chemical Entity Mention in Patents; Natural Language Processing

1 Introduction

The identification and organization of information within patent texts can be a crucial step [1, 2] for facilitating novelty checking, validation, and identification of starting points for knowledge discovery [3]. However, the extraction of chemical and biological entities from patent text is a challenging task due to significant differences in linguistic structures [4]. Additional challenges arise due to complexities in chemical names, term ambiguity, complex syntactic structures, and Optical Character Recognition errors [5]. Nonetheless, cataloging of patent information can be crucial in accelerating research, devising Intellectual Property management strategies, and promoting technology transfer [2].

In light of the increasing amount of digitally available text data, there is opportunity for the development of automated methods for mining key entities and relationships. Such mining can support and enhance retrieval and curation of relevant documents. To develop natural language processing (NLP) strategies for extracting named entities from patent texts, the availability of annotated corpora is an essential step. Significant effort has been invested in creation of annotated corpora (gene, protein, chemicals) in the biomedical domain (such as from scientific literature indexed in MEDLINE) [6–8]. With the recognition of the importance of mining patent data, there has been some recent effort in this direction [1, 5]. Towards encouraging the development of tools and methods for automated recognition of chemical and biological entities from medicinal chemistry patents, the BioCreative-related initiatives were organized to provide manually annotated patent text corpora for supporting the development of NLP tools that could be effectively benchmarked [4].

Named entity recognition techniques often use dictionaries containing domain specific vocabularies [9]. Systems such as Peregrine [10], TaxonGrab [11], and LINNAEUS [12] have demonstrated the utility of dictionary matching approaches to identify disease and organism names. In context of chemical entities, dictionary-matching approaches can be challenging and require post-processing rules [13]. Furthermore, high coverage of chemical concepts with dictionaries alone can be difficult due to the range and volume of novel compound names [14]. To address this challenge, rule-based and machine learning techniques have been shown to improve performance when used in combination with dictionary-based approaches [13, 15]. Finally, ensemble approaches, which combine multiple machine learning models have also shown promising results [16].

Towards achieving the goal of recognition of chemical name entity mentions for the Chemical Entity Mention in Patents (CEMP) task, this study focused on leveraging a combination of dictionary-matching and rule-based entity recognition approaches. The results from evaluation on a training and test set highlight the efficacy of approach and identify several areas for improvement.

2 Methods

ChemGrab is a system that was developed in this study, which relies on dictionaries for matching and identification of feature rules for identification of chemical tokens. The identified chemical name tokens are expanded to chemical entities, and include the start and end indices. The implementation of ChemGrab was done using Julia (v.0.5).

2.1 Dictionary of chemicals

The Jochem dictionary was used as the look-up dictionary for this study [17]. This dictionary is based on combination of information from Unified Medical Language System (UMLS), Medical Subject Headings (MeSH), Chemical Entities of Biological Interest (ChEBI), DrugBank, Kyoto Encyclopedia for Genes and Genomes (KEGG), Human Metabolome Database (HMDB), and ChemIDPlus.

2.2 Negative Chemical Dictionary

A combination of word lists from WordNet [18] and SPECIALIST Lexicon [19] was used to develop a negative dictionary of non-chemical entities. WordNet is a lexical database of English words, synonyms, and their variations. The SPECIALIST Lexicon, a UMLS knowledge source, consists of common English words and biomedical vocabulary. The lexicon includes words as well as their spelling and grammatical variants. From the combined list of these two sources, chemicals were excluded by comparison with Jochem based on an exclusion criteria of pair-distance between two strings (s1 and s2) calculated using similarity metric with a cut-off of 0.90:

$$pair\ distance\ score = \frac{2\times|pairs(s1)\ \cap\ pairs(s2)|}{|pairs(s1)|\ +\ |pairs(s2)|} \qquad (1)$$

Using the pair distance score, the chemical names from the negative dictionary were eliminated.

2.3 Word level features

Word level feature identification for chemical entity recognition was done based on a comparison of the negative and chemical dictionaries. The characters for each word were divided into three groups: (1) Alphabet (a-z); (2) Number (0-9); and (3) Other characters (excluding whitespace). From each respective dictionary, the occurrence frequency of characters with a specified separation distance within a given word (chemical or non-chemical) entity (referred to as "elements" hereafter) was recorded. A separation distance used was in the range of 1 to 5. A *tf-idf* inspired scoring method was used to downweight trivial elements and upweight rare ones.

$$score = \frac{(xy)_t^i}{(xy)_t} \times \log \frac{\sum_t (x\,*)_t}{\sum_t (x\,*)_t^i} \qquad (2)$$

where, $t \in$ {*chem dataset, non-chem dataset*}, i is the character separation distance, x and $y \in$ {*a-z, 0-9, other characters (excluding whitespace)*}, $(xy)_t^i$ is frequency of co-occurrence of character x and y at a given separation distance i within the chemical dictionary or the word dictionary. $(x\,*)_t^i$ is the frequency of co-occurrence of character x with any other character at a given separation distance i within the chemical dictionary or the word dictionary. When i is not specified, the term indicates co-occurrence irrespective of separation distance.

Each element received a chemical score and a non-chemical score as obtained from their respective dictionary. For each set of two-character combinations (e.g., a-a) at a given separation distance (e.g., 1-5) the difference of scores were normalized using a z-score calculation and standard normal curve area, which was used for calculating the final score (described in section 2.5). Critical elements were identified using a significance level of 0.05 for non-chemicals and 0.95for chemicals.

2.4 Tokenization

The goal of this step was to identify chemical entity names that contained whitespace. From the dictionary of chemical names containing whitespaces, a list was generated that contained three characters prior to whitespace and all of the following characters. The list was manually evaluated to remove entries that did not seemed relevant (e.g., “ium ion” or “hyl group”). Matching and replacement of whitespaces contained within entities was done using the above described list as a scaffold. The text segments were tokenized at ‘whitespace’, ‘/’, ‘-’ and ‘;’. Additional processing steps involved removal of non-alphabet and non-digit characters from the start and end positions of tokens. Non-chemical tokens were excluded based on comparison with the negative word list described in section 2.1.

2.5 Recognition of chemical entities

After tokenization, chemical entities were identified from those strings that were not identified in the negative dictionary using following three successive steps: (1) *Direct Lookup*: tokens were compared to the Jochem dictionary to identify direct matches; (2) *Rule-based identification*: word-level features as described in section 2.3 were used to determine whether a token was a chemical. The scores of critical elements were used to calculate a final score:

$$token\ score\ =\ 1 - \prod_{i\,=\,1}^{m}(1-E) \qquad (3)$$

where, *m* is the total number of critical elements (2) and *E* is the element score. A token score of one reflected a perfect chemical token. Thresholds of 0.97 and 1.00 were tested for identification of chemical tokens; (3) *Approximate matching*: each chemical entity from Jochem was indexed according to those that occurred five or less times and were the using lowest scoring elements (e.g., ‘o’, ‘l’, ‘1’). Using these features, potentially matching candidates were queried for further scoring. The similarity scoring was based on the Levenshtein metric, using a threshold of 0.7 (1 being exact match and 0 indicating no match). Following identification of all chemical tokens from the above four steps, the start and end indices of neighboring tokens were used to expand as a single chemical entity giving due consideration to the punctuations that occurred between them.

3 Results and Discussion

The evaluation was performed using the evaluation functionality within Biomedical Annotation Metaserver (BeCalm) participant account. This evaluation was quantified based on calculation of micro precision, recall and F-score. Two thresholds of token scores were tested during the training set submission and the best scoring system was used for final test set submission (Table 1).

Table 1: Results from evaluation on training and test data submissions

	Micro precision		Micro recall		Micro F-score	
Threshold	Training	Test	Training	Test	Training	Test
0.95	0.5379	0.5293	0.6783	0.6726	0.6000	0.5924
1.00	0.5835	-	0.5851	-	0.5843	-

The system described here for the CEMP task relied heavily on dictionary for matching, tokenization as well as rule identification. Although preliminary in nature, the performance of ChemGrab highlights several areas for improvement. The tokenization step implemented in this study relied on direct matches to characters surrounding whitespace, which may be a limiting factor. This step could be generalized by identifying patterns instead from the token windows. Future work will aim at identifying corpus specific rules to address the problem of whitespace separation.

Improvement in performance may also be achieved by using the training corpus to learn chemical-specific tokenization rules in conjunction with negative dictionary. The scoring of word level features described here may possibly be enhanced in future by incorporating the chemical ontology structure. Such a scoring system may result in imparting higher weights to specific elements from specific chemical groups.

Additional work is required in post-processing step of combining chemical tokens to expand over multi-token mention of chemical entities. This step would involve checking balanced parenthesis, square brackets, curly brackets, punctuations, and other non-word characters, which are commonly found in chemical names. The system developed for this study does not include tagging of acronyms, addition of which may improve the recall. Future work is expected to include using corpus specific contextual features. The system is available as a REST-compliant web service (`http://bcbi.brown.edu/chemgrab`).

4 Conclusion

The identification of chemical entities from biomedical literature and patents can help guide effective management and retrieval of relevant information that offer potential to guide future investigation. Here, a chemical named entity recognition approach was developed, ChemGrab, which relies on dictionary and language lexicon for look-up, matching and feature identification. The promising evaluation results of ChemGrab, relative to the CEMP reference data set from BioCreative V.5, suggest that it may serve as a foundation for automating identification of chemical mentions in text.

5 Acknowledgement

This work was supported by the United States National Library of Medicine of the United States National Institutes of Health under Award Number R01LM011963.

REFERENCES

1. 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, e107477 (2014).
2. Clark, K., Cavicchi, J., Jensen, K., Fitzgerald, R., Bennett, A., Kowalski, S.P.: Patent data mining: a tool for accelerating HIV vaccine innovation. Vaccine. 29, 4086–4093 (2011).
3. Tyrchan, C., Boström, J., Giordanetto, F., Winter, J., Muresan, S.: Exploiting structural information in patent specifications for key compound prediction. J. Chem. Inf. Model. 52, 1480–1489 (2012).
4. Krallinger, M., Rabal, O., Lourenço, A., Perez, M.P., Rodriguez, G.P., Vazquez, M., Leitner, F., Oyarzabal, J., Valencia, A.: Overview of the CHEMDNER patents task. In: Proceedings of the fifth BioCreative challenge evaluation workshop. pp. 63–75 (2015).
5. Kiss, M., Nagy, Á., Vincze, V., Almási, A., Alexin, Z., Csirik, J.: A Manually Annotated Corpus of Pharmaceutical Patents. In: Text, Speech and Dialogue. pp. 135–142. Springer, Berlin, Heidelberg (2012).
6. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics. (2003).

7. Kolárik, C., Klinger, R., Friedrich, C.M., Hofmann-Apitius, M., Fluck, J.: Chemical names: terminological resources and corpora annotation. In: Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference) (2008).
8. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: Overview of the chemical compound and drug name recognition (CHEMDNER) task. In: BioCreative challenge evaluation workshop. p. 2 (2013).
9. Erhardt, R.A.-A., Schneider, R., Blaschke, C.: Status of text-mining techniques applied to biomedical text. Drug Discov. Today. 11, 315–325 (2006).
10. Schuemie, M.J., Jelier, R., Kors, J.A.: Peregrine: Lightweight gene name normalization by dictionary lookup. In: Proceedings of the Biocreative 2 workshop (2007).
11. Koning, D., Sarkar, I.N., Moritz, T.: Taxongrab: Extracting Taxonomic Names from Text. (2008).
12. Gerner, M., Nenadic, G., Bergman, C.M.: LINNAEUS: a species name identification system for biomedical literature. BMC Bioinformatics. 11, 85 (2010).
13. Rocktäschel, T., Weidlich, M., Leser, U.: ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics. 28, 1633–1640 (2012).
14. Eltyeb, S., Salim, N.: Chemical named entities recognition: a review on approaches and applications. J. Cheminform. 6, 17 (2014).
15. Lowe, D.M., Sayle, R.A.: LeadMine: a grammar and dictionary driven approach to entity recognition. J. Cheminform. 7, S5 (2015).
16. Leaman, R., Wei, C.-H., Lu, Z.: tmChem: a high performance approach for chemical named entity recognition and normalization. J. Cheminform. 7, S3 (2015).
17. Hettne, K.M., Stierum, R.H., Schuemie, M.J., Hendriksen, P.J.M., Schijvenaars, B.J.A., Mulligen, E.M. van, Kleinjans, J., Kors, J.A.: A dictionary to identify small molecules and drugs in free text. Bioinformatics. 25, 2983–2991 (2009).
18. Miller, G.A.: WordNet: A Lexical Database for English. Commun. ACM. 38, 39–41 (1995).
19. Browne, A.C., McCray, A.T., Srinivasan, S.: The specialist lexicon. National Library of Medicine Technical Reports. 18–21 (2000).