A dictionary- and grammar-based chemical named entity recognizer

Saber A. Akhondi[1], Bharat Singh[1], Eelke van der Host[2], Erik van Mulligen[1], Kristina M. Hettne[2], Jan A. Kors[1]

1-Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherland
2-Department of Human Genetics, Leiden University Medical Center, The Netherlands

[1]{s.ahmadakhondi,b.singh,e.vanmulligen,j.kors}@erasmusmc.nl;
[2]{e.van_der_horst,k.m.hettne}@lumc.nl;

Abstract.
The past decade has seen a massive increase in the number of chemical-related publications. Automatic identification and extraction of compounds and drugs mentioned in these publications can greatly benefit drug discovery research. The BioCreative CHEMDNER task focuses on recognizing and ranking mentions of these compounds in text (CDI) and extracting mention locations (CEM). We investigated an ensemble approach where dictionary-based named entity recognition is used along with grammar-based recognizers to extract compounds from text. Using an open source indexing system, we assessed the performance of ten different commercial and publicly available lexical resources in combination with three different chemical compound recognizers. The best combination along with a set of regular expressions was used to extract the compounds. To rank the different compounds found in a text, a normalized ratio of frequency of mention of chemical terms in chemical and non-chemical journals was calculated. When tested on the training data, our final system obtained an F-score of 88.5% for the CDI task, and 81.0% for the CEM task.

Keywords: Named entity recognition, Molecular structure, Chemical databases, Chemical identifiers

1 Introduction

The past decade has seen a massive increase in the number of chemical -related publications in scientific journals and patents. With the increase in the amount of available literature, it becomes harder to find interesting, relevant and novel information from these unstructured texts [1]. The ability to automatically index every individual publication with the chemical entities mentioned in them, can ease finding of relevant information. Ranking these chemical entities based on accuracy can also increase the certainty of the relevance of the publication. Also, identifying and extracting the location of every mention of chemical compounds in each of these publications can later be used to establish relationship with other entities or concepts [2].

Different text-mining approaches can be taken to extract chemical named entities from text. Approaches have previously been categorized as dictionary-based, morphology-based, and context-based [2]. In dictionary-based approaches, different matching methods are used to lookup matches of the dictionary terms in the text [2]. This approach requires well-defined and good-quality dictionaries. The dictionaries are usually produced from well-known chemical databases. This approach may well capture non-systematic chemical identifiers, such as brand or generic drug names, which are source dependent and are generated at the point of registration. The drawback of a dictionary approach is that it is nearly impossible to include all systematic chemical identifiers, such as IUPAC names [3] or SMILES [4], which are algorithmically generated, based on compound structure, and follow a specific grammar [5]. Instead, morphology- or grammar-based approaches try to capture systematic terms by utilizing the particular grammar, for example through finite state machines [6]. Both of these approaches may suffer from tokenization problems [2]. On the other hand, context-aware systems use machine learning techniques and natural language processing (NLP) to capture chemical entities. The drawback of machine learning approaches is the need of a gold standard for training the system.

The BioCreative CHEMDNER task [7] aims at encouraging the implementation of systems that can index chemical entities (especially the ones that are associated with a structure) in scientific journals. Participants were invited to submit results for two different tasks. The chemical document indexing sub-task (CDI) focuses on the extraction

of a ranked list of chemical entities occurring in each of a set of documents [7]. The chemical entity mention recognition sub-task (CEM) aims at providing the location of every mentioned chemical entity within a document [7]. For this the CHEMDNER organizers have provided participants with a manually annotated gold standard corpus.

2 Methods & Discussion

We investigated an ensemble-based approach where dictionary-based named entity recognition is used along with grammar-based recognizers and chemical toolkits to extract compounds from text. Using Peregrine, an open source indexing system [8, 9], we analyzed the performance of ten different commercial and publicly available lexical resources along with three different chemical compound recognizers. This was done with different indexing settings using different tokenizers (for example case sensitive and case insensitive matching). The best combination along with a set of self-defined regular expressions was used to extract the compounds.

For the dictionary-based approach we tried to focus on compounds that are associated with a structure. We only extracted information from databases with compound records that had MOL files [10]. The following databases were used: Chemical Entities of Biological Interest (ChEBI) three-star compounds [11], ChEMBL [12], Chemspider [13], DrugBank [14], Human Metabolome Database (HMDB) [15], NIH Chemical Genomics Center Pharmaceutical Collection (NPC) [16], Therapeutic Target Database (TTD) [17], and a subset of PubChem [18] compounds likely to have structure-activity relationships and/or other biological annotations [19] with all of their corresponding synonyms derived from PubChem substances. The extracted information contained brand names, synonyms, trade names, generic names, research code, Chemical Abstracts Service (CAS) numbers, etc.

In addition, we used the complete Jochem joined lexical resource [20] and a subset of chemical-related semantic types from UMLS [21], although these resources do not contain MOL files. To capture family names, we also created a dictionary from the ChEBI ontology where we only took parent compounds that did not appear in the ChEBI three-star

database, assuming that these terms have a high likelihood of being a family name. We call this dictionary ChEBI family.

Our preliminary approach was to extract non-systematic and systematic chemical identifiers using both dictionary- and grammar-based approaches, family names using the ChEBI family dictionary, and database identifiers using a set of manually defined regular expressions.

Using the Peregrine tagger all the terms from the mentioned resources were used to index the training and the development sets. This was done in separate runs with different settings: case insensitive, case sensitive, and only case-sensitive for abbreviations (defined as terms where the majority of characters consists of capitals or digits). Assuming chemical compounds will mostly be present in noun phrases of a sentence, the above experiments were also repeated by only feeding noun phrases extracted with the OpenNLP chunker [22] to Peregrine.

We also used a number of public and commercial software packages that can find chemical entities in text: NextMove's LeadMine [23], ChemAxon's Document to Structure toolkit [24], and OSCAR 4 [25]. These tools have implemented grammar-based recognition of systematic chemical identifiers.

The following stop words were used for all of the mentioned experiments: 100 English basic words [26], PubMed stop word list [27], Jochem stop word list [20], and stop words from the CHEMDNER guideline [7].

We used the BioCreative evaluation library script [28] to calculate precision, recall, and F-score. Based on the obtained scores on the training and development sets, we decided on the best ensemble system. Our results showed that a combination of ChEBI three-star compounds and HMDB (in case-sensitive mode) for the dictionary approach along with LeadMine and the self-defined regular expressions performed best. For the CDI task, the micro-averaged F-score was 66.7%, with 70.0% precision and 63.6% recall; for the CEM task, the F-score was 62.2% with 66.3% precision and 58.6% recall. Including the subset of PubChem to this set improved recall with around 9%, while decreasing precision by about 10%, yielding a small drop in F-score. The use of the stop words list greatly improved overall performance, whereas the use of noun phrases did not.

In the final setup we tried to further improve our systems by extending our dictionary with all gold-standard annotations from the training and development sets that our systems initially missed. With this addition, the best system (combination of CHEBI three-star, HMDB, LeadMine and regular expressions) reached 81.2% F-score, 73.4% precision, and 90.1% recall for the CDI task; these values were 75.0%, 68.5%, and 82.9%, respectively, for the CEM task. Furthermore, we added all false-positive terms that were never annotated by the annotators of the training and development sets, to our stop word list. This further improved performance to 88.5% F-score (87.6% precision and 89.4% recall) in the CDI task, and 81.0% F-score (80.9% precision and 81.1% recall) for the CEM task.

For the CDI ranking task, PubMed abstracts were divided into three subgroups based on subject categories from the ISI Web of Knowledge [29]. The first group consisted of abstracts from chemical journals, using the same categories as described in the CHEMDNER guidelines [7]. The second group contained abstracts from non-chemical journals (e.g., “Agricultural economics & policy” related journals). The final group contained the remaining abstracts. The first two groups were indexed using Peregrine and all vocabularies. We assumed that chemical terms should be present more frequently in chemical abstracts than in non-chemical abstracts. Based on the frequency of indexed terms, a normalized ratio was calculated between zero and one. Whenever available this normalized ratio was used for ranking a term. If the ratio was not available (for terms not contained in the vocabularies), we used the precision of the capturing system for that term. A term with high ratio is found more frequently in chemical abstracts than in non-chemical abstracts and therefore is likely to be a chemical term. Vice versa, a term with low ratio is likely to be non-chemical, or highly ambiguous. This approach may also be used to detect ambiguous or erroneous terms in chemical databases.
Finally the following five runs were submitted for both the CDI and the CEM subtasks:
Run 1 - system with the best F-score (case-sensitive matching)
Run 2 - system with lower F-score, but higher recall (case-sensitive matching)
Run 3 - system with the best F-score (partially case-sensitive matching)
Run 4 - system with lower F-score, but higher recall (partially case-sensitive matching)

Run 5 - system with the best F-score (case-sensitive matching), with an extended stop words list based on the ratio previously discussed. In this run whenever an overlap was seen between indexed terms the longer term was chosen.

Finally it should be noted that our system can provide structures for all found terms with the dictionary-based approach, as only terms with MOL files were included from ChEBI, HMDB, and subset of PubChem. Also the LeadMine tool can provide structure for extracted terms. Only the missed annotated terms in our training data, which were added to our vocabulary to improve the performance of the system, are not linked to structure information.

3 Acknowledgment

This study was made possible by a grant provided by AstraZeneca. The authors would like to acknowledge NextMove Software for providing access to Leadmine. We also would like to thank ChemAxon for providing us license to their cheminformatics software. Finally we would like to acknowledge the Royal Society of Chemistry for making Chemspider available to us for research purposes.

REFERENCES

1. Yeh A, Morgan A, Colosimo M, Hirschman L. “BioCreAtIvE task 1A: gene mention finding evaluation.” BMC Bioinformatics 2005, 6 Suppl 1:S2.
2. Miguel Vazquez, Martin Krallinger, Florian Leitner, Valencia A. “Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications.” Molecular Informatics 2011, 30(6-7): 506-519.
3. About IUPAC, “http://www.iupac.org/home/about.html”.
4. Weininger D. “SMILES, a chemical language and information system.1.Introduction to methodology and encoding rules.” J Chem Inf Comput Sci 1988, 28:31-36.
5. Akhondi SA, Kors JA, Muresan S. “Consistency of systematic chemical identifiers within and between small-molecule databases.” J Cheminform 2012, 4(1):35.
6. Sayle R, Xie PH, Muresan S. “Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction.” J Chem Inf Model 2012, 52(1):51-62.
7. Krallinger M, Rabal O, Leitner F, Vazquez M, Oyarzabal J, Valencia A. “Overview of the chemical compound and drug name recognition (CHEMDNER) task.” Proceedings of the fourth BioCreative challenge evaluation workshop 2013, 2.
8. Peregrine, “https://trac.nbic.nl/data-mining/"
9. Schuemie MJ, Jelier R, Kors JA. “Peregrine: Lightweight gene name normalization by dictionary lookup.” Proceedings of the Biocreative 2 workshop 2007 April 23-25(Madrid):131-140.

10. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, J L. “Description of several chemical structure file formats used by computer programs developed at molecular design limited.” J Chem Inf Comput Sci 1992(32):244-255.
11. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C “Chemical Entities of Biological Interest: an update.” Nucleic Acids Res 2010, 38(Database issue):D249-254.
12. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B et al. “ChEMBL: a large-scale bioactivity database for drug discovery.” Nucleic Acids Res 2012, 40(Database issue):D1100-1107.
13. Pence HE, Williams A. “ChemSpider: An Online Chemical Information Resource.” Journal of Chemical Education 2010, 87(11):1123-1124.
14. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V et al “DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.” Nucleic Acids Res 2011, 39(Database issue):D1035-1041.
15. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S et al “HMDB: a knowledgebase for the human metabolome.” Nucleic Acids Res 2009, 37(Database issue):D603-610.
16. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, Nguyen DT, Austin CP. “The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics.” Sci Transl Med 2011, 3(80):80ps16.
17. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C et al. “Update of TTD: Therapeutic Target Database.” Nucleic Acids Res 2010, 38(Database issue):D787-791.
18. Bolton E, Wang Y, Thiessen P, S B. “PubChem: integrated platform of small molecules and biological activities.” Annual reports in computational chemistry 2008, 12th edition, Washington, DC: American Chemical Society.
19. Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, Tyrchan C, Varkonyi P, Xie PH. “Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data.” Drug Discov Today 2011, 16(23-24):1019-1030.
20. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Mulligen EM, Kleinjans J, Kors JA. “A dictionary to identify small molecules and drugs in free text.” Bioinformatics 2009, 25(22):2983-2991.
21. Bodenreider O. “The Unified Medical Language System (UMLS): integrating biomedical terminology.” Nucleic Acids Res 2004, 32(Database issue):D267-270.
22. Apache OpenNLP library, “http://opennlp.apache.org/”.
23. NextMove Softwar- LeadMine, “http://www.nextmovesoftware.com/products/LeadMine.html”.
24. ChemAxon- Document to Structure, “http://www.chemaxon.com/products/document-to-structure/”.
25. Jessop DM, Adams SE, Willighagen EL, Hawizy L, Murray-Rust P. “OSCAR4: a flexible architecture for chemical text-mining.” J Cheminform 2011, 3(1):41.
26. 100 English basic words, “http://en.wiktionary.org/wiki/Category:100_English_basic_words”.
27. PubMed Stopwords list, “http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_170.html”.

28. BioCreative evaluation library scripts, “http://www.biocreative.org/resources/biocreative-ii5/evaluation-library/ ”.
29. Web of Knowledge, “http://webofknowledge.com/”.