Patent mining: combining dictionary-based and machine-learning approaches

Saber A. Akhondi*[1], Ewoud Pons[1], Zubair Afzal[1], Herman van Haagen[1], Benedikt Becker[1], Kristina M. Hettne[2], Erik M. van Mulligen[1], Jan A. Kors[1]

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

[1]{s.ahmadakhondi, e.pons, m.afzal, h.vanhaagen, b.becker, e.vanmulligen, j.kors}@erasmusmc.nl
[2]{k.m.hettne}@lumc.nl

Abstract. Exploration of the chemical patent space is essential for early-stage medicinal chemistry activities. The BioCreative CHEMDNER-patents task focuses on the recognition of chemical compounds in patents. This includes recognition of chemical named entities in patents (CEMP), classification of chemical-related patent titles and abstracts (CPD), and recognition of genes and proteins in patent abstracts (GPRO). In this study we tackled the CEMP and CPD tasks. We investigated an ensemble system where a dictionary-based approach is combined with a machine-learning approach to extract compounds from text. For this the performance of several lexical resources was assessed using Peregrine, our open source indexing engine. We combined our dictionary-based results on the patent corpus with the results of tmChem, a CRF-based chemical recognizer. To improve the performance of tmChem, three additional feature types where introduced (POS tags, lemmas, and word-vector clusters). When evaluated on the training data, our final system obtained an F-score of 85.21% for the CEMP task, and an accuracy of 91.53% for the CPD task. On the test set, our system ranked sixth among 21 teams for CEMP with an F-score of 86.82%, and second for CPD with an accuracy of 94.23%.

Keywords. Patent mining; Named entity recognition; Chemical databases; Conditional random fields; Word vectors; Text mining; CHEMDNER-Patents; CEMP; CPD

1 Introduction

Exploration of the chemical and biological space covered by patents is essential for early-stage medicinal chemistry activities [1]. Analyzing

patents can help understand compound prior art, and lead to identification of alternative starting points for chemical research. Typically, patent information is manually extracted [2]. This process is time-consuming and expensive due to the complexity of chemical patents (e.g., large size, unstructured, spelling mistakes, and OCR text). Automatic approaches can help to ease this process, but have proven to be complex and challenging [3]. One of the issues is that they require a gold-standard corpus for algorithm training and evaluation [4].

The CHEMDNER-patents track 2 challenge in BioCreative V [5] focuses on the extraction of chemical and biological data from medicinal chemistry patents, and consists of three tasks: CEMP (Chemical Entity Mention in Patents), focusing on chemical named entity recognition in patents; CPD (Chemical Passage Detection), focusing on the classification of patent titles and abstracts according to whether they contain chemical entities; and GPRO (Gene and Protein Related Object), focusing on the recognition of gene and protein mentions in patents.

The Erasmus MC team participated in the CEMP and CPD tasks. For the CEMP task we used an ensemble approach where dictionary-based approaches are combined with a machine-learning approach. For the dictionary-based approach we used Peregrine, our open-source indexing software [6] along with seven lexical resources. For the machine-learning approach we used the tmChem chemical recognizer system [7], which uses a conditional random field (CRF) classifier to extract chemical named entities. We improved the performance of tmChem by including additional features such as lemma, POS tags and word clusters [8]. We applied the same system in the CPD task to classify chemical-related titles and abstracts.

2 Methods

Data

The corpus that the task organizers made available for training consisted of 14,000 manually annotated patent titles and abstracts, divided into a training and development set of 7,000 patents each [5]. The final test set contained 40,000 titles and abstracts of which only 7,000 were annotated. During system development, only the annotations of the training and development sets were made available to the challenge participants. To generate the evaluation results we used the BioCreative evaluation software [9], and focused on micro-averaged recall, preci-

sion, and F-score to assess system performance. We also used the Markyt prediction analysis toolkit [10] to visualize the results.

Pre-processing

The tmChem application could not handle some of the input characters (the tmChem pre-processing step did not remove some of the special characters). Therefore we standardized all input text. For this the input text consisting of title and abstract was converted into byte arrays (using Java libraries) and then converted back into UTF-8.

Dictionary-based approach

Peregrine was used to analyze the performance of different lexical resources. The tool was used with the same settings and tokenizers previously used for the BioCreative CHEMDNER challenge [11].

Lexical resources

We extracted chemical terms from the seven lexical resources: ChEBI, ChEMBL, DrugBank, HMDB, NPC, TTD, and a subset of PubChem containing compounds with structure-activity relationships and/or other biological annotations [1, 11]. Chemical terms were only extracted if structure information was available [12].

Exclusion list

To improve the precision of the dictionary-based approach we expanded our term exclusion list previously defined for chemical named entity recognition [11], with exclusion terms mentioned in the CEMP annotation guidelines. Any term that was in the exclusion list was automatically removed from the output.

Exclusion ratio

To further improve the precision of the dictionary-based approach, for each term in the training set the ratio of true-positive and false-positive detections was calculated. Any recognized term in the development set with a ratio lower than 0.3 was disregarded. If the ratio was not available (the term was not seen in the training set), the term was not excluded. Before processing the test set, ratios were calculated for all terms in the combined training and development sets.

Machine-learning approach

We used the tmChem chemical recognizer system, the best performing system in the previous BioCreative chemical named entity recognition challenge [13]. The tmChem system is an ensemble system that combines the output of two CRF-based systems. The first system is a modified version of BANNER, the second approach is based on the tmVAR system [7]. Previous results of tmChem showed that the performance of the second system outperformed the first and the ensemble system. Therefore we only used the second system, which employs CRF++ libraries [14].

Features

Our initial feature set consisted of all features extracted by tmChem, including stemming, word morphology, prefixes and suffixes, character counts (digit, uppercase, lowercase), semantic affixes (such as trivial rings), and chemical elements.

Three additional types of features were calculated and used with tmChem:

(a) POS tags and lemmatization features

We determined POS tags using the MaxentTagger [15] and lemmas using BioLemmatizer [16]. For this, the training and development sets were converted to BioC, an XML format for document annotation [17], and a pipeline with BioC-compliant tools was set up to generate the POS tags and lemmas.

(b) Word embedding cluster features

Recent studies have shown that features based on clusters of word vectors can improve classification performance [8, 18]. We used the word2vec tool [19] to generate word vectors. The tool uses *K*-means clustering to generate clusters of the word vectors. We used the number of the cluster to which a word belonged as a feature.

We generated separate word clusters during the training and test phase of the challenge. The training clusters were generated from the 14,000 titles and abstracts in the training and development sets. We extended these data with 200 full chemical patents used in a previous study [4]. We experimented with different values of *K* (300, 500, 1000). The clusters for the test set were generated using all 40,000 abstracts plus the 200 full patents with $K = 1000$.

Post-processing

We applied different post-processing steps. For the dictionary-based approach, we identified all missed terms (false negatives) and re-indexed the text for these terms. Only terms with an exclusion ratio of 0.5 were accepted. For the machine-learning approach, the tmChem post-processing steps were used [7].
The best combination of dictionary-based approach and machine-learning approach was used for the final ensemble system.

Text classification

For the CPD classification task we used the output of the CEMP task. If a chemical term was recognized in the title or abstract, it was categorized as chemical-related text.

3 Results

Table 1 shows the performance of the dictionary-based approach on the development set. Use of the exclusion list gives a substantial precision improvement for most dictionaries. PubChem provides the highest recall and F-score among the individual lexical resources, while ChEMBL provides the highest precision. The table also shows the performance of several combinations of lexical resources. We previously used ChEBI-HMDB in BioCreative 4 [11]. In this study, we selected ChEMBL-DrugBank, the combination with the highest precision.

Table 1: Performance of different dictionaries and dictionary combinations with and without removal of exclusion terms.

	Without exclusion			With exclusion		
Dictionary	**P**	**R**	**F**	**P**	**R**	**F**
ChEBI	56.51	29.47	38.74	78.87	28.42	41.79
ChEMBL	84.53	20.46	32.94	85.11	19.87	32.22
DrugBank	68.20	17.28	27.58	85.15	16.89	28.19
HMDB	66.11	29.38	40.68	79.59	28.19	41.63
NPC	30.90	44.85	36.59	55.23	30.61	39.39
TTD	66.89	14.07	23.24	80.90	13.89	23.71
PubChem	34.30	47.11	39.69	67.03	45.64	54.30
Combined	30.85	50.32	38.25	53.66	48.59	51.00
ChEBI-HMDB	55.46	36.98	44.37	78.12	35.45	48.77
ChEMBL-DrugBank	70.51	23.94	35.74	83.02	23.16	36.21

Table 2 shows the performance of the ensemble system trained on the training corpus and evaluated on the development corpus. We only present dictionary-based results for the combination of ChEMBL and DrugBank as they provided the highest F-score on the training data when combined with the CRF. The best ensemble system obtained an F-score of 85.21% with a precision of 84.88% and a recall of 85.55%.

Table 2: Performance of the ensemble system trained on the training set and tested on the development set.

System	P	R	F
Dictionary based (ChEMBL-DrugBank)	70.51	23.94	35.74
+ Exclusion list	83.02	23.16	36.21
+ Term removal (exclusion ratio 0.3)	88.85	23.09	36.65
+ CRF original features	84.96	83.83	84.39
+ Post-processing (CRF)	84.50	84.91	84.70
+ POS + lemmatization features	84.72	85.09	84.90
+ Word cluster features	84.88	85.55	**85.21**
+ Missed terms (exclusion ratio 0.5)	75.88	88.63	81.76

Using the best ensemble system, for the CPD task we obtained a sensitivity of 95.31%, a specificity of 84.87%, an accuracy of 91.53%, and a MCC of 81.51%.
On the test set, among the 21 teams that submitted results, our system obtained the sixth place for CEMP (F-score 86.52%) and the second place for CPD (accuracy 94.23%, MCC 87.03%).

4 Discussion

Our final system for the CEMP task obtained an F-score of 85.21% on the development set, with balanced precision and recall. This yielded an accuracy of 91.53% on the CPD task. The decision to include machine learning approaches was made based on the high performance of CRF systems in the BioCreative CHEMDNER challenge [13].

The recall of our lexical resources is low, and even a combination of all dictionaries gives a recall and precision of only 50%. The low recall can be explained because the majority of systematic chemical terms are not present in lexical resources. Meanwhile the machine-learning approach provided a much higher precision and recall (86% and 81%, respectively). In order to maintain the high precision of the ensemble

system, we used the lexical resources the highest precision (ChEMBL and DrugBank). This provided us with a system that improves the machine-learning performance, albeit slightly.

Although providing structure information about the recognized chemicals was not part of the challenge, this information is often important in practical applications. We can readily associate dictionary terms with structures since we limited our lexical resources to chemical records with structures (which provides structural information for 23% of the recognized chemical terms.) For the machine-learning approach, the mapping of recognized terms to structures is less straightforward, but part of these terms will be systematic chemical identifiers. These can also be converted into chemical structures using chemical naming conversion software [20].

Contrary to our expectation, the inclusion of false-negative terms that were missed in the training set decreased the performance on the development set. This is mainly caused by new terms for which an exclusion ratio could not be computed because they have not been found before due to tokenization issues. We included the missing terms in two of our rounds. This was done based on the assumption that the exclusion ratio list will improve when trained on training and development set.

We plan to make the system available as a web service, accessible through our website [www.biosemantics.org]. Peregrine is available for download from the same location. The ontologies and the word vector clusters can be made available upon request.

5 Acknowledgment

This study was made possible by a grant provided by AstraZeneca.

References

1. 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:**1019-1030.
2. Zimmermann M, Fluck J, Thi le TB, Kolarik C, Kumpf K, Hofmann M: **Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology.** *Curr Top Med Chem* 2005, **5:**785-796.
3. Jessop DM, Adams SE, Murray-Rust P: **Mining chemical information from open patents.** *J Cheminform* 2011, **3:**40.

4. Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, Zimmermann M, Jagarlapudi SA, Sayle R, Kors JA, Muresan S: **Annotated chemical patent corpus: a gold standard for text mining.** *PLoS One* 2014, **9:**e107477.
5. **CHEMDNER-patents** [http://www.biocreative.org/tasks/biocreative-v/track-2-chemdner/]
6. Schuemie MJ, Jelier R, Kors JA: **Peregrine: Lightweight gene name normalization by dictionary lookup.** In *Proc of the Second BioCreative Challenge Evaluation Workshop*. 2007: 131-133.
7. Leaman R, Wei CH, Lu Z: **tmChem: a high performance approach for chemical named entity recognition and normalization.** *J Cheminform* 2015, **7:**S3.
8. Deng L, Yu D: **Deep learning: methods and applications.** *Foundations and Trends in Signal Processing* 2014, **7:**197-387.
9. **BioCreative Evaluation Software** [http://www.biocreative.org/resources/biocreative-ii5/evaluation-library/]
10. **Markyt prediction analysis toolkit** [http://www.markyt.org/biocreative/analysis]
11. Akhondi SA, Hettne KM, van der Horst E, van Mulligen EM, Kors JA: **Recognition of chemical entities: combining dictionary-based and grammar-based approaches.** *J Cheminform* 2015, **7:**S10.
12. Dalby A, Nourse JG, Hounshell WD, Gushurst AK, Grier DL, Leland BA, Laufer J: **Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited.** *Journal of chemical information and computer sciences* 1992, **32:**244-255.
13. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, Valencia A: **CHEMDNER: The drugs and chemical names extraction challenge.** *J Cheminform* 2015, **7:**S1.
14. **CRF++: Yet Another CRF toolkit** [https://taku910.github.io/crfpp/]
15. Toutanova K, Klein D, Manning CD, Singer Y: **Feature-rich part-of-speech tagging with a cyclic dependency network.** *Hlt-Naacl 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference* 2003**:**252-259.
16. Liu H, Christiansen T, Baumgartner WA, Jr., Verspoor K: **BioLemmatizer: a lemmatization tool for morphological processing of biomedical text.** *J Biomed Semantics* 2012, **3:**3.
17. Comeau DC, Islamaj Dogan R, Ciccarese P, Cohen KB, Krallinger M, Leitner F, Lu Z, Peng Y, Rinaldi F, Torii M, et al: **BioC: a minimalist approach to interoperability for biomedical text processing.** *Database (Oxford)* 2013, **2013:**bat064.
18. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G: **Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features.** *J Am Med Inform Assoc* 2015, **22:**671-681.
19. **word2vec** [https://code.google.com/p/word2vec/]
20. Akhondi SA, Kors JA, Muresan S: **Consistency of systematic chemical identifiers within and between small-molecule databases.** *J Cheminform* 2012, **4:**35.