Identification of Chemical Entities in Patent Documents

Tiago Grego[1], Piotr Pęzik[2], Francisco M. Couto[1], and Dietrich Rebholz-Schuhmann[2]

[1] Faculty of Sciences, University of Lisbon, Campo Grande, 1749-016 Lisboa, Portugal
[2] EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
tgrego@fc.ul.pt, pezik@ebi.ac.uk, fcouto@di.fc.ul.pt, rebhozl@ebi.ac.uk

Abstract. Biomedical literature is an important source of information for chemical compounds. However, different representations and nomenclatures for chemical entities exist, which makes the reference of chemical entities ambiguous. Many systems already exist for gene and protein entity recognition, however very few exist for chemical entities. The main reason for this is the lack of corpus to train named entity recognition systems and perform evaluation.

In this paper we present a chemical entity recognizer that uses a machine learning approach based on conditional random fields (CRF) and compare the performance with dictionary-based approaches using several terminological resources. For the training and evaluation, a gold standard of manually curated patent documents was used. While the dictionary-based systems perform well in partial identification of chemical entities, the machine learning approach performs better (10% increase in F-score in comparison to the best dictionary-based system) when identifying complete entities.

Keywords: Chemical Named Entity Recognition, Conditional Random Fields, Text Mining.

1 Introduction

Every year life sciences produce a huge amount of new publications, including patent registration documents. The tasks of document categorization, recognizing entities and related information, is becoming increasingly important in the daily research and work life of scientists, but it is a challenge. The performance of state of the art text mining tools that recognize gene and proteins is still far from achieving the levels of performance achieved by text mining tools applied to other areas such as news text [1, 2, 3]. For gene and protein entity recognition, some competitions such as the BioCreAtIvE [4] provide an evaluation of state-of-the-art systems for text mining publicly available data sources. Two main approaches are being used by those systems: dictionary-based and dictionary-independent approaches.

Dictionary-based approaches allow a direct mapping of the recognized entities to reference objects (i.e. identifiers for public databases). However this approach has the drawback of being dependent on the completeness and quality of the dictionary and the methods to resolve the high ambiguity and spelling variants inherent to biomedical entities. The other approach is dictionary-independent, and includes rule-based as

S. Omatu et al. (Eds.): IWANN 2009, Part II, LNCS 5518, pp. 942–949, 2009.

well as case-based systems. This approach is better suited to find named entities when no comprehensive dictionary is available.

Machine-learning approaches are based on an annotated training set from which statistical information can be obtained about the inherent dependencies in the data. This extracted information is used to create a probabilistic model that can be applied on unseen data to perform the named entity recognition task. The main bottleneck of this approach is the selection and creation of a training set large enough to enable the creation of a model sufficiently accurate. In the BioCreAtIvE sub-task of gene mention recognition, the best approaches achieved an F-measure of 86% [5].

However, gene and protein mentions are not the only important entities in the biomedical field. Other chemical substances [6] such as drugs, metabolites, nutrients, enzyme cofactors and experimental reagents are also relevant. The annotation of chemical entities enables a number of applications, such as the creation of semantically enhanced articles, with benefits for the readers. The entities found can be linked to their properties, reactions and applications, and co-occurrence with other entities can reveal new relations between chemical data and other bioinformatics data.

We present a chemical entity recognition system capable of finding a wide variety of chemical molecules, classes of chemicals, ligands and chemical formulas of biological interest. A corpus of patent documents was used for the evaluation. Dictionary-based approaches were used to have a baseline evaluation and compared to a machine-learning approach using CRFs.

2 Related Work

Recently, some chemical named entity systems were developed that we describe as follows: ProMiner [7], a dictionary-based system that uses DrugBank for recognition of drug names in MEDLINE; EBIMed uses the drug dictionary from MedlinePlus as source [8]; Narayanaswamy *et al* [9] describes a system based on a manually developed set of rules that rely heavily upon some crucial lexical information, linguistic constraints of English, and contextual information; Kemp and Lynch [10] proposes a system that identifies chemical names in patent texts with handcrafted rules using dictionaries with chemical name fragments; OSCAR3 [11] relies on an internal lexicon of chemical names and structures initially populated using ChEBI [12]. OSCAR3's performance was evaluated on different corpora with F-score rates between 60-80%. This open source program is one of the few available to the academic community [13]. Klinger *et al* [14] presents a machine-learning approach based on conditional random fields (CRF), and a performance of 80-85% F-score. However this system is an IUPAC-like [15] named entity recognition system only, and it is usual for a chemical to be referenced by the trivial name or other synonyms.

There are few annotated corpus available to evaluate (and train if needed) chemical named entity recognition systems. The GENIA corpus [16] includes some chemical annotations; however it is a generic corpus for molecular biology and includes many entities besides chemicals. Only a few MEDLINE abstracts in that corpus are recognizable as chemistry abstracts. Thus, the lack of proper corpus leads the authors of most systems to generate their own.

3 Method

3.1 Corpus

A joint team of curators from the ChEBI (Chemical Entities of Biological Interest) and EPO (European Patent Office) have manually annotated a corpus of 40 patent documents (this annotated corpus will be our gold standard). Those documents were selected to be a representative set of the universe of chemical patent documents. EPO is interested in providing such a corpus so that improvements can be made in information retrieval systems applied to the existing documents.

ChEBI is a freely available dictionary of small molecular entities such as any constitutionally or isotopically distinct atom, molecule, ion, radical, etc. In addition to molecular entities, ChEBI contains groups (parts of molecular entities) and classes of entities, enabling ChEBI to be organized as a chemical ontology, structuring molecular entities into subsumption classes and defining the relations between them.

Patent documents describe inventions that are required to be new, involve an inventive step, and are susceptible of industrial application. Because of its innovative nature, patent documents have the potential to be a good source of new chemical entities that can be used to extend the ChEBI ontology.

The corpus of 40 documents contains 4985 sentences and 182258 words. The number of entities annotated in the gold standard was of 11162, which gives an average 280 entities per document.

3.2 Dictionary-Based Approach

The approach followed was a direct search in the documents for the entities present in each of the dictionaries used. This simple approach is easy to implement and can be used as a baseline to have preliminary results and to understand how existing dictionaries cover the entities present in the gold standard.

Three different dictionaries were used as resource: ChEBI, DrugBank and Oscar3.

ChEBI contains approximately 15000 entries, and 40000 synonyms. This number shows the problem of multiple names for one entity (polysemy). For example, adrenaline and epinephrine refer to the same chemical entity. Chemical names, particularly common names, may contain ambiguity as to the exact chemical which is intended by the use of the name. For example the term adrenaline may refer to either one of the enantiomers (*S*)-adrenaline and (*R*)-adrenaline. This accounts for the complexity of identifying chemical entity mentions in text.

DrugBank combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database contains nearly 4800 drug entries [17].

OSCAR3 keeps an internal lexicon of chemical names and structures that have been initially populated using ChEBI, and further extended. This lexicon was also used as a dictionary resource [18].

3.3 Machine-Learning Approach

Conditional random fields (CRF) were used for building probabilistic models to automatically annotate the corpus [19]. CRFs are a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and lattices. In the

context of named entity recognition they can be used to label a sequence of tokens, each token being represented as a set of features

The primary advantage of CRFs over hidden Markov models (HMM) is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models.

The gold standard was tokenized using a general tokenizer, and the sequence of tokens used by MALLET [20], a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text, which includes an implementation of CRFs.

The corpus was divided in a training-set and a testing-set. From the 40 documents, 20 were randomly selected for training, and the remaining used for testing.

To select an adequate set of features, several models were created using different sets of features. Tested features include the token, the stem of the token, lowercased versions of the token and stem, prefix and suffix of the token and indication if the token is a number. For example, for the piece of sentence "...cosmetic compositions containing colostrum, tocopherols, zinc oxide and hyaluronic acid..." (the chemical entities present are underlined) the list of tokens obtained by the tokenizer and some possible features are shown in Table 1. The CRF implementation uses a sequence of sets of such features, plus a label (for the documents in the training set) for the training step. The resulting model can then be used to predict the label of another sequence of features (the testing-set).

With this approach a token can be classified as being part of a chemical entity, or not. In this way it is impossible to identify the boundaries of an entity, and the results are not comparable to the ones given for dictionary-based approaches.

3.4 Evaluation

For evaluation of the dictionary-based systems, the corpus was automatically annotated using the different resources. The obtained annotations were then compared with the ones of the gold standard.

Table 1. Example of a sequence of features, and the corresponding label

Token	Stem	Prefix	Suffix	Label
cosmetic	cosmet	cos	tic	Not Chemical
compositions	composit	com	ons	Not Chemical
containing	contain	con	ing	Not Chemical
colostrum	colostrum	col	rum	Not Chemical
tocopherols	tocopherol	toc	ols	Chemical
zinc	zinc	zin	inc	Chemical
oxide	oxid	oxi	ide	Chemical
and	and	and	and	Not Chemical
hyaluronic	hyaluron	hya	nic	Chemical
acid	acid	aci	cid	Chemical

Many entities are composed by more than one token (for example the entity "hyaluronic acid"), so we make the distinction between partial and a complete match.

When the complete entity is correctly identified by the automatic annotation system, we consider having a complete match (in the example given before, "hyaluronic acid" had to be annotated to be considered a complete match). When only part of an entity is identified (for example "acid" in the example given) we have a partial match. When nothing of an entity is annotated we have a missed annotation, and when a piece of unannotated text in the gold standard is annotated by the system we have an annotation error. We can than obtain precision, recall and F-score measures.

To evaluate the machine-learning approach based on CRF we need a way to measure the complete match results. To be able to do so we changed the number of labels from two ("Not Chemical" and "Chemical") to five ("Not Chemical", "Single", "Start", "End" and "Middle"). This way a token can still be labeled "Not Chemical", but the label "Chemical" was split into four new labels. "Single" identifies a token that is itself a chemical entity (single token entities). For the multi token entities, the first token is labeled as "Start", and the last one to as "End". For entities composed by more than two tokens, the remaining tokens are labeled as "Middle". This increase in the number of classes can put an overhead in the system, slightly decreasing the performance, but is necessary to allow boundary detection.

4 Results

Dictionary-Based Approach. From Table 2 we can see that for the dictionary-based approach the highest F-score (88%) was achieved using OSCAR as resource, immediately followed by ChEBI with only 3% less. Recall is extremely low for DrugBank, which can be explained by the specificity of this resource (only contains drug names). The resource with higher recall is OSCAR, which leads us to conclude that the internal lexicon of OSCAR3 is the most extensive of the three. Precision is higher than 80% for any of the resources used. For complete match we have a different scenario, the highest F-score is achieved using ChEBI (37%) followed by OSCAR (27%). These results show that the dictionary-based systems can identify where a chemical entity is, but fail to completely identify it, having only a partial match on the entity.

The better results achieved when using ChEBI as a resource were not expected given that the lexicon of OSCAR3 is more extensive. However the gold standard was annotated by ChEBI curators which may explain why this resource performs better in the context of this corpus.

The complete identification of an entity is however important, and improvements in the complete match results are highly desirable.

Machine-Learning Approach. The results of annotating the testing set with models using different sets of features, for selection, are presented in Table 3. The best results were obtained by using a feature set composed by the stem of the token, the prefix, the suffix, and the indication if the token is a number or not. This will be the feature set used in the following experiments.

Table 2. Evaluation results for partial and complete match using dictionary-based approach

Resources	Partial Match			Complete Match		
	Precision	Recall	F-Score	Precision	Recall	F-Score
ChEBI	80,9	91,2	85,7	35,9	38,7	37,2
DrugBank	80,1	28,1	41,6	55,8	6,02	10,9
OSCAR	81,7	96,5	88,5	21,6	37,5	27,4

Table 3. Effect of using different feature sets in CRFs. (lc indicates that uppercase characters were set to lowercase)

Feature Set	Precision	Recall	F-Score
lc_token	82,2	51,1	63,1
lc_stem	82,9	52,8	64,5
prefix suffix	80,9	63,7	71,3
lc_stem prefix suffix	82,5	63,2	71,6
stem prefix suffix	85,0	61,5	72,5
stem prefix suffix digit	83,0	64,6	72,7

Table 4. Evaluation results for complete match using condition random field. Results using ChEBI dictionary are shown for comparison

Approach	Precision	Recall	F-Score
ChEBI resource	35,9	38,7	37,2
50/50 split	62,4	43,6	51,3
CRF crossvalidation	58,5	39,5	47,2

The results obtained for complete match using a random split of 20 documents as training-set, and the remaining documents as testing-set (50/50 split) are given in Table 4.

To have results over the complete corpus we used cross-validation. Each one of the 40 documents was classified using a model constructed using the remaining 39 documents as a training-set. Cross-validation is more robust than simple random 50/50 split, and the results obtained this way for complete match in the identification of the chemical entities in the complete corpus are shown in Table 4.

5 Discussion

Dictionary-based approaches perform well in the task of identifying at least part of a chemical entity, having F-scores of 88% using OSCAR3. However many entities are composed by more than one token and in this case the complete identification of the entities lowers their effectiveness. The performance decreases considerably to a best F-score of 37% using the ChEBI resource. Using OSCAR3 the obtained F-score is 27%.

A machine learning-approach using an implementation of conditional random fields was developed. In feature selection and evaluation we conclude that a feature

set composed by the stem, prefix, suffix and the information if the token is a number or not is the best feature set we tested. However many improvements can be accomplished here. The tokenizer that we used is a generic one, and splits many chemical entities into two or more tokens when this could be avoided. This happens because the tokenizer splits text separated by a hyphen into different tokens. An improved tokenizer, designed specifically for chemical entity tokenization can thus improve the sequence of features sets, and the overall results. Also the features prefix and suffix are the first and last three characters of the token, respectively. Improvements can be made to provide better prefixes and suffixes, and thus improve the feature set for better performance of the machine-learning approach.

Despite these bottlenecks, the machine-learning approach outperforms the dictionary-based ones. The machine learning method improves the F-score by 10% in relation to the results using the best dictionary resource. Besides the better performance, the machine-learning method has other advantages: It can identify novel chemical entities in documents, which in turn can be used to extend dictionaries. A dictionary-based approach can only identify molecules already present in the resource.

6 Conclusion

We have presented a chemical entity recognizer that uses a machine learning approach based on CRFs and compared its performance with dictionary-based approaches using several terminological resources.

The dictionary-based systems performed well in partial identification of chemical entities, while the machine learning approach performed better (10% increase in F-score in comparison to the best dictionary-based system) when identifying complete entities.

In the future it would be interesting to map the identified chemical entities to the ChEBI database. This way a user that is looking at a patent document can view information related to a chemical molecule present in the text. Also when no mapping can be made, we have an indication of a potentially novel chemical entity.

It would be also interesting to apply the model learned from the gold standard to other corpus, such as MEDLINE abstracts, to determine if the system can be used as an all-purpose chemical entity recognizer.

Acknowledgments. We thank the ChEBI and EPO curators for providing the gold standard. This work was supported by FCT, through the project PTDC/EIA/67722/2006, the Multiannual Funding Programme, and the PhD grant SFRH/BD/36015/2007.

References

1. Yeh, A., Hirschman, L., Morgan, A.: Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. Bioinformatics 19(1), i331–i339 (2003)
2. Hersh, W., Cohen, A., Roberts, P., Rekapalli, H.: TREC 2006 genomics track overview. In: Proc. of the 15th Text REtrieval Conference (2006)
3. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6, S1 (2005)

4. Hirschman, L., Krallinger, M., Valencia, A.: Proc. of the Second BioCreative Challenge Evaluation Workshop. Centro Nacional de Investigaciones Oncologicas (2007)
5. Smith, L., Tanabe, L., Ando, R., Kuo, C., Chung, I., Hsu, C., Lin, Y., Klinger, R., Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C., Povinelli, R., Vlachos, A., Baumgartner, W., Hunter, L., Carpenter, B., Tsai, R., Dai, H., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, M., Mana-Lopez, A., Mata-Vazquez, J., Wilbur, W.: Overview of BioCreative II gene mention recognition. Genome Biology 9(suppl. 1), S2 (2008)
6. Reyle, U.: Understanding chemical terminology. Terminology 12, 111–126 (2006)
7. Hanisch, D., Fundel, K., Mevissen, H., Zimmer, R., Fluck, J.: ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics 6(suppl. 1), S14 (2005)
8. Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., Stoehr, P.: Ebimed - text crunching to gather facts for proteins from medline. Bioinformatics 23 (2007)
9. Narayanaswamy, M., Ravikumar, K., Vijay-Shanker, K.: A biological named entity recognizer. In: Proc. of the Pacific Symposium on Biocomputing, pp. 427–438 (2003)
10. Kemp, N., Lynch, M.: The extraction of information from the text of chemical patents. 1. identification of specific chemical names. J. Chem. Inf. Comput. Sci. 38, 544–551 (1998)
11. Corbett, P., Murray-Rust, P.: High-throughput identification of chemistry in life science texts. In: Berthold, M.R., Glen, R.C., Fischer, I. (eds.) CompLife 2006. LNCS (LNBI), vol. 4216, pp. 107–118. Springer, Heidelberg (2006)
12. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344–D350 (2008)
13. Corbett, P., Copestake, A.: Cascaded classifiers for confidence-based chemical named entity recognition. BMC Bioinformatics 9(suppl. 11), S4 (2008)
14. Klinger, R., Kolá, C., Fluck, J., Hofmann-Apitius, M., Friedrich, C.: Detection of IUPAC and IUPAC-like chemical names. ISMB 2008. Bioinformatics 24, i268–i276 (2008)
15. International Union of Pure and Applied Chemistry, `http://www.iupac.org`
16. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: GENIA corpus – a semantically annotated corpus for bio-textmining. Bioinformatics 19(suppl. 1), i180–i182 (2003)
17. Wishart, D., Knox, C., Guo, A., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 34, D668–D672 (2006)
18. Corbett, P.: OSCAR3 (Open Source Chemistry Analysis Routines) - software for the semantic annotation of chemistry papers,
`http://sourceforge.net/projects/oscar3-chem`
19. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th ICML, pp. 282–289 (2001)
20. McCallum, A.: MALLET: A Machine Learning for Language Toolkit,
`http://mallet.cs.umass.edu`