Chemical and Biological Entity Recognition System from Patent Documents

Hongchang Lai
Information Technology
Supporting Center,
Institute of Scientific and Technical
Information of China
No. 15 Fuxing Rd,.Haidian Distirct,
Beijing 100038, P.R. China
+86 10 5888 2447
laihc2013@istic.ac.cn

Shuo Xu
Information Technology
Supporting Center,
Institute of Scientific and
Technical Information of China
No. 15 Fuxing Rd,.Haidian Distirct,
Beijing 100038, P.R. China
+86 10 5888 2447
xush@istic.ac.cn

Lijun Zhu
Information Technology
Supporting Center,
Institute of Scientific and
Technical Information of China
No. 15 Fuxing Rd,.Haidian Distirct,
Beijing 100038, P.R. China
+86 10 5888 2447
zhulj@istic.ac.cn

ABSTRACT

It is crucial to explore the chemical and biological space covered by patent documents. In order to recognize chemical and biological entities, a recognition system is developed on the basis of open-source machine learning and natural language processing (NLP) toolkits. The system processing pipeline consists of three major components: pre-processing (sentence detection, tokenization), recognition (conditional random field (CRF) based approach), and post-processing (rule-based approach). The paper introduces each part in detail. Finally, extensive experiments on annotated chemical patent corpus are conducted, and the balanced-F measure is 69.20% with 10-fold cross validation. The results indicate that the performance on patent documents is slightly lower than that of counterpart on paper and news corpus.

Keywords

Conditional Random Field (CRF); Chemical and Biological Entity Recognition; Patent Mining; Cross Validation

1. INTRODUCTION

It is crucial to explore the chemical and biological space covered by patent documents. For example, it can help speed-up the early-stage medicinal chemistry activities [1] [2]. Though patent documents contain many valuable chemical and biological entities, such as chemical compounds, genes, proteins, drug and so on, automatic recognition systems from patent documents are still very limited.

However, as for paper and news documents, many identification approaches are proposed and resulting systems are also developed.

In our opinion, the reasons are two-fold: (a) the annotated patent corpus are not available to public; (b) the patents are complex legal documents which are very difficult to understand. But the situation will be improved continually, since there is an increasing interest on patent mining, such as BIOINFORMATICS [3], BioCreative [4], JNLPBA [5] and iPaMin [6].

An Annotated Chemical Patent Corpus [8] was published by Akhondi, which enables the development of the chemical and biological entity recognition system. Even so, it is still a rather challenging task to automatically recognize chemical and biological entities from non-structural documents, especially patents [7], since patents are complex legal documents that even contain up to hundreds of pages.

In this paper we explore the chemical and biological entity recognition system from patent documents using similar approaches in [11]. Thus, one can see whether it is feasible by just borrowing some methods. The organization of the rest of the article is as follows. Section 2 summarized the overview of the annotated patent corpus. Section 3 introduces the recognition system and the methods we used. Section 4 decrypts the annotated corpus we used and some information of our experiments.

2. DATASETS OVERVIEW

Akhondi et al have produced gold standard chemical patent corpus of which 47 patents have been annotated by at least three annotators. The full-text patents and annotated entities are publicly accessible at www.biosemantics.org.

We analyzed the training and harmonized dataset and found some nested chemical and biological entities in the harmonized set. In our system, CRF++ is adopted for the actual implementation to process the sequence label problem. Since CRF++ cannot identify the nested entities, we just omit the less spanned entities. .There are 1239 entities of the type "OCRERRORSPELL" and "OCRERRORLINE" in the original annotation corpus, however some of them are nested. Finally, we reduced the entities amount from 37,776 down to 37,288, removing 488 nested entities. The harmonized set was produced from the 47 common patents, including a total of fourteen classes, 9857 unique terms and 37,288 annotated terms (see table 1).

Published at Ceur-ws.org
Proceedings of the Second International Workshop on Patent Mining and its Applications (IPAMIN). May 27–28, 2015, Beijing, China.

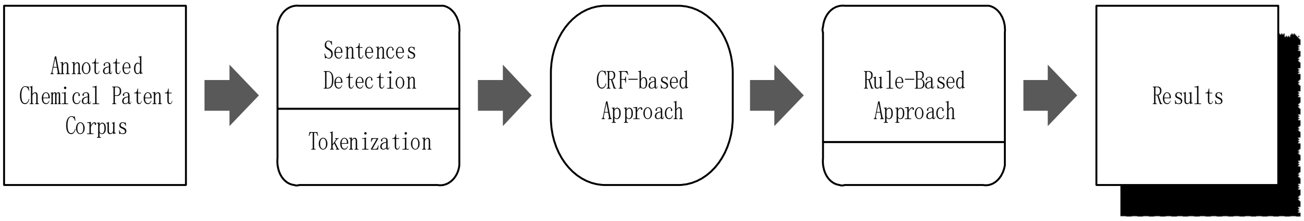

Figure 1 The simplified system processing modules. Pipeline includes three major components: pre-processing (sentence detection, tokenization), recognition (CRF-based approach) and post-processing (rule-based approach)

The results indicate that IUPAC (International Union of Pure and Applied Chemistry) entities and generic names have been annotated obviously more than any other chemical type. On the other hand, InChI (International Chemical Identifier), CAS (Chemical Abstracts Service) registry numbers and SMILES (Simplified molecular input line entry specification) appear rarely in the chemical patents. Since we removed one of the label tag of entities which have two or more tags, the count of results would be a little bit different with [8].

Table 1 Number of annotated terms and unique terms in the harmonized set of the gold standard corpus after removing nested entities

	Description	*Annotated Terms*	*Unique Terms*
M	IUPAC	13943	4592
I	SMILES	20	20
Y	InChi	0	0
D	Trademark	2355	897
B	Abbreviation	2087	146
C	CAS number	6	5
F	Formula	1115	160
R	Registry Number	140	95
G	Generic	8381	811
T	Target	3221	654
Disease	Disease	3765	1205
MOA	MOA	1016	197
OCRERROR-SPELL	Spelling error	1189	1029
OCRERROR-LINE	Spurious line break	50	46
	Total	37288	9857

3. SYSTEMS DESCRIPTION AND METHODS

Based on the summary of the principal methods used in the MUC-4 (Mucin 4) systems, Hobbs proposed a generic information extraction system [10] which consists of ten modules. It is the theoretical basis based on a large amount of practice for our system. On the other hand, we refer to the recognizing chemical entities system published in [11].

The process as showed in Figure 1, our system looks like a serialized pipeline consisted by three major components. At first, annotated chemical patents would be detected sentence boundary. The sentences would be split by tabs ("\t"). And then, each detected sentenced is tokenized as many tokens one by one. Secondly, chemical and biological entity is extracted from corpus with a CRF-based approach. A 10-fold cross validation method is adopted in order to evaluate the effect of our recognition system. Finally, some post-processing steps include a rule-based approach. Each step would be outlined in details in the following subsections.

3.1 Pre-process: sentence detection, and tokenization

There are two kinds of document for each patent in annotated chemical patent corpus, the original patents and the entity annotated for each part. In corpus, each patent was divided into several different partitions. Each partition contains different parts of the patent document. Generally, each subdocument is irregular in each line which is a sentence or not. For example, in the document named *US4659716_0001* of the training_set, some lines are the metadata features of patent such as the title, abstract, inventors and so on; some lines have two or more sentences.

In the system, the openNLP sentence detector toolkit is utilized. Detecting sentence boundary is a challenging work by the reason of the ambiguous punctuation marks. For the further performance of the sentence boundary detection, we gathered the many abbreviations sets of the corpus in advance, such as *var., e.g., sp.* Especially in annotated documents, such as the entity contains the full point marks, for example "EC 3.4.24.11" or "MgCl2.6H2O" etc. Then we generated several rules, for instance if current sentence ends with these abbreviations or comma, the current and subsequent sentences are merged into a new one. And the metadata features in patent as mentioned before, each line is regarded as a sentence because the metadata features are shorter than other sentences and have less information about the entity.

In the end, all the sentences were combined into a bulky document. Each line of the document is a subpart of the patent. The line format is as follows:

fileID sentence. sentence.

Each line begins with the file id of the source of the sentence followed by one tab. Sentences split by space " ".

The tokenizer in the system is based on the OpenNLP toolkit. It can divide the sentence above into some reasonable tokens what we need. However, it would get a poor result by using the original tokenizer, which cannot be applied to sequence labeling problem. Then some improvement approaches are expected to be adopted, and we get much better fine-grained tokens. Such as the entity " (S)-(-)-α,α-diphenyl-2-pyrrolidinemethanol" in *US5650521_ 0003*, the entity type is "M" which means "*IUPAC*".

Table 2 Examples of Chemical component entity labels

token	…	(	S	)	-	(	-
label	*O*	*B*	*I*	*I*	*I*	*I*	*I*
token	)	-	α	,	α	-	dipheny
label	*I*	*I*	*I*	*I*	*I*	*I*	*-I*
token	-	2	-	pyrrolidinem ethanol	….		
label	*I*	*I*	*I*	*E*	*O*		

As shown in Table 2, the punctuation marks (brackets, dashes, etc.), Greek symbols, numbers are regarded as the isolate tokens. In the annotation documents, the type "OCRERRORSPELL" and "OCRERRORLINE" are marked in the end of each document. Meanwhile, the entities of these two types also have the right entity types. Such as in the *US20050222261_0003*:

T109 D 4726 4738 siruvastatin

T343 OCRERRORSPELL 4726 4738 siruvastatin

However, some of OCRERRORSPELL entities have only one type. It means some of them are nested in entity types, but others have a unique type label. For consistency, the uniform type labels are given for each entity to get rid of nested types. There are 1239 entities of the type "OCRERRORSPELL" and "OCRERRORLINE" in the original annotation corpus, however some of them are nested. Finally, we reduce the entities amount from 37,776 down to 37,288, removing 488 nested entities.

3.2 Recognition: crf-based approach

As mentioned above, the chemical and biological entity recognition problem is treated as a sequence label problem (Table 2). Conditional random fields, as a framework for building probabilistic models to segment and label sequence data[13], avoids a fundamental limitation of MEMMs (maximum entropy Markov models) and other discriminative Markov models based on directed graphical models, and offers several advantages over hidden Markov models and stochastic grammars.

CRF can pick up the context into account; e.g., the linear chain CRF in natural language processing predicts sequences of labels for sequences of input samples. There are observations $\mathcal{X}$ and random variables $\mathcal{Y}$, the random variables $\mathcal{Y}$ are conditioned on $\mathcal{X}$. the conditional distribution $p(\mathcal{Y}|\mathcal{X})$is then modeled. Due to some polynomial equations easily computed by Newton's method, the CRF++ adopts the L-BFGS (Limited-memory BFGS (Broyden–Fletcher–Goldfarb–Shanno)) method to do the unconstrained optimization for parameter estimation. On the other hand, CRF++ use line search to compute the step size of the unconstrained optimization problem.

The annotated entity in patent corpus can be classified into one of the fourteen classes:

$$C = \{M, I, Y, D, B, C, F, R, G, T, Disease, MOA, OCRERRORSPELL, OCRERRORLINE\}$$

4-tag method is used to label the chemical entity with B I E O, which means "beginning of the entity", "word in the entity", "end of the entity" and "the other words". And some nested annotated entities mentioned in section 3.1 are uniform to a same type, because the CRF++ cannot process the nested entities. Harmonized set merged by the annotations of the 47 patents annotated by more than three groups is used as the training set with different entity types (chemicals and their sub entities, diseases, MOAs, and targets).

3.3 The features for CRF

Our system exploits four different types of features:

General linguistic features. Our system includes the original tokens, as well as stemmed tokens, as features using the Porter's stemmer from Stanford CoreNLP.

Characteristic features. Since many entities contains numbers, Greek letters, Roman numbers, amino acids, chemical elements, and special characters, our system calculates several statistics as features for each token, including its number of digitals, number of upper- and lower-case letters, number of all characters and presence or absence of specific characters or Greek letters, Roman numbers, amino acids, or chemical elements.

Case pattern features. Similar to [12], the upper case alphabetic character, the lower case one and any number (0-9) are replaced by 'A', 'a', '0' respectively. Moreover, our system also merges consecutive letters and numbers and generated additional single letter 'a' and number '0' features.

Contextual features. For each token, our system includes the linguistic features of two neighboring tokens from each side.

There is an example of the entities features:

Table 3 An example for entity features

Stemmer	Amino Acid	Element	Symbol
Lymphocyte	true	true	false
Roman	Num Of Digitals	Num Of Upper Case Letters	Num Of Lower Case Letters
False	0	0	11
Length	case Pattern	brown	label tag

3.4 Post-processing: rule-based approach

On closer examination, we find that the results of CRF approach include some false positive chemical and biological entities. So, we developed several additional rules to remove them. In addition, our post-processing step also helps adjust text spans of entities, such as adding a missing closing parenthesis.

But we found some false cases in our results:

Such as in the file *EP1481667_0004,* the entity "dopamine receptor" occurs two times but annotated once. In our opinion, it violated the first rule in annotation guideline in paper [8]: When an entity is nested or has an overlap with another entity, the entity should be annotated as more specific and informative.

And in *US20050222261_0004*, "ACE inhibitors" was annotated as two entities. But in *WO2004000294_0004*, it was regarded as the only one. Some entities like "AMcAMP", "IcAMP" (Abbreviation), "amino acids", "agonist", "methane sulfonic acid" were not annotated in some document. "BMS- 204352" and "methyl testosterone" was not annotated in *EP1481667_0004*, but our system recognizes it as an entity. These cases would influence the results to some degree.

4. EXPERIMENTS

The patent corpus is available in 3 different sets: 1-Harmonized_set; 2-Full_set; 3-Training_set. We analyzed the training and harmonized dataset, and found some nested entities in the harmonized set as discussed in section 3.1. Since CRF++ cannot identify the nested entities, we just omit the less spanned entities. Then, we insert the original text of patents and the annotated entities into the mysql database to do the experiments.

Each document is saved as a record in database, the sentences split by space " ". Each term is stored in another table with the classes, offsets, fileID and so on.

The dataset is split for the 10-cross validation, and the training set. Each round contains about 12,000 sentences and 500,000 features.

In CRF++, there are 4 major parameters ("-a", "-c", "-f" and "-p") to control the training condition. CRF++ uses the features "-f" as the cut-off threshold features, that occurs no less than NUM times in the given training data. "-p" is the number of threads. In our submitted predictions, the parameters: "-a", "-f" and "-p" are set to default (CRF-L2), 2 and 4, respectively. The option "-c" trades the balance between over-fitting and under-fitting. The predicted results will significantly be influenced by this parameter. It is better to find an optimal value by cross validation. We just set "-c" option to $\{2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}\}$ due to the constraints of experimental time. Our submitted 5 runs corresponds to different values of "-c" option.

And we use brown clustering [14] to improve the recognition's effect. Brown clustering is an agglomerative, bottom-up form of clustering that groups words into a binary tree of classes, using a merging criterion based on the log-probability of a text under a class-based language model. Our system uses the cluster memberships of words resulting from Brown clustering as features of each entity. At last, we run for 5 times in different ways: without brown clustering, 500 clusters, 1,000 clusters, 1,500 clusters, 2,000 clusters. Experiments with brown clusters have one more feature than "without brown clusters" in CRF++ template file "brown tokens".

However, our results are not so good as we expect (Table 4). In the analogous experiment, the entity subtask in the BioCreative IV CHEMDNER competition, the official scores are higher than us. The average precision, recall, F1 score are at about 89.21%, 66.41%, 76.11% respectively[1].

In addition to our system own reasons, some factors that may affect the results. The research using paper corpus often do experiments with the title, abstract and keywords of paper and it has less noise data. However, we use the patent corpus with full text. Patents are focused on the protection of intellectual property rights but papers on the knowledge dissemination and sharing. In order to protect the intellectual property rights and innovation, patent documents will write in a special way. On the contrary, the author can choose the way that readers make it easier to understand in the paper.

5. CONCLUSIONS

We develop a chemical and biological entity recognition system and use the annotated chemical patent corpus to do the experiment with the system. In our recognition system, we regard it as a sequence labeling problem instead of extracting the whole entity at once. We utilize some open-source NLP toolkits, such as OpenNLP, Stanford CoreNLP, and do some modification to appropriate for the patent corpus with some additional rules. In our system, CRF++ is adopted for the actual implementation to process the sequence label problem. However, the results are not so good as we expect. As it shows in Table 4, we get too much FP results and nothing in FN. Maybe the entities annotated in one patent but not annotated in another one influence the experiment results. We will define some suitable rules to improve the recognition system in the future.

Table 4 Performance results in our system for the gold standard patent corpus[2].

	Run 1	Run 2	Run 3	Run 4	Run 5
best cost	2^1	2^1	2^0	2^1	2^1
TP	28981	29655	29473	29502	29451
TN	10517	15262	15626	15568	15668
FP	16607	11131	10790	11027	10875
FN	0	0	0	0	0
Precision (%)	63.57	72.71	73.20	72.79	73.03
Recall (%)	73.37	66.02	65.35	65.46	65.27
F1 score (%)	68.12	69.20	69.05	68.93	68.94

6. ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China: Research on Technology Opportunity Detection based on

[1] The experiment data using the official dataset is available at website: http://www.sciteminer.org/XuShuo/Demo/CEM .

[2] Run 1 is the experiment without brown clusters. The other four runs are respectively brown clustering's number of 500, 1,000, 1,500, 2,000 clusters.

Paper and Patent Information Resources under grant number 71403255, and Key Work Project of Institute of Scientific and Technical Information of China (ISTIC): Research and Development on Knowledge Organization System and Intelligent Analysis Service Demonstration Platform for Science and Technology Literature in New Material Domain under grant number ZD2014-7-7.

7. REFERENCES

[1] Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, et al. (2011) Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. Drug Discov Today 16: 1019–1030.

[2] Southan C, Boppana K, Jagarlapudi SA, Muresan S (2011) Analysis of in vitrobioactivity data extracted from drug discovery literature and patents: Ranking 1654 human protein targets by assayed compounds and molecular scaffolds. J Cheminform 3: 14.

[3] De Ridder, D. et al. 2013. Pattern recognition in bioinformatics. *Briefings in Bioinformatics*. 14, 5 (Sep. 2013), 633–647.

[4] Grego, T. et al. 2009. Identification of Chemical Entities in Patent Documents. *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, Pt Ii, Proceedings*. S. Omatu et al., eds. Springer-Verlag Berlin. 942–949.

[5] Campos, D. et al. 2013. Gimli: open source and high-performance biomedical name recognition. Bmc Bioinformatics. 14, (Feb. 2013), 54.

[6] Han, H. et al. 2014. Mining technical topic networks from Chinese patents. 1st International Workshop on Patent Mining and Its Applications, IPaMin 2014, Co-located with Konvens 2014, October 6, 2014 - October 7, 2014 (2014).

[7] Roman Klinger, Corinna Kolarik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich, 2008. Detection of IUPAC and IUPAC-Like Chemical Names. Bioinformatics, Vol. 24, No. 13, pp. i268-i276.

[8] Akhondi, S.A. et al. 2014. Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. PLoS ONE. 9, 9 (2014), e107477. DOI: 10.1371/journal.pone.0107477

[9] Zimmermann, M. et al. 2005. Information Extraction in the Life Sciences: Perspectives for Medicinal Chemistry, Pharmacology and Toxicology. Current Topics in Medicinal Chemistry. 5, 8 (Aug. 2005), 785–796.

[10] Hobbs J R. The generic information extraction system[C]//MUC. 1993: 87-91.

[11] Xu S, An X, Zhu L, et al. A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature[J]. Journal of Cheminformatics, 2015 (Suppl 1): S11.

[12] Wei, C.H., Harris, B.R., Kao, H.Y., Lu, Z.: tmVar: A text mining approach for extracting sequence variants in biomedical literature. Bioinformatics 129(11) (2013) 1433–1439

[13] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML'01. (2001) 282–289

[14] Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 384-394). Association for Computational Linguistics.