Comparison of different strategies for utilizing two CHEMDNER corpora

Thaer M Dieb and Masaharu Yoshioka

Graduate School of Information Science and Technology, Hokkaido University
{diebt@kb.ist.hokudai.ac.jp
yoshioka@ist.hokudai.ac.jp}

Abstract. To identify chemical entities and drug names in patent according to CHEMDNER patent task-CEMP subtask, we use machine learning technique to construct a chemical named entity recognition (CNER) system. It is desirable for machine-based CNER system to have large training examples. Two CHEMDNER corpora have been developed. One is the corpus for the patent task and the other is the CHEMDNER corpus for PubMed abstract constructed for CHEMDNER task in BioCreative IV. Both corpora were constructed based on very similar guidelines. However, the style of writing is different. In this paper, we are discussing different strategies to utilize these two corpora to identify chemical entities in patent. Our basic system uses conditional random field (CRF) as a machine learning technique that uses linguistic features in addition to domain knowledge feature produced by ChemSpot. We compare the results of these strategies using simple system performance measures (e.g., recall, precision, and F-score) and analysis on the unique findings of each system.

Key words: Chemical entity recognition, Chemdner corpora, Conditional random field

1 Introduction

Chemical named entity recognition (CNER) systems that use machine learning techniques would prefer to have large training examples that cover wide varieties of entities related to the task in order to generate efficient rules to identify such entities' mentions. For the BioCreative V, CHEMDNER track-CEMP subtask: the detection of chemical named entity mentions in patents, there are two corpora available. One is the corpus for the patent task and the other is the CHEMDNER corpus for PubMed abstract constructed for CHEMDNER task in BioCreative IV. Even though these corpora were constructed based on a very similar annotation guideline, the style of writing for patent is different from the one for abstract of the research paper. In this research, we discuss different strategies for utilizing these two CHEMDNER corpora using machine learning system, and compare results for clarifying the issues related to these strategies.

In order to compare these strategies, we implemented a CNER system that uses machine learning technique. We use three strategies to train the system : 1)

Simple: Use CHEMDNER patent corpus only, 2) Merge: Merge the two CHEMDNER corpora for training to enlarge training examples, 3) Domain adaptation: Use output of basic system trained on CHEMDNER PubMed corpus as an additional feature of the CRF that uses CHEMDNER patent corpus for training. This can help learning any consistent differences between annotation schemas of both corpora.

We compare the results of these strategies using simple system performance measures (e.g., recall, precision, and F-score). In addition, we use unique findings analysis of each strategy to characterize its behavior.

Since both corpora use almost the same guidelines, we confirm that merging both of them can leverage the performance without any harm.

2 Systems description and methods

2.1 System outline

Our system uses CRF++ [1] (Version 0.58), an implementation of conditional random field (CRF) [2] as a machine learning system. We use two kinds of features, one is linguistic features generated based on the output of GPoSTTL tagger (Version 0.9.3) which is a basic text tokenizer and part-of-speech tagger, and the other is domain knowledge feature generated by ChemSpot [4], a common CNER tool. We have used the latest version of ChemSpot (ChemSpot 2.0 with updated dictionary file and ids list). However, since this version is developed after CHEMDNER PubMed corpus, it might have used CHEMDNER PubMed corpus data for training. Because of that, we implemented another version of the system with ChemSpot 1.6 (an older version developed prior to CHEMDNER PubMed corpus). Both tools are available at [5]. All features are encoded as IOB format in the CRF++ system. Below is a list of the features used:

- Surface symbol: symbol used to represent a term.
- Part-of-speech (POS) tag: result from the GPoSTTL tagger [3].
- Lemmatization: symbol that is a result from the POS tagger.
- Orthogonal feature: was added using regular expressions based on the definition in [6].
- ChemSpot tag: output of ChemSpot tool.

2.2 Text tokenization

For chemical named entity recognition tasks, a general, linguistic oriented POS tagger (such as the GPoSTTL tagger) might not be good enough to tokenize text in regards with chemical entities' boundaries. For example, some entities might fall within a token and can't be labeled correctly. Similar problems in the biomedical domain have already been discussed [7].

To solve this problem, it is necessary to apply a particular tokenization technique to achieve better labeling results. We have applied a post tokenization mechanism [8] to generate a greater number of smaller tokens with shorter sizes for a more flexible labeling. Basically, we further tokenized chunks containing "–", "+" and " / " into multiple tokens.

2.3 Experiment

Before we discuss the results of our system with the three strategies, we mentioned in the Introduction section, we present the performance of both versions of ChemSpot on the CHEMDNER PubMed corpus and CHEMDNER patent corpus. Table 1 shows these results.

Table 1. Performance of ChemSpot on CHEMDNER PubMed and patent corpora

Version	**PubMed corpus**						**Patent corpus**					
	Macro-average			**Micro-average**			**Macro-average**			**Micro-average**		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
ChemSpot2.0	0.75	0.85	0.77	0.76	0.84	0.80	0.70	0.73	0.70	0.72	0.69	0.71
ChemSpot1.6	0.67	0.60	0.60	0.72	0.61	0.66	0.61	0.55	0.56	0.70	0.51	0.59

Since the official test dataset of CHEMDNER patent is not yet released, we evaluated the three strategies on the development dataset. The training data for each strategy is as follows:

- Simple: we used the CHEMDNER patent corpus-train dataset for training.
- Merge: we used both CHEMDNER patent corpus-train dataset and full CHEMDNER PubMed corpus for training.
- Domain adaptation: first, we used the output of the basic system trained on CHEMDNER PubMed corpus as an additional feature, and then we used CHEMDNER patent corpus-train dataset for training.

For all the three strategies, we used both versions of ChemSpot for the evaluation. Table 2 shows the results or our system using the three strategies.

Table 2. Performance of our system with different training strategies

		Macro-average			**Micro-average**		
	Strategy	Precision	Recall	F-score	Precision	Recall	F-score
System with ChemSpot2.0	**Simple**	0.82	0.80	0.80	0.86	0.80	0.83
	Merge	0.82	0.81	0.80	0.85	0.81	0.83
	Domain adaptation	0.82	0.80	0.80	0.86	0.81	0.83
System with ChemSpot1.6	**Simple**	0.80	0.74	0.75	0.86	0.75	0.80
	Merge	0.81	0.76	0.77	0.86	0.77	0.81
	Domain adaptation	0.81	0.74	0.75	0.86	0.76	0.81

For a machine-learning based CNER system, it is easier to identify entities exist in training data. However, for new entities that uniquely exist in the test data (and not in the training data), identification can be more challenging. To further characterize the three strategies of utilizing both CHEMDNER corpora, we have analyzed their performance on identifying unique entities in the experimental dataset (in this case, entities which exist in the development dataset and not in the train dataset of CHEMDNER patent corpus). Additionally, we have performed such analysis for ChemSpot tools. Table 3 shows the results of this analysis for ChemSpot tools. Table 4 shows the results for our system using different training strategy.

Table 3. Unique entities identification analysis for ChemSpot tools

Mentions	32142
Unique	13009
ChemSpot2.0	7979
Coverage	0.61
ChemSpot1.6	5842
Coverage	0.45

Table 4. Unique entities identification analysis for different training strategies

	Simple	**Merge**	**Domain adaptation**
Mentions	32142	32142	32142
Unique	13009	13009	13009
System with ChemSpot2.0	9471	9458	9437
Coverage	0.73	0.73	0.73
System with ChemSpot1.6	8268	8311	8252
Coverage	0.64	0.64	0.63

Merging both corpora resulted in reducing the unique terms in experiment dataset (development dataset of CHEMDNER patent corpus) to 11451 instead of 13009, and thus slightly improved the identification of unique entities.

For the CHEMDNER patent task-CEMP subtask, we have submitted 3 runs based on best performed strategies as follows:

- Run 1: first, we used the output of the basic system trained on CHEMDNER PubMed corpus as an additional feature, then we used the CHEMDNER patent corpus (train and development datasets) for training.
- Run 2: we used both CHEMDNER patent corpus (train and development datasets) and full CHEMDNER PubMed corpus for training. For the features, we used the same basic features as discussed in the section System outline.

– Run 3 : we used the same basic features as discussed in the section System outline. For training, we used CHEMDNER patent corpus (train and development datasets) for training.

The evaluated results of the three runs [9] are shown in table 5.

Table 5. Evaluated results of Submitted runs

Run	Precision	Recall	F-score
Run 1	0.87031	0.83811	0.85391
Run 2	0.86437	0.8425	0.85329
Run 3	0.87002	0.83617	0.85276

3 Discussion

From table 2 data, we can notice that all three strategies have similar results when we use ChemSpot 2.0 output as a feature in our system. Since ChemSpot 2.0 is developed after CHEMDNER PubMed corpus, there might be a possibility that it uses that corpus data for training, thus utilizing CHEMDNER PubMed corpus has almost no effect with ChemSpot 2.0 as a feature. However, in the case of ChemSpot 1.6 (which was developed prior to CHEMDNER PubMed corpus), there is a little improvement in the performance with the merge strategy. Domain adaptation strategy did not have leveraging impact above the simple approach that utilizes only one corpus data. Since both corpora are based on very similar guidelines, we conclude that merge approach strategy might be useful without harming the performance.

Unique entities identification analysis can show that merge approach has decreased the number of unique terms and thus slightly improved the performance.

Generally speaking, when merging several corpora for a certain CNER task, it is recommended to consider the guidelines of these corpora. In case these guidelines are very similar, merge approach can be useful with harming the performance.

4 Conclusion

In this report, we discussed different strategies to utilize both CHEMDNER corpora available (PubMed and patent) in order to identify chemical entities in patent. We found that since both corpora are based on very similar guidelines, merging both corpora can be useful without harming the performance.

Acknowledgments This work was partially supported by JSPS KAKENHI Grant Number 25280035.

References

1. CRF++ tool: http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar, accessed Aug. 30, 2015.
2. Lafferty, J.D, McCallum, A., Pereira, F. : Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. the Eighteenth International Conference on Machine Learning. ICML 01, San Francisco, CA, USA, 282289, (2001).
3. GPoSTTL http://gposttl.sourceforge.net, accessed Aug. 30, 2015.
4. Rocktschel, T., Weidlich, M., and Leser, U. : ChemSpot: A Hybrid System for Chemical Named Entity Recognition. Bioinformatics 28 (12): 1633-1640 (2012).
5. ChemSpot tools: https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/chemspot/chemspot
accessed Aug. 30, 2015.
6. Takeuchi, K. and Collier, N. : Bio-medical entity extraction using support vector machines. Artif. Intell. Med. 33 (2) : 125-137 (2005).
7. Yeh, A., Morgan, A., Colosimo, M., Hirschman, L. :BioCreAtIvE Task 1A: gene mention finding evaluation. BMC Bioinformatics, 6(Suppl 1):S2 (2005) doi:10.1186/1471-2105-6-S1-S2.
8. Dieb, T.M., Yoshioka, M. :Extraction of Chemical and Drug Named Entities by Ensemble Learning Using Chemical NER Tools Based on Different Extraction Guidelines. Trans. on machine learning and data mining. (Accepted for publication)
9. Krallinger et al. Overview of the CHEMDNER patents task. Proceedings of the Fifth BioCreative Challenge Evaluation Workshop (2015).