Ensemble Approach to Extract Chemical Named Entity by Using Results of Multiple CNER Systems with Different Characteristic

Masaharu YOSHIOKA Thaer M. DIEB

Graduate School of Information Science and Technology, Hokkaido University
N-14 W-9, Kita-ku, Sapporo 060-0814, JAPAN

Abstract. We propose a novel ensemble approach chemical named entity recognition (CNER) tool that uses different CNER tools such as OSCAR4 and ChemSpot with different characteristics by using machine learning (ML) technique. Since this tool may identify typical errors of one CNER by using other tools' output, our system outperforms ChemSpot (ML-based) and OSCAR4 (rule-based) in original setting.

1 Introduction

This reports describes our system that participated in the BioCreative IV, Track 2- CHEMDNER Task: Chemical compound and drug name recognition task, Chemical entity mention recognition sub-task [1].

There are various chemical named entity recognition (CNER) tools to identify chemical entities in the research articles. There are two main approaches for this task. One is rule and dictionary based approach that uses chemical dictionary and syntactic patterns for representing chemical named entity. OSCAR4 [2] is one of the best open system based on this approach. The other approach is machine learning approach that uses several linguistic features such as POS, lematization form, and orthogonal features to identify chemical named entity. ChemSpot [3] is one of the best open system based on this approach that also uses dictionary based feature.

Based on the analysis of OSCAR4 and ChemSpot in this task, ChemSpot is good at precision, but is not good at recall. On the contrary, OSCAR4 is good at recall, but is not good at precision. In addition, when we merge both results, recall increased because there are many unique answers that only one system can identify.

However, since there are typical cases in the unique answers, it is not good to use simple ensemble approach based on voting [4]. Therefore, we decide to use output of these CNER tools as features of the keyword by using Conditional Random Field (CRF) [5] as a machine learning technique.

We confirm our approach outperforms ChemSpot and OSCAR4 that are not trained for this corpus. In addition, we also confirm our framework is consistently improve the performance by integrating output from different characteristic systems.

2 Systems description and methods

2.1 Preliminary experiments and methods

OSCAR4 [2] and ChemSpot [3] is well known open CNER system. Both systems have good performance on SCAI corpus (F1 is larger than 0.7). However, due to the difference between SCAI corpus guideline and this task, it is necessary to understand the basic performance of these systems.

Table 1 shows preliminary evaluation results by two fold cross validation (average of result using train as training and development for evaluation and one using development as training and train for evaluation).

From this results, we confirm the performance of these two system is not equivalent with one for SCAI corpus. We assume this may comes from the difference between corpus construction guideline. Therefore, it is necessary to have a mechanism to adopt this difference.

"ChemSpot+OSCAR4" in Table 1 shows evaluation results that merge results of two systems and exclude duplication. Based on the comparison between recall, we confirm both systems have unique answers that other system cannot recognize.

As we confirmed in this preliminary experiment, ChemSpot and OSCAR4 have different characteristic and find out unique answers. Therefore, in this paper, we construct CNER tool that integrates output of two different systems.

Table 1. Preliminary evaluation by two fold cross validation

Name	Macro			Micro		
	Prec.	Rec.	F	Prec.	Rec.	F
OSCAR4	0.43	0.63	0.48	0.41	0.62	0.49
ChemSpot	0.66	0.58	0.58	0.72	0.59	0.65
ChemSpot+OSCAR4	0.45	0.79	0.54	0.44	0.79	0.56

Prec.: Precision, Rec.: Recall, Avg-P: Average precision

2.2 Systems description

We employ common approach to use CRF with multiple characteristic features (e.g., linguistic features, and others) to predict the target token type.

We use following modules for proposed system.

- CRF++ [1]: Implementation of CRF classifier
- ChemSpot[2]: A CNER tool based on a hybrid approach combining a Conditional Random Field with a dictionary.

[1] CRF++-0.58:http://crfpp.googlecode.com/svn/trunk/doc/index.html

[2] Chemspot1.5:http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/resources/chemspot/chemspot/

– OSCAR4[3]: A CNER tool that uses dictionary, rule and NLP techniques. There are several categories for OSCAR4 output. In this system we use CM (compound) for annotation. In addition to CM itself, RN(reaction) and CJ (adjective) followed with CM are also used as a part of chemical named entity.

In addition to the system output, we also uses linguistic features (symbol, POS, lematization form, and orthogonal features) and output of our rule-based CNER tools [6]. This CNER tools uses similar rules of OSCAR4, but it tends to exclude common ordinary word (e.g., In, ...) for obtaining better precision.

One of the difficulties to integrate results of different system is an inconsistency of segmentation. For example, "cyclooxygenase-2" in abstract 22080037 is treated as one element in OSCAR4 and selected as chemical, but ChemSpot separate it into "cyclooxygenase", "-", and "2". In order to handle this type of inconsistency, we translate results into same format output into "22080037 A:245-261" and mapped into the final POS tagging results. In this mapping process, all segmented elements that include a part of the tagged offset results of other CNER are treated as tagged for that CNER.

Followings are list of features used in the system.

– Surface symbol: symbol that uses for representing term.
– Part-Of-Speech (POS) tag: result of GPoSTTL tagger [4], which is a brill's Parts-of-Speech tagger, as built-in tokenizer and lemmatizer.
– Lemmatization: symbol that is a result from POS tagger.
– Orthogonal feature: symbol that represent styles of surface symbol (such as all capitals, lowercase, and digits). This information is identified using regular expressions based on the POS tag.
– CNER tag: output of our CNER system in IBO format. our CNER is a rule-based chemical entity recognizer that uses regular expression to identify chemical compounds. in addition, CNER is uses syntactical rules to solve some mismatches that might occur between chemical compounds and normal text. For example, we assume that short words falling in the beginning of a sentence such as "In" is not chemical compound (such as Indium). also we try to identify abbreviations of technical terms identified within the document and avoid tagging them as chemical compounds.
– ChemSpot tag: output of ChemSpot in IBO format
– OSCAR4 tag: output of OSCAR4 in IBO format

Table 2 shows an example of the IOB style data for CRF training and evaluation by using these information.

Template for generating features is almost same as one for Japanese NE of CRF++ tool kit (modification of template for handling large number of features for one element). Confidence for the extracted terms are calculated based on the

[3] chemicalTagger-1.3: http://chemicaltagger.ch.cam.ac.uk/ We use output of OSCAR4 related output of ChemicalTagger
[4] gposttl-0.9.3: http://gposttl.sourceforge.net

Table 2. Sample of the input file format for CRF (part of the text)

Surface	POS	Lemmatization	Orthogonal	CM	ChemSpot	OSCAR4	CEM
cyclooxygenase-2	JJ	cyclooxygenase-2	OtherHyphon	O	O	B	O
(	(	(	Other	O	O	I	O
COX-2	NP	cox-2	TwoCaps	O	O	I	O
)	)	)	Other	O	O	I	O
,	,	,	Comma	O	O	O	O

output of CRF. For the confidence value for multiple terms are calculated by multiplication of confidence value for all values for "B" and "I".

In this sub-task, we submit following 3 results with different configuration. All of them using the above mentioned basic features.

1. Run1: uses all features discussed in before and training data and development data for CRF learning. However, it includes some errors for text normalization (Mostly for back up purpose).
2. Run2: uses all features discussed in before and training data and development data for CRF learning. Text normalization error of Run 1 are mostly solved.
3. Run3: uses same data of Run2, but uses different CRF parameter (hyperparameter c=1.5 minimum number of features f=3).

3 Discussion

Since there are no information about final data, we conduct experiments by using train and development data as two fold cross validation data.

Table 3 shows result of several evaluational results for different settings. Run1-3 corresponds to the submitted system and Run2-OSCAR4 and Run2-ChemSpot is for Run2 based system without OSCAR4 or ChemSpot respectively. Run2Basic-OSCAR4 and Run2Basic-ChemSpot is for basic elements of Run2 system (Surface symbol, POS, lemmatization form and Orthogonal feature) and OSCAR4 or ChemSpot respectively. Since there are no rank for the output of OSCAR4 and ChemSpot, average precision is marked as "-".

Basically, our submitted systems ("Run1-3") have similar results that outperforms other systems and settings especially in recall.

Performance of "ChemSpot" is worse than one for the case of SCAI corpus. Since it may comes from the inconsistency between the tagged guideline between SCAI corpus (training data of ChemSpot) and one for this task. Since "Run2Basic+ChemSpot" uses task corpus and performance is better than "ChemSpot". This difference shows that it is necessary to use training data for assimilating different guideline.

In addition, from the comparison between "Run2" and other "Run2" based system, we confirm different characteristic features supports to improve recall without loosing precision. There are several cases that ensemble based approach detect appropriate term boundary to improve recall and precision at the same

Table 3. Performance evaluation by two fold cross validation

Name	Macro				Micro			
	Prec.	Rec.	F	Avg-P	Prec.	Rec.	F	Avg-P
Run1	0.80	0.69	0.72	0.59	0.83	0.69	0.75	0.66
Run2	0.77	0.68	0.70	0.56	0.80	0.68	0.73	0.57
Run3	0.76	0.68	0.70	0.56	0.79	0.68	0.73	0.57
Run2-OSCAR4	0.74	0.64	0.67	0.52	0.78	0.65	0.71	0.54
Run2-ChemSpot	0.76	0.64	0.67	0.53	0.80	0.63	0.71	0.53
Run2Basic+OSCAR4	0.76	0.64	0.67	0.52	0.81	0.62	0.70	0.53
Run2Basic+ChemSpot	0.74	0.64	0.66	0.52	0.78	0.64	0.71	0.53
OSCAR4	0.43	0.63	0.48	-	0.41	0.62	0.49	-
ChemSpot	0.66	0.58	0.58	-	0.72	0.59	0.65	-

Prec.: Precision, Rec.: Recall, Avg-P: Average precision

time. However adding similar types of CNER system (our CNER system and OSCAR4) may not be so much effective than adding different characteristic systems.

4 Conclusion

In this report, we have discussed an ensemble approach using machine learning system and aggregates different chemical named entity recognizers with different characteristics in order to identify chemical entities. We confirm the usefulness of our approach by using training and development data for BioCreative IV, Track 2- CHEMDNER Task.

Acknowledgments This work was partially supported by JSPS KAKENHI Grant Number 24240021.

References

1. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Oyarzabal, J., Valencia, A.: Overview of the chemical compound and drug name recognition (CHEMDNER) task. In: Proceedings of the fourth BioCreative challenge evaluation workshop, vol. 2. (2013) (to appear).
2. Jessop, D., Adams, S., Willighagen, E., Hawizy, L., Murray-Rust, P.: OSCAR4: a flexible architecture for chemical text-mining. Journal of Cheminformatics **3** (2011) 41
3. Rocktäschel, T., Weidlich, M., Leser, U.: ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics **28** (2012) 1633–1640
4. Zhou, G., Shen, D., Zhang, J., Su, J., Tan, S.: Recognition of protein/gene names from text using an ensemble of classifiers. BMC Bioinformatics **6** (2005) 1–7
5. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289

6. Dieb, T.M., Yoshioka, M., Hara, S.: Automatic information extraction of experiments from nanodevices development papers. In: Proceedings of The 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM2012). (2012) 42–47