Transactions on Machine Learning
and Data Mining
Vol. 8, No. 2 (2015) 61-76
ISSN 1864-9734
ISBN: 879-3-942952-39-2

ibai Publishing
www.ibai-publishing.org

Extraction of Chemical and Drug Named Entities by Ensemble Learning Using Chemical NER Tools Based on Different Extraction Guidelines

Thaer M. Dieb and Masaharu Yoshioka

Graduate School of Information Science and Technology, Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
diebt@kb.ist.hokudai.ac.jp

Abstract. Chemical named-entity recognition (chemical NER) is the task of extracting chemical information and chemical-related entities such as drug names and source materials from text in several domains such as bioinformatics and nanoinformatics. There have been several attempts to construct corpora for handling such chemical-related information based on different corpus-construction guidelines. Even though these guidelines contain common types of chemical information, they differ in several ways. As a result, chemical NER tools developed for a particular guideline might be able to extract common chemical named entities, but they may have problems extracting other chemical-related entities. Assuming the differences between these guidelines are consistent, the pattern of success and failure of the chemical NER tools might also be consistent. In this paper, we present an ensemble-learning approach that uses the conditional random field (CRF) as a machine-learning technique to fuse a variety of different characteristic chemical NER tools based on different guidelines to construct a chemical NER for a particular guideline. To achieve consistent tokenization across these different tools, we applied a post-tokenization mechanism. We evaluated the system using the BioCreative IV, CHEMDNER task datasets. We confirmed that the ensemble-learning approach using a combination of chemical NER tools is better than a simple domain-adaptation approach using just one chemical NER tool. We also confirmed that the ensemble-learning approach could improve the performance of a well-tuned rule-based chemical NER tool on certain tasks.

Keywords: Chemical named entities recognition, Ensemble learning, Conditional random field, Text tokenization.

1 Introduction

Recently, we have become able to use large quantities of textual data for extracting useful information. As an example, we can use a research article database as "big data" for understanding research trends and new research results. This is a new frontier for utilizing machine-learning techniques. There are two main approaches in this domain, namely analyzing bibliographic information to identify research trends [1, 2] and extracting useful information by using text-mining techniques [3, 4].

Chemical named-entity recognition (chemical NER) is an application domain for extracting chemical information from text. Extraction of all the chemical named entities from a paper is desirable in finding articles that are related to particular chemical named entities. In addition, chemical information plays a significant role in a variety of related disciplines such as bioinformatics [5] and nanoinformatics [6]. For example, chemical information could help detect drug-protein interaction in the bioinformatics domain or source materials in nanodevice-development experiments within the nanoinformatics discipline.

Chemical NER tasks began with extracting general chemical named-entity information and expanded to meet the demand for extracting chemical-related named entities (such as drugs) that are used in particular research domains. In such an expansion, new guidelines were created to include the new types of entities. Because these guidelines also aim to extract general chemical named entities, they are similar to those used for the general chemical NER task but also include guidelines for the extraction of new types of chemical-related named entities. At an early stage, the SCAI corpus [4] of general chemical named entities was created to identify the International Union of Pure and Applied Chemistry (IUPAC) entities [7]. Another approach was to create a chemical named-entity dictionary such as the Chemical Entities of Biological Interest (ChEBI) [8].

To support these chemical NER tasks, several chemical NER tools were developed to extract chemical named entities from text. Because there was no standard guideline of the chemical NER task, these tools were developed based on one of the guidelines. At an early stage of chemical NER tool development, most tools were evaluated using the SCAI corpus, which was a large corpus that was freely available for such tasks. This means that most chemical NER tools, such as OSCAR4 [9], and ChemSpot [10], were developed primarily for the chemical NER task defined by the SCAI corpus.

In this paper, we propose a method for applying these chemical NER tools to the BioCreative IV, CHEMDNER task [5] based on an ensemble-learning technique. This task aims to extract drug names in addition to general chemical named entities. BioCreative IV, CHEMDNER corpus uses the abstracts of chemical-related papers in MEDLINE. Chemical-related terms are identified as the offset information of such terms from the beginning position of the text. Left side of Figure 1 shows an example of those terms' information. Right side of the figure represents illustrated interpretation using abstract text information.

For this very recent task, a simple ensemble-learning approach based on voting [11] is not appropriate. Therefore, we use all the system outputs of these chemical NER tools as features of a conditional random field (CRF) model [12], in addition to linguistic features, such as lexical and orthogonal features, that are widely used for chemical

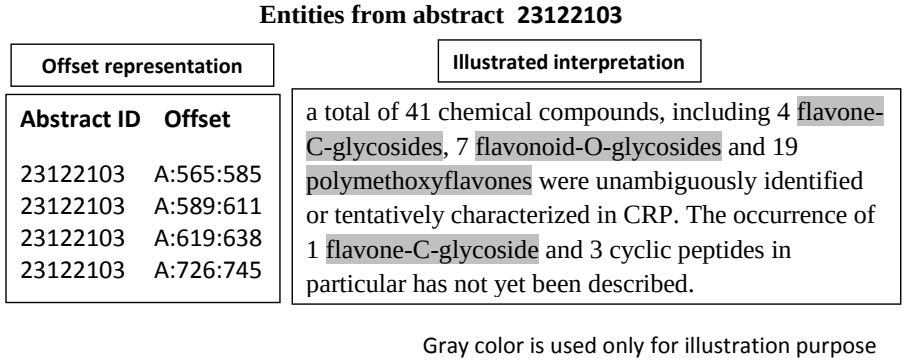

Fig. 1. BioCreative IV, CHEMDNER corpus data snapshot

NER tasks. This approach is similar to the concept of domain adaptation [13] in natural language processing (NLP), which uses machine-learning results from corpora of a variety of domains to analyze texts in a new domain.

Because different chemical NER tools use different tokenization schemas for text tokenization, it is necessary that the ensemble-learning approach should handle any inconsistency between the outputs of the chemical NER tools. We apply a post-tokenization mechanism to generate more flexible tokenization schema that can adapt to a variety of chemical NER-tool tokenization schemas.

This paper has five sections. Following this Introduction, Section 2 reviews related work involving chemical NER tools and ensemble learning in NLP. Section 3 discusses the construction of a chemical NER tool using the output of other chemical NER tools. We also discuss how to integrate tokenization results from different tools. In Section 4, our chemical NER tool is evaluated in terms of the BioCreative IV, CHEMDNER corpus. To investigate the effectiveness of our ensemble-learning approach, we conducted three experiments. The first used chemical NER tools developed prior to the BioCreative IV, CHEMDNER task. The second used rule-based chemical NER tools developed for the BioCreative IV, CHEMDNER task. In the third experiment, we evaluated our system using the official test dataset of the BioCreative IV, CHEMDNER task. We confirm that the ensemble-learning approach outperforms the standalone performance of each chemical NER tool by a statistically significant amount. Even for a rule-based system that is tuned for a specific task, the ensemble-learning approach can offer a slight but statistically significant improvement in precision and F-score. In addition, we confirm that our tokenization mechanism considerably improved the performance of the ensemble-learning approach. Section 5 concludes the paper.

2 Related Work

There are two major approaches to implementing chemical NER tools. The first is a machine-learning approach that uses several linguistic features such as POS, lemma-

tization form, and orthogonal features to identify chemical named entities. For example, ChemSpot [10] is one of the best open systems based on this approach. It also uses dictionary-based features. Because the performance of the released version of ChemSpot is mainly evaluated in terms of the SCAI corpus, this tool has been developed to extract chemical named entities based on SCAI guidelines. The second approach is rule based, using a chemical dictionary and syntactic patterns to represent chemical named entities via regular expressions. For example, OSCAR4 [9] is one of the best open systems implementing this approach (OSCAR4 also uses machine-learning-based methods in the form of a maximum-entropy Markov model). OSCAR4 uses the guidelines of the ChEBI database. Other tools also use a hybrid approach, combining rule-based and machine-learning-based methods.

Because there are many research domains that use chemical information, these chemical NER tools have been applied to a variety of research domains. However, because of the variations in chemical-related entities across domains, these tools may not be sufficient to extract all chemical-related information in a particular domain. For example, drug names are chemical-related named entities, but general chemical NER tools cannot recognize all of them. To meet these new demands, a new corpus for extracting chemical and drug names was developed in the BioCreative IV, CHEMD-NER task [14]. Even though the corpus guidelines share certain chemical entities with general chemical NER guidelines, several differences remain. Because of these differences, chemical NER tools that were developed using general chemical NER guidelines, such as ChemSpot and OSCAR4, might not perform well when tested with the new task. Some implementations for named-entity recognition have adopted an ensemble-learning approach. For example, Dimililer et al. [15] describe classifier subset selection for biomedical named-entity recognition. In this work, a vote-based classifier selection scheme has an intermediate level of search complexity between static classifier selections and real-value and class-dependent weighting approaches. Zhou et al. [16] describe voting-based ensemble classifiers to detect hedges and their scopes. Another ensemble-learning approach assumes that instead of searching for the best-fitting feature set for a particular classifier, an ensemble of several classifiers that are trained using different feature representations could be a more fruitful approach. For example, Ekbal et al. [17] apply this approach to named-entity recognition. However, all of these approaches assume that all machine-learning systems are constructed for the same task.

3 Framework for Ensemble-learning Approach

3.1 Framework Architecture

The ensemble approach we are proposing uses CRF to fuse several chemical NER tools that use different recognition schemas. This framework decomposes input text into a sequence of tokens (tokenization), generates characteristic features for each token, namely linguistic features, and the results of chemical NER tools for this token, and then uses CRF to predict the label of the token. Based on CRF results, the system can identify chemical entities and drug names in a text.

A General text tokenizer (e.g., POS tagger) might not be good enough to adapt to multiple tokenization schemas applied by different chemical NER tools. Our system im-

plements a post-tokenization mechanism to overcome such problems. First, we discuss the tools we are using in more detail.

Chemical NER Tools We have used the following chemical NER tools:

- **SERB-CNER (Syntactically enhanced rule-based chemical NER)** is a rule-based chemical-entity recognizer that uses regular expressions to identify chemical compounds. This tool also uses syntactical rules to solve any mismatches that might occur between chemical compounds and normal text. For example, we assume that short words at the beginning of a sentence, such as "In", do not represent a chemical compound such as Indium. We also try to identify abbreviations of technical terms used within the document and avoid tagging them as chemical compounds. We originally developed this tool to identify chemical compounds in nanocrystal-development research papers [18].
- **ChemSpot** is a named-entity recognition tool for identifying mentions of chemicals in natural language texts, including trivial names, drugs, abbreviations, molecular formulas and IUPAC entities. ChemSpot uses a hybrid approach that combines a CRF with a dictionary. ChemSpot is trained by using SCAI corpora [19] annotated mainly with IUPAC [7] entities.
- **OSCAR4** is an open-source extensible system for the automated annotation of chemistry in scientific articles. It uses a rule-based approach, in addition to machine-learning-based methods in the form of a maximum-entropy Markov model, to identify chemical entities.

Linguistic Features We have used GPoSTTL as a basic text tokenizer and part-of-speech tagger to define the basic type of each token. GPoSTTL is an enhanced version of Eric Brill's rule-based tagger. In addition to the POS tag, GPoSTTL generates a lemmatization feature. Based on GPoSTTL results, we use regular expressions to generate orthogonal features as defined in [20]. An orthogonal feature is a symbol that represents various styles of surface symbols (such as all capitals, lowercase, or digits).

Conditional Random Field (CRF) A CRF [12] is a probabilistic sequence-labeling model commonly used in NER tasks. In such a task, a CRF model takes an input of a text token sequence and seeks to assign a categorical label for each member of a sequence relying on statistical inference. Because a named entity might span over multiple tokens, IOB format is used to define entity boundaries, where "B" identifies the beginning of a named entity, "I" declares that the token is inside the named entity and "O" means that the token is outside the named entity. To label the token sequence, a CRF model builds a set of inference rules using a training dataset in which each token is attached to a feature set and labeled correctly. As noted linguistic features such as token surface, POS tag, lemmatization, and orthogonal features, are commonly used in NER tasks.

The inference rules take into consideration the target label of a token in relation to both its own feature set and also the feature sets of neighboring tokens within a certain feature window size. The feature window is defined as a function of the target label to n-gram feature combination. For example, in a bigram, the current target label is defined

as a function of the combination of two features, one from the current token's feature set plus another one from a neighboring token's feature set. This makes CRF well suited for natural language processing applications [21, 22]. Figure 2 shows an outline of the CRF model.

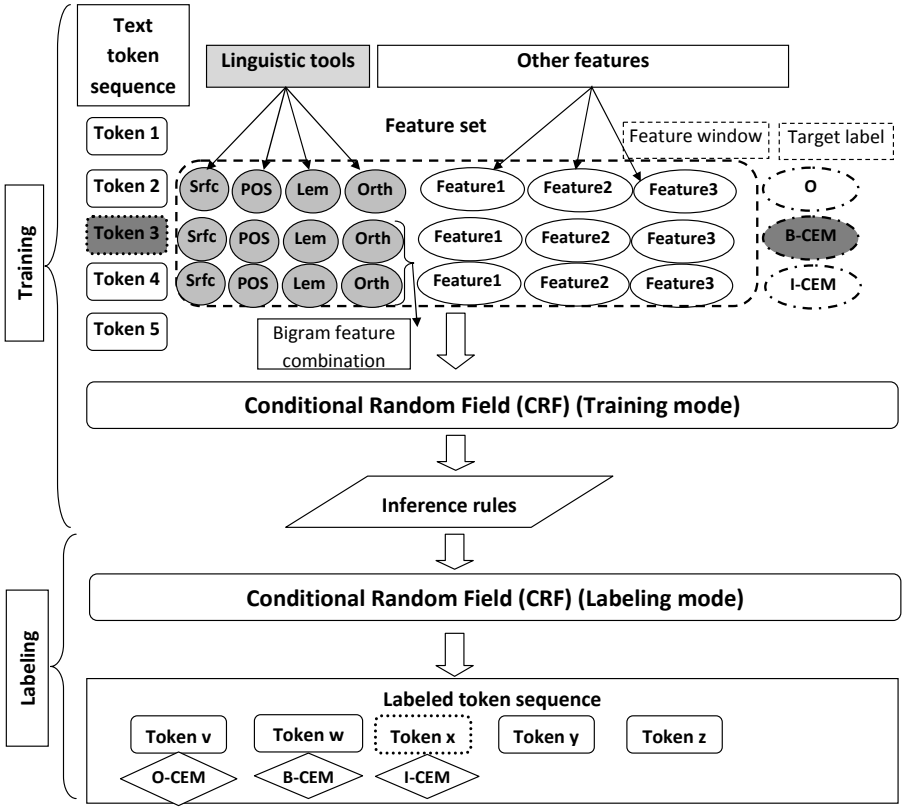

Fig. 2. Outline of the CRF model.
Srfc=token surface, POS=POS tag, Orth=orthogonal feature, Lem=lemmatization, CEM is target label.

3.2 System Implementation

Figure 3 shows an overall activity diagram for the system. In our system, in addition to the linguistic features, we use the results of the chemical NER tools for the CRF feature set. For the feature template, we use a template that is compatible with the CoNLL 2000 shared task and the Base-NP chunking task [23]. We use unigram, bigram, and trigram feature combinations. This template can handle a large number of features for

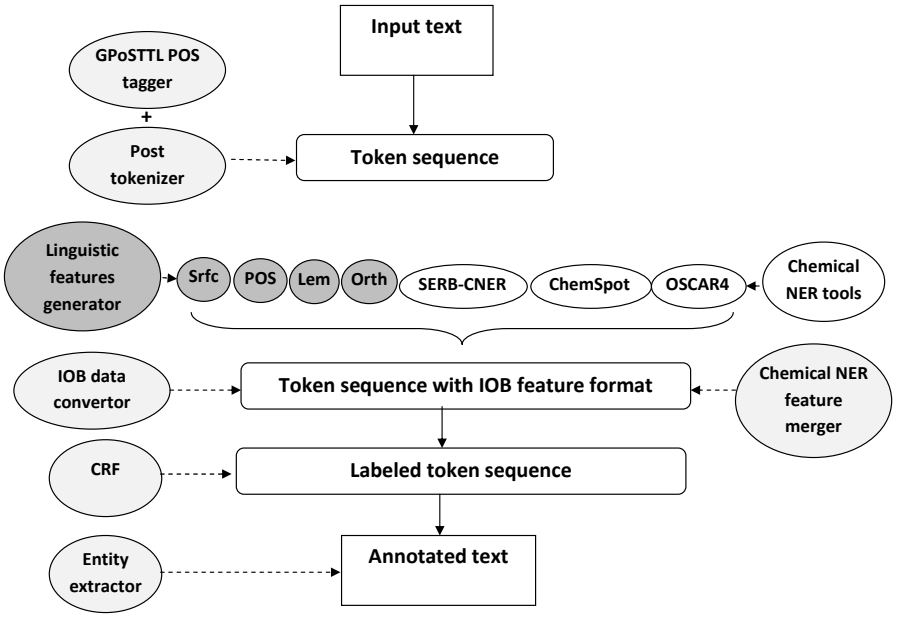

Fig. 3. A system overall activity diagram
POS=POS tag, Orth=orthogonal feature, Lem=lemmatization.

one element. Table 1 shows an example of training data for CRF. The system uses CRF++ (Version 0.58) [24], an implementation of CRF, as a tool for the sequence-labeling task. The features of CRF++ are:

- Surface symbol: symbol used to represent a term.
- Part-of-speech (POS) tag: result from the GPoSTTL tagger (Version 0.9.3) [25].
- Lemmatization: symbol that is a result from the POS tagger.
- Orthogonal feature: identified using regular expressions based on the POS tag.
- SERB-CNER tag: output of the SERB-CNER system in IOB format.
- ChemSpot tag: output of ChemSpot (Version 1.5) [26] in IOB format.
- OSCAR4 tag: output of OSCAR4 in IOB format. (We use the output of OSCAR4-related ChemicalTagger (Version 1.3) [27].)

Orthogonal features are defined in [20]. The tokenization mechanism and merging chemical NER results are discussed in detail in Section 3.3. Confidence values for the extracted terms are calculated based on the CRF output. The confidence values for multiple terms are calculated by multiplying of confidence values for all values of "B" and "I".

3.3 Tokenization Mechanism

In a sequence-label task, a certain tool returns the labeling result as a feature of each token, word boundaries of the recognized named entities are defined by using tokenization

Table 1. A sample training data for CRF

Token	POS	Lem	Orth	SERB-CNER	ChemSpot	OSCAR4	CEM
chemical	NN	chemical	Lowercase	O	O	O	O
compounds	NNS	compound	Lowercase	O	O	O	O
,	,	,	Comma	O	O	O	O
including	VVG	include	Lowercase	O	O	O	O
4	CD	4	DigitNumber	O	O	O	O
flavone	NNS	flavone	OtherHyphon	O	B	B	B-CEM
-	NNS	-	OtherHyphon	O	I	I	I-CEM
C	NNS	C	OtherHyphon	O	I	I	I-CEM
-	NNS	-	OtherHyphon	O	I	I	I-CEM
glycosides	NNS	glycosides	OtherHyphon	O	I	I	I-CEM

POS=POS tag, Orth=orthogonal feature, Lem=lemmatization, CEM=target label.

results. In chemical NER, parts of long complex terms are often annotated as chemical named entities. However, because it is usually not necessary to analyze the inner structure of a term in a general POS tagging task, a general POS tagger (such as the GPoSTTL tagger) tends to treat such a long complex term as one token. Similar problems in the biomedical domain have already been discussed [28].

In this study, we aim to aggregate the results of different chemical NER tools. Depending on recognition schema, each chemical NER tool has its own text tokenizer. In many cases, these tokenization schemas are inconsistent. For example, in Abstract 23122060 of the BioCreative IV, CHEMDNER corpus, “d-glucose” is tokenized as one entity by the POS tagger and OSCAR4 tokenizer and labeled as a chemical by OSCAR4, whereas the ChemSpot tokenizer considers only “glucose” as a chemical entity. Because of this discrepancy, this result from ChemSpot cannot be matched with the POS tagger tokenization. Figure 4 illustrates this case of inconsistent tokenization.

To solve this problem, it is necessary to apply particular tokenization techniques to generate a greater number of tokens. This will achieve better labeling results, and has the advantage of being highly consistent [29].

We have analyzed the matching ratio between the tokenization of the text by GPoSTTL and chemical entities, and the drug-name boundaries in the annotation results of other chemical NER tools, including the “gold standard” manual annotation of the BioCreative IV, CHEMDNER corpus. The tokenization of GPoSTTL did not achieve a high matching ratio, particularly with the “gold standard” annotation of the BioCreative IV, CHEMDNER corpus. A low matching ratio between POS tagger tokens on the one hand and results from chemical NER tools and “gold standard” annotation on the other will cause inappropriately noisy training data. For example, unmatched results from chemical NER tools will not be labeled correctly (either unlabeled or loosely labeled as a chemical entity). Therefore, the performance will not be satisfactory [30].

To handle this issue, we analyzed the tokenization schemas of chemical entities and drug names annotated by the chemical NER tools including the “gold standard” annotation of the BioCreative IV, CHEMDNER corpus. We found that ChemSpot and the

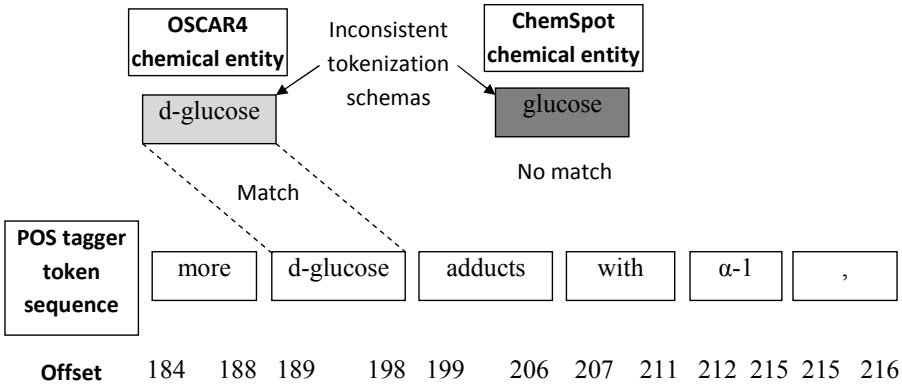

Fig. 4. Inconsistent tokenization schemas.

"gold standard" annotation of the BioCreative IV, CHEMDNER corpus tend to tokenize chunks of text containing "–", "+" and " / " into multiple tokens using these elements as separators. Because GPoSTTL does not decompose such chunks, the boundaries of chemical entities and drug names in the output of other chemical NER tools cannot be matched correctly. To solve this problem, it is necessary to generate a greater number of smaller tokens to adapt to such mismatches in tokenization schemas and to achieve better labeling results.

We implemented a post-tokenization mechanism for the POS tagger by adding three new tokenization rules for post-tokenization processing using the GPoSTTL tagger. We partitioned chunks containing "–", "+" and " / " into different tokens as for other chemical NER tools. We then checked the matching ratio after this post-tokenization process on the three datasets of the BioCreative IV, CHEMDNER corpus (training dataset, development dataset and test dataset). Table 2 shows the results of this analysis.

Table 2. Tokenization matching ratio analysis

	Training dataset		**Development dataset**		**Test dataset**	
	GPoSTTL	Post-tokenization	GPoSTTL	Post-tokenization	GPoSTTL	Post-tokenization
ChemSpot	0.93	1	0.92	1	0.92	0.99
OSCAR4	0.99	0.99	0.99	0.99	0.98	0.98
Gold standard	0.87	0.99	0.88	0.99	0.88	0.99

It is clear from Table 2 that the matching ratio between the POS tagger tokens and the chemical entities and drug names boundaries has increased considerably. For the "gold standard" annotation data, we achieved a matching ratio of 0.99. This matching ratio has increased the performance of the overall system.

4 Experiments and Discussion

4.1 First Experiment: Evaluation of the Ensemble-learning Approach and Post-tokenization Mechanism

The goal of the first experiment was to compare the performance of the ensemble-learning approach with a simple domain-adaptation approach that used only one chemical NER tool combined with CRF, on the BioCreative IV, CHEMDNER corpus. In addition, we wanted to check the effectiveness of post-tokenization on the performance. The BioCreative IV, CHEMDNER corpus contains three datasets (training, development and test). Each of the training and development datasets contains 3,500 abstracts, and the test dataset contains 3,000 abstracts.

We compared the system performance of the ensemble-learning approach before and after post-tokenization to evaluate the effectiveness of the post-tokenization process. We also compared the ensemble-learning approach with the results of a simple domain-adaptation approach that used CRF plus one of the chemical NER tools at a time (ChemSpot, OSCAR4 and SERB-CNER), together with post-tokenization. We used ten fold cross-validation on the combined three sets of the BioCreative IV, CHEMDNER corpus (training, development and test). In each fold, we used 90% of each of the three sets as training data and the remaining 10% as test data. We measured the performance using both macro- and micro-averages for precision, recall and F-score. The macro-average uses the performance of each abstract in the test dataset for calculating the average for all test data, whereas the micro-average uses all abstracts as one document for calculating the performance. Table 3 shows the macro-average and micro-average results for the ten fold cross-validation.

Table 3. Average system performance on the BioCreative IV, CHEMDNER corpus

	Macro-average			**Micro-average**		
	Precision	Recall	F-score	Precision	Recall	F-score
SERB-CNER+CRF	85.31	69.52	74.23	89.24	68.15	77.28
ChemSpot+CRF	85.26	76.77	78.84	88.10	76.21	81.72
OSCAR4+CRF	86.00	76.41	78.88	88.65	74.67	81.06
Ensemble	78.72	70.83	72.72	82.26	70.86	76.13
Ensemble/p.tok	86.62 $*	79.46 $*#	81.13 $*#	88.76 *	78.60 $*#	83.37 $*#

CRF: Conditional Random Field.
Ensemble = (SERB-CNER+ChemSpot+OSCAR4+CRF) without post-tokenization.
Ensemble/p.tok = (SERB-CNER+ChemSpot+OSCAR4+CRF) with post-tokenization.
Underlining indicates significant values for the ensemble system compared with the performance before post-tokenization. A dollar sign ($) indicates a significant value compared with SERB-CNER combined with CRF. An asterisk (*) indicates a significant value compared with ChemSpot combined with CRF. A hash (#) indicates a significant value compared with OSCAR4 combined with CRF. All significance measures were at the 0.05 level ($P < 0.05$).

Considering Table 3, we can observe the following:

- Tokenization considerably affects the performance. Comparing the performance of the system before and after the post-tokenization process, it is clear that tokenization of text by the POS tagger can significantly affect the annotation of chemical entities and drug names. The use of a chemical-oriented POS tagger can improve the system performance because it can overcome some of the tokenization mismatches that can occur between normal text and chemical entities.
- Our system (the ensemble-learning approach with CRF) has, in general, obtained better F-scores than any of the simple domain-adaptation approaches. The system clearly outperforms the original chemical NER tools. There might be some discrepancies between the definitions of what is considered a chemical entity by different recognizers. However, we find that the use of orthogonal features has helped to reduce the effect of this problem by enabling the CRF system to learn rules that include both lexical and chemical tags. We found that Ensembling different chemical NER tools with different characteristics and different annotation criteria could leverage the performance because each tool can add new information to the system.

4.2 Second Experiment: Use of the Ensemble-learning Approach for a Well-tuned Rule-based Chemical NER

The goal of the second experiment was to check the ability of the ensemble-learning approach to leverage the performance of a well-tuned rule-based system for a specific task.

To investigate this effectiveness, we used one of the best performing rule-based chemical NER systems in the official BioCreative IV, CHEMDNER task, namely LeadMine [31]. LeadMine is a grammar- and dictionary-driven approach to chemical entity recognition. We asked the developer of LeadMine to provide the results data officially used for the BioCreative IV, CHEMDNER task and used this data for this experiment. In the experiment, we added the output of LeadMine as a feature, in addition to the features of the other chemical NER tools we discussed before. We used a ten fold cross-validation test on the BioCreative IV, CHEMDNER corpus. Because LeadMine was tuned using the training and development datasets of the corpus, it was not appropriate to use these datasets in the evaluation. Therefore, in each fold, we trained both systems (ensemble and LeadMine with CRF) on a combination of the full training and development datasets and 90% of the test dataset. We then tested the systems in each fold on 10% of the test dataset. Table 4 shows the macro-average and micro-average results for the ten fold cross-validation.

From Table 4, it is clear that the ensemble-learning approach slightly leverages the performance of a rule-based system tuned for a specific task. Even though the improvement is small, it is statistically significant for precision and F-scores.

It is also clear that the ensemble-learning approach can help find new rules by checking terms that can only be extracted by the CRF. Analyzing the performance of LeadMine (a rule-based system, and one of the best systems in the BioCreative IV, CHEMDNER task), we find that approximately 6% of the “gold standard” entities were recalled

Table 4. Average system performance including LeadMine on the BioCreative IV, CHEMDNER test dataset

	Macro-average			**Micro-average**		
	Precision	Recall	F-score	Precision	Recall	F-score
LeadMine+CRF	90.34	85.88	86.88	91.46	85.42	88.33
Ensemble/LeMi	<u>90.67</u>	85.97	<u>87.14</u>	<u>91.91</u>	85.63	<u>88.65</u>

CRF: Conditional Random Field.
Ensemble/LeMi = (SERB-CNER+ChemSpot+OSCAR4+LeadMine+CRF) with post-tokenization.
Underlining indicates significant values at the 0.05 level ($P < 0.05$).

by the ensemble-learning approach with CRF but not with LeadMine. However, because we apparently also lost a different 6% of the "gold standard" entities, the recall stayed almost the same, while the precision improved. The CRF could identify some entities that would not be identified by the rule-based system. For example, in Abstract 22173956 in the chunk, "heterocyclic amines" is an entity in the "gold standard" annotation. However, the word " heterocyclic" was not identified by any chemical NER tool as a chemical entity or drug name, whereas "amines" was identified as such by all rule-based tools. The CRF enables us to identify such cases by learning them from the training dataset. Table 5 illustrates this case in IOB format.

Table 5. Gold standard entity recognized by CRF.

Tkn	**B-POS**	**E-POS**	**POS**	**Lem**	**Orth**	**CNER**	**ChemSpot**	**OSCAR4**	**Lead**	**CEM**
heterocyclic	469	481	JJ	heterocyclic	Lowercase	O	O	O	O	B-CEM
amines	482	488	NNS	amine	Lowercase	B	O	B	B	B-CEM

Tkn=token, B-pos=beginning of position, E-pos=end of position, POS=part-of-speech tag, Lem=lemmatization, Orth=orthogonal feature, CNER=SERB-CNER, Lead=LeadMine, CEM=gold standard.

4.3 Third Experiment: System Evaluation using the Official BioCreative IV, CHEMDNER Test Dataset

The goal of this experiment was to evaluate our final system in terms of the official test dataset of the BioCreative IV, CHEMDNER task. We also evaluated each chemical NER tool performance. The results of this experiment can be used as a reference for comparison between our system and other systems.

We trained the system on a combination of the training dataset and the development dataset provided by BioCreative IV, CHEMDNER corpus. We tested the system using different combinations of chemical NER tools with the layouts described above (linguistic features + chemical NER-tool combinations) using the official test dataset

of the BioCreative IV, CHEMDNER corpus. Table 6 shows the performance of the various chemical NER systems on the official test dataset. For the LeadMine system, because we could only obtain the final output of the system, we show the performance as reported in the official BioCreative IV, CHEMDNER task [5]. For SERB-CNER, the performance, particularly for recall, is low because this tool uses only very simple rules to identify chemicals. These simple rules fail to generalize towards more chemical entities.

Table 6. Performance of different chemical NER systems for the official test dataset

System	Macro-average			Micro-average		
	Precision	Recall	F-score	Precision	Recall	F-score
SERB-CNER	23.79	11.37	13.42	43.95	11.26	17.93
ChemSpot	66.92	57.59	58.52	72.94	58.87	65.15
OSCAR4	42.71	62.88	47.34	40.66	62.08	49.13
LeadMine	87.25	81.41	82.72	89.25	81.48	85.19
Ensemble	87.36	78.17	80.68	89.41	77.47	83.01
Ensemble/LeMi	**90.09**	**85.09**	**86.34**	**91.52**	**84.85**	**88.06**

CRF: Conditional Random Field.
Ensemble = SERB-CNER+ChemSpot+OSCAR4+CRF.
Ensemble/LeMi = SERB-CNER+ChemSpot+OSCAR4+LedMine+CRF.

4.4 Discussion

In this paper, we propose a framework for using chemical NER tools as constituents for ensemble learning. Since these tools did not aim to extract drug names as chemical-related entities, they cannot extract such drug names appropriately. Therefore, we confirmed that a simple domain-adaptation method that uses linguistic features and one tool output for learning improves the performance of the automatic extraction. This result shows that it is better to use such a domain-adaptation method for chemical NER tasks that aim to extract new chemical-related entities.

We also confirmed that the ensemble-learning approach that uses multiple outputs of chemical NER tools further improves the performance and that the improvement is statistically significant. This result shows that consistent differences between target task guideline and each chemical tool can be used to construct new inference rules to add more precise annotation. In addition, the ensemble-learning approach can find new entities that a rule-based system tuned for a specific task cannot. Therefore, the ensemble-learning approach can be used to construct new rules that can be added to the rule-based system. This approach can also be used in an expansion toward different kinds of chemical-entity-related domains in the future. For example, in the nanoinformatics domain, researchers use chemicals as source materials for their experiments. It

is necessary to extract chemical entities in this domain when analyzing experimental results.

5 Conclusion

In this paper, we have discussed an ensemble-learning approach that aggregates different chemical NER tools with different characteristics and different annotation criteria. This approach combines simple domain adaption and general ensemble-learning features. We confirmed that this approach is generally promising, because each chemical NER tool can contribute some unique new findings, thereby leveraging the performance. This approach can also be used in enhancing the performance of a well-tuned rule-based chemical NER system by providing information to enable the creation of new rules. Finally, we have found that the text-tokenization method considerably affects the performance of the system.

In future work, we plan to use the ensemble-learning approach to analyze the differences between similar but not identical guidelines. Resolving such differences could support the optimization of a system to identify chemical entities in the context of a particular objective.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 24240021 and 26540165. We would like to thank Dr. Daniel M. Lowe and his team for providing the results of LeadMine.

References

1. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., Yon Rhee, S. : Big data: The future of biocuration. Nature, 455, 47-50, doi:10.1038/455047a (2008).
2. Otsuki, A., Kawamura, M. :The Study of the Role Analysis Method of Key Papers in the Academic Networks. Trans. on machine learning and data mining. Vol. 6, No.1, 3-18 (2013).
3. Kim, J. D., Ohta, T., Tateisi, Y., Tsujii, J. "GENIA Corpus-semantically annotated corpus for bio-textmining," Bioinformatics 19, i180i182, (2003).
4. SCAI corpora: http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/research-development/ information-extraction-semantic-text-analysis/named-entity-recognition/chem-corpora.html accessed Mar. 1, 2015.
5. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M, Sayle, R.A., Batista-Navarro, R.T., Rak, R., Huber, T., Rocktaschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Munkhdalai, T., Ryu, K.H., Ramanan, S.V, Nathan, S., zitnik, S., Bajec, M,. Weber, L., Irmer, M., Akhondi, S.A., Kors, J.A., Xu, S., An, X., Sikdar, U.K., Ekbal, A., Yoshioka, M., Dieb, T.M., Choi, M., Verspoor, K., Khabsa, M., Giles, C.L., Liu, H., Ravikumar, K.E., Lamurias, A., Couto, F.M., Dai, H., Tsai, R.T., Ata, C., Can, T., Usie, A., Alves, R., Segura-Bedmar, I., Martinez, P., Oyarzaba, J., Valencia, A. J. :The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics 7(Suppl 1):S2 (2015). doi:10.1186/1758-2946-7-S1-S2

6. De la Iglesia, D., Harper, S., Hoover, M. D., Klaessig, F., Lippell, P., Maddux, B., Morse, J., Nel, A., Rajan, K., Reznik-Zellen, R., Tuominen, M. :Nanoinformatics 2020 roadmap. National Nanomanufacturing Network. Amherst, MA 01003. http://eprints.internano.org/607/1/Roadmap_FINAL041311.pdf (2011) accessed Apr. 26. 2015. doi: 10.4053/rp001-110413
7. IUPAC http://www.iupac.org/, accessed Apr. 1, 2015.
8. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcntara, R., Darsow, M., Guedj, M., Ashburner, M. :ChEBI: a database and ontology for chemical entities of biological interest. Nucl. Acids Res. 36 (suppl 1): D344-D350 (2008). accessed Apr. 26. 2015. doi:10.1093/nar/gkm791
9. Jessop, D., Adams, S., Willighagen, E., Hawizy, L.,Murray-Rust, P. :OSCAR4: a flexible architecture for chemical text-mining. Journal of Cheminformatics 3, 41, (2011).
10. Rockt aschel, T., Weidlich, M., Leser, U. :ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics 28, 16331640, (2012).
11. 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, 17, (2005).
12. 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).
13. Blitzer, J., McDonald, R., Pereira, F. : Domain adaptation with structural correspondence learning. Proc. the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06), Association for Computational Linguistics, Stroudsburg, PA, USA, 120-128, (2006).
14. BioCreative IV-CHEMDNER Corpus http://www.biocreative.org/ resources/biocreative-iv/chemdner-corpus/ accessed Apr. 1, 2015.
15. Dimililer, N., Varolu, E., Altnay, H. :Classifier subset selection for biomedical named entity recognition. Applied Intelligence Volume 31, Issue 3, pp 267-2825, (2009).
16. Zhou, H., Li, X., Huang, D., Yang, Y., Ren, F. :Voting-Based Ensemble Classifiers to Detect Hedges and Their Scopes in Biomedical Texts. IEICE TRANSACTIONS on Information and Systems Vol.E94-D No.10 pp.1989-1997, (2011).
17. Ekbal, A.,Saha, S. :Multiobjective optimization for classifier ensemble and feature selection: an application to named entity recognition. International Journal on Document Analysis and Recognition (IJDAR), Volume 15, Issue 2, pp 143-166, (2012).
18. Dieb, T.M., Yoshioka, M., Hara, S. :Automatic information extraction of experiments from nanodevices development papers. Proc. The 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM2012), 4247, (2012).
19. Kolarik, C., Klinger, R., Friedrich, C.M., Hofmann-Apitius, M., Fluck, J. :Chemical names: terminological resources and corpora annotation. Proc. the Workshop on Building and Evaluating Resources for Biomedical Text Mining, Marrakech, Morocco, pp. 5158, (2008).
20. Takeuchi, K., Collier, N. :Bio-medical entity extraction using support vector machines. Artificial Intelligence in Medicine, vol. 33, Issue 2, pp. 125137, (2005).
21. Klinger, R., Tomanek, K.: Classical probabilistic models and conditional random fields. Technical Report TR07-2-013. Department of Computer Science, Dortmund University of Technology; ISSN 1864-4503, (2007).
22. McDonald, R., Pereira, F.:Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics. 6(Suppl. 1):S6, (2005).
23. CoNLL 2000 http://www.cnts.ua.ac.be/conll2000/chunking/, accessed Apr. 1, 2015.
24. CRF++ tool: http://crfpp.googlecode.com/ svn/trunk/doc/index.html?source=navbar, accessed Apr. 1, 2015.
25. GPoSTTL http://gposttl.sourceforge.net, accessed Apr. 1, 2015.

26. Chemspot1.5: http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/resources/chemspot/chemspot/ accessed May. 1, 2015.
27. chemicalTagger-1.3: http://chemicaltagger.ch.cam.ac.uk/
28. 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.
29. Leaman, R., Gonzalez, G. :BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008:652-63. (2008).
30. Yoshioka, M., Dieb, T.M. :Ensemble Approach to Extract Chemical Named Entity by Using Results of Multiple CNER Systems with Different Characteristic. Proc. the fourth BioCreative challenge evaluation workshop, vol. 2, (2013).
31. Lowe, D.M., Sayle, R.A. :LeadMine: A grammar and dictionary driven approach to chemical entity recognition. Proc. the fourth BioCreative challenge evaluation workshop, vol. 2, (2013).