Complaint and Diagnosis Extraction System Utilizing Rule-based Term Extraction System

Koichi Takeuchi
Okayama University, Japan
koichi@cl.cs.okayama-u.ac.jp

Shozaburo Minamoto
Okayama University, Japan
minamoto@cl.cs.okayama-u.ac.jp

Motoki Yamasaki
Okayama University, Japan
yamasaki@cl.cs.okayama-u.ac.jp

Team Name

oka1

Subtasks

Complaint and diagnosis task

Keywords

Rule-based term extraction, Simple rule language, CRFs

1. INTRODUCTION

In our laboratory, we have been developing a rule-based term extraction system in biomedical domain that contains hand-coded terms and domain specific suffixes, thus we construct a complaint and diagnosis extraction system utilizing the term extraction system. Since the rule-based system was constructed according to the annotated newspaper corpus [3], it must be inappropriate to direct application of the rule-based system to the target task, i.e., extraction of complaint and diagnoses expressions. To the other task that the rule-based system has expected, thus we propose a statistical learning method-based term extraction system using the rule-based term extraction system as major features.

2. COMPLAINT AND DIAGNOSIS EXTRACTION SYSTEM USING CRFS WITH RULE-BASED TERM EXTRACTION SYSTEM

Extraction of Complaint and Diagnosis expressions (after this, CD expressions) must be not easy because of their wide variety of expressions; however, most of the expressions must contain some sort of domain specific words such as “症状” and “異常”. Derivational approaches that can capture complex terms from base terms succeeded in much of previous work [1][2][4] in terminology; thus in this context, we have been developing a rule-based term extraction system on biomedical domain for a disease surveillance system. Since the target terms of the extraction system contains CD expressions but are not identical with the CD expressions, we do not apply the term extraction system to this task but we apply CRFs to capture the CD expressions utilizing the outputs of the term extraction system as a principal feature.

Figure 1 shows the overview of the proposed system. The proposed system consists of two phases, i.e., a learning phase and an applying phase. In the learning phase, the system applies a morphological analyzer ChaSen to the training text of MedNLP, and then applies a simple rule based term extraction system SRL to the segmented texts; Finally, CRFs learns CD expressions with annotated tags and the features by ChaSen and SRL. In the applying phase, the system applies CRFs to plain text to extract CD expressions with the features of ChaSen and SRL. The details of the modules are below.

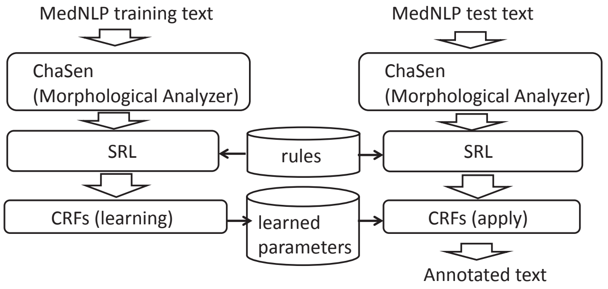


Figure 1: Overview of the proposed extraction system of complaint and diagnosis expressions

2.1 SRL for Biomedical Domain

SRL is a free software of rule-based term extraction system with graphical interface for aiding rule construction. Figure 2 shows the view of editing rules with MedNLP corpus. The lower window of Figure 2 shows the results of extracted terms with xml schema on the corpus; for example “ネフローゼ症候群” is extracted as a disease term. The rules are composed of entity rules and template rules which indicate patterns of internal term structure and external term structure, respectively. The structure of rules is below.

```
         rule  ::=  entity_rule | template_rule
template_rule  ::=  entity_rule
               ::=  | entity_rule base
               ::=  | base entity_rule
  entity_rule  ::=  base | entiry_rule base
               ::=  | base entity_rule
         base  ::=  string | character_type
               ::=  | word_list_set
```

The entity rules (i.e., indicate internal structure) are composed of combination of base rules that are strings or character types or word list sets. This indicates that each term should be composed of combination of domain specific words or morphemes. Thus we categorize words and morphemes on the basis of their meanings and then we designate entity rules to extract terms. For example, to construct an entity rule for disease terms, firstly we manually collect specific suffixes such as "症候群" and "症" that indicate disease, and then we add these morphemes to a suffix list; Second, the other specific modifiers such as "ネフローゼ" and "ダウン" are registered to a modifier list; Finally we make the entity rule based on the word lists:

Re: :− name(disease,X){list(@disease_modifiers)
list(@disease_suffixes)}.

The entity rule Re indicates that morpheme strings that the first morpheme is registered at the @disease_modifiers list and the second morpheme is registered at the @disease_suffixes list are extracted as disease terms. Thus the Re can extract "ネフローゼ 症候群", "ネフローゼ症", "ダウン 症候群" and "ダウン 症". Since SRL has optional function, thus we can designate extendable pattern to capture more longer terms such as "難治性 ネフローゼ 症候群" by adding optional modifiers to the entity rule as

Re': :− name(disease,X){optionlal(@first_modifiers)
list(@disese_modifiers) list(@disease_suffixes)}.

Where the word list @first_modifiers contains words that often come to the first position in terms to modify the meaning of the following words and morphemes.

The template rules designate external patterns of terms in texts. For example, terms occur with the following case marker "が" we can designate this using the entity rules with string as

Rt: :− name(disease,X) "が".

Currently we assume that terms occur everywhere in texts, and thus we just only designate entity rules for a template rule as below:

Rt: :− name(disease,X).

In previous work [7] entity rules and template rules are manually constructed on the 18 categories of biomedical domain.

In the 18 categories, the categories from Anatomy to Virus are domain specific, while the others are general categories. We assume that both of categories must be good features to extract the CD expressions because general categories (as well as domain specific categories) must be a clue to exclude unrelated compounds in the statistical learning method as described in the next section.

2.2 CRFs-based CD Extraction system

We applied Conditional Random Fields [5] as a CD extraction system.

CRFs evaluates a possibility of an output label sequence $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n)$ given an input word sequence $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$ by the following equations:

$$P(\mathbf{Y}|\mathbf{X}) = \frac{\exp(\Lambda \cdot \mathbf{F}(\mathbf{Y}, \mathbf{X}; \mathbf{D}))}{\mathbf{Z_X}},$$

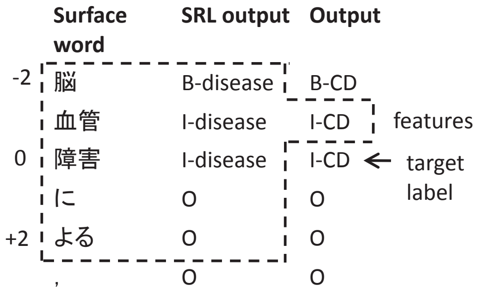
$$\mathbf{Z_X} = \sum_{\mathbf{Y}_h} \exp(\Lambda \cdot \mathbf{F}(\mathbf{Y}_h, \mathbf{X}; \mathbf{D})).$$

Where the second equation indicates that the symbol $\mathbf{Z_x}$ denotes the sum of a numerator of the first equation on all possible label candidates $\mathbf{Y_h}$. The symbol Λ denotes a parameter of CRFs that is estimated using a training corpus. This training scheme is depicted at the left-hand side module in Figure 1.

Table 1: Categories on biomedical domain and statistics of entity rules

Category	Examples	# entity rules
Anatomy	のど，血液	47
Bacteria	ボツリヌス菌	77
Chemical	塩酸，アルコール	51
Condition	完治，重傷,	368
Control	検査，摂取	38
Disease	鳥インフルエンザ	365
DNA	遺伝子	10
Outbreak	集団食中毒	25
Product	ワクチン，抗生物質	34
Protein	たんぱく質，リシン	9
RNA	リボ核酸	5
Symptom	悪寒，嘔吐	28
Virus	アデノウィルス，H5N1 型	65
Non human	サル，家畜	20
Location	東京，日本	321
Organization	厚生労働省	373
Person	教授，小学生 15 人	250
Time	2013 年，4 月	221
Total		2307

The function $\mathbf{F}$ generates a feature vector of the two arguments $\mathbf{Y}$ and $\mathbf{X}$. The function $\mathbf{F}$ is designated by the feature descriptor $\mathbf{D}$ that designates which feature should be taken into consider. We apply as features surface words, the output of SRL, and the combinations of previous and the following words. The details are depicted in Figure 3.

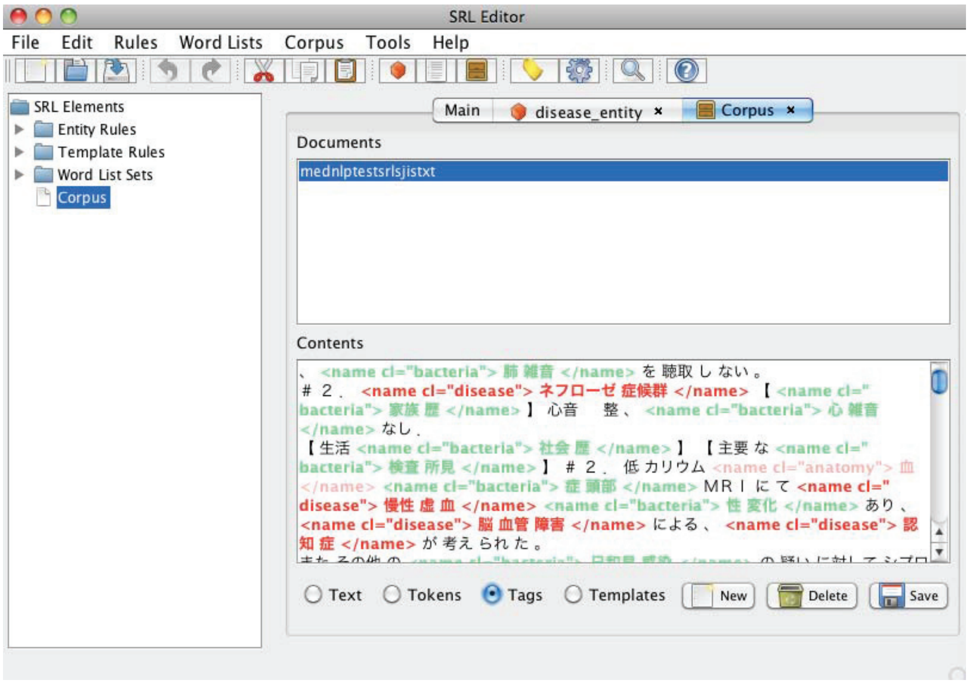


	Surface word	SRL output	Output
-2	脳	B-disease	B-CD
	血管	I-disease	I-CD
0	障害	I-disease	I-CD
	に	O	O
+2	よる	O	O
	,	O	O

Figure 3: Feature vector for an output candidate label

The first column in Figure 3 denotes the surface words and morphemes (i.e., tokens) of input texts segmented by ChaSen. The second column shows the output of SRL; the figure depicted that SRL correctly extracted the sequence " 脳 血管 障害" (cerebral vascular disease) as a disease term. The third column indicates the output label sequence of the input tokens.

In Figure 3, if we focusing on the output label "I-CD" at the third token, the features for the label are the surface



mednlptestsrlsjistxt

```
、 <name cl="bacteria"> 肺 雑音 </name> を 聴取 し ない 。
＃ ２ ． <name cl="disease"> ネフローゼ 症候群 </name> 【 <name cl="
bacteria"> 家族 歴 </name> 】 心音　 整 、 <name cl="bacteria"> 心 雑音
</name> なし .
【 生活 <name cl="bacteria"> 社会 歴 </name> 】 【 主要 な <name cl="
bacteria"> 検査 所見 </name> 】 ＃ ２ ． 低 カリウム <name cl="anatomy"> 血
</name> <name cl="bacteria"> 症 頭部 </name> ＭＲＩ に て <name cl="
disease"> 慢性 虚 血 </name> <name cl="bacteria"> 性 変化 </name> あり 、
<name cl="disease"> 脳 血管 障害 </name> による 、 <name cl="disease"> 認
知 症 </name> が 考え られ た 。
また その他 の <name cl="bacteria"> 日和見 感染 </name> の 疑い に対して シプロ
```

Figure 2: Overview of SRL editor for pattern construction on training corpus

words and SRL's outputs that are located from -2 to +2 position of the target label, and also the previous output label "I-CD" at the second token. For the implementation of this CRFs-based CD extraction system, we apply a free software crf++[1] These features are constructed using the templates of crf++ that enables us to designate complex patterns of feature vectors.

3. EXPERIMENTAL RESULTS AND DIAGNOSES

We apply two-fold cross validation to evaluate the proposed method on the training corpus. The statistics of the separated training and test set is shown in Table 2. In Table 2 "correct chunks" indicates that number of the CD expressions that are manually annotated in each corpus.

Table 2: Statistics of the two training/test corpus

	Corpus A	Corpus B
tokens	25703	25678
correct chunks	1356	1461

Evaluation measures are precision rate, recall rate and f-measure.

$$\text{precision} = \frac{\#\text{correctly detected chunks}}{\#\text{system recognized chunks}}$$

$$\text{recall} = \frac{\#\text{correctly detected chunks}}{\#\text{correct chunks}}$$

$$\text{f-measure} = \frac{2(\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}$$

The scores are calculated an evaluation script distributed by CoNLL shared task[2].

To see how the SRL feature improves the accuracy of extracting CD expressions, we set a baseline system that uses only surface words in CRFs system. Table 3 shows that the extraction results of the baseline system and the proposed extraction system. In Table 3 the baseline system performed in high precision rate; this is because most of the CD expressions would be the same expressions and then the surface-word-based system can correctly detect the same expressions in the test corpus. As for recall rate, however, the baseline system shows low performance. On the other hand, the proposed system shows high recall rate but low precision rate, and then overcomes the baseline system in f-measure score. The reason of increasing recall rate must be the manually-constructed rules of biomedical terms in SRL; the rules enable us to detect CD-expressions that might not occur in training corpus. The precision rate, however, decreases comparing with the performance of the surface-based system. This is because of (1) uncontrollable of word compositionality of rules in SRL and (2) inappropriate rules in SRL.

Table 3: Extraction results using surface with SRL's outputs

	precision	recall	f-measuer
Srf	85.4	55.4	67.2
Srf with SRL	82.9	63.3	71.8

Table 4 shows examples of CD expressions that are ex-

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html.

[2] http://ifarm.nl/signll/conll/.

tracted by the proposed system correctly but not by the baseline system; and while extracted by the baseline system correctly but not by the proposed system. As we see at the first line in Table 4, the proposed system can correctly extract compound expressions comparing with the baseline system; this is the benefit of rule-based term extraction system because of generativeness of the extraction rules. As at the second line in Table 4, however, some of the CD expressions are not extracted because of the noisy outputs of SRL; the details are below.

Table 4: Examples of CD expressions extracted by the proposed system correctly but not by the baseline system and vice versa

correctly extracted only by the proposed system	脳卒中，骨粗鬆症，神経系異常所見，甲状腺機能亢進症
by the baseline system	高血圧，糖尿病，胸部症状

The reason of missing “高血圧” is uncontrollability of SRL’s rules. The word “高血圧” is correctly recognized as *symptom* in SRL; but the case embedded in the special context, SRL recognizes the word as different category. For example, a complex sequence “高血圧確定診断名” is wrongly recognized as *person* because of the final morpheme “名” is recognized as the head suffix that indicates name of a person. This is obviously wrong because “名” indicates only a general meaning of name but “診断名” must be the correct suffix that indicates “name of diagnosis”. Besides the sequence should be composed of the two words “高血圧” and “確定診断名”. SRL recognizes the longest sequence matched in the rules as terms thus this wrong extraction has been applied.

The other reason that SRL wrongly recognize the terms’ categories is insufficient term lists. For example, the words “糖尿病” and “胸部症状” are wrongly categorized into *bacteria* because (1) the words are not registered in the rules of SRL; and (2) SRL has an inappropriate entity rule that extracts all of the word sequences consisting of two morphemes with only Chinese characters. These incorrect categories make noisy features, and thus the the proposed system decreases the precision rates comparing with the baseline system. For this problem, we have to have a rule-revision system to find the inappropriate rules in SRL.

The total f-measure score of the system with SRL features overcomes that of the system with surface words, and then the system with SRL is applied to formal run. In formal run CRFs learned from all of the training corpus i.e., both corpus A and B are applied to test corpus in formal run. Table 5 shows the result of the formal run evaluated by the task organizer[6]. All of the scores are improved comparing with the test results in Table 3. Since the rest corpus in formal run contains 22782 morphemes, the ratio of training corpus to the test corpus is 3 : 1; this ratio is three times as much training corpus as that in Table 3.

Table 5: Results of extracting CD expressions in formal run

Precision	Recall	F-measure	Accuracy
86.52	70.13	77.47	95.74

4. CONCLUSIONS

In the manuscript, we propose a CRFs-based term extraction system using a rule-based term extraction system (called SRL) as major features. The rule-based term extraction system has been manually constructed in biomedical domain with 18 categories. The proposed system is applied to complaint and diagnosis extraction task. In the preliminary experiments, we found that the CRFs-based system with SRL overcomes the baseline system with surface words in recall rate and f-measure score. In the formal run the proposed system performed 77.47 in f-measure score. In the experimental results we found that some inappropriate rules decrease precision rates of the proposed system; then in future work we will apply a term extraction system utilizing rule-based term extraction system with organized rules to CD extraction task.

5. ADDITIONAL AUTHORS

6. REFERENCES

[1] B. Daille. Conceptual sctructuring through term variations. In *Proceedings of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 9–16, 2003.

[2] C. Jacquemin. *Spotting and Discovering Terms through Natural Language Processing.* Cambridge, Mass: MIT Press, 2001.

[3] A. Kawazoe, L. Jin, M. Shigematsu, R. Barrero, K. Taniguchi, and N. Collier. The development of a schema for the annotation of terms in the biocaster disease detection/tracking system. In *Proceedings of the International Workshop on Biomedical Ontology in Action (KR-MED 2006)*, pages 77–85, 2006.

[4] T. Koyama and K. Takeuchi. Enhancing Multi-word Term Extraction for Designated Theme Embedded in a Domain Corpus. In *Proc. of The 9th International Conference on Terminology and Artificial Intelligence*, pages 73–79, 2011.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289, 2001.

[6] M. Morita, Y. Kano, T. Ohkuma, M. Miyabe, and E. Aramaki. Overview of the ntcir-10 mednlp task. In *Proceedings of NTCIR-10*, 2013.

[7] K. Takeuchi, T. Shinnou, and N. Collier. Bio-medical term extraction on simple rule language. In *Proceedings of the 3rd International Symposium on Languages in Biology and Medicine*, pages 132–134, 2009.