

SRL を利用した規則ベースの感染症用語抽出

新納 貴志[†] 竹内 孔一^{††} コリアーナイジェル^{†††}

[†] 岡山大学工学部情報工学科 〒700-8530 岡山県岡山市津島中 3-1-1

^{††} 岡山大学大学院自然科学研究科 〒700-8530 岡山県岡山市津島中 3-1-1

^{†††} 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: [†]{shinnou,koichi}@cl.cs.okayama-u.ac.jp, ^{††}collier@nii.ac.jp

あらまし 我々は感染症情報を Web 上から集めて提示する BioCaster システムを構築している。感染症情報は各国のローカルニュースに速報が出ることが予測されることから英語のみならず日本語を含めたアジア言語での開発を進めている。核となる技術は感染症に関する用語を記事から見つける用語抽出であるが、既存の手法では学習データを利用した統計的学習モデルを利用して構築した。しかしながら、新たな病気など学習データに無い用語が現れた際うまく獲得できないことが予測されるため規則に基づく用語抽出システムの構築を行う。規則ベースで用語を抽出するシステムとして SRL (Simple Rule Language) が公開されており、ユーザは語構成ならびに文脈を規則で記述することで用語を抽出できる。そこで本研究では感染症情報に必要な用語についてどのように SRL 上で定義できるかについて明らかにする。

キーワード SRL, 感染症情報, 用語抽出, BioCaster

Bio-medical Term Extraction with Morpho-Syntactic Rules on Simple Rule Language

Takashi SHINNOU[†], Koichi TAKEUCHI^{††}, and Nigel COLLIER^{†††}

[†] Department of Information Technology, Faculty of Engineering, Okayama University

Tsushimanaka 3-1-1, Okayama-shi, Okayama, 700-8530 Japan

^{††} Graduate School of Natural Science, Okayama University

Tsushimanaka 3-1-1, Okayama-shi, Okayama, 700-8530 Japan

^{†††} National Institute of Informatics 2-1-2, Hitotsubashi, Tokyo, 101-8430 Japan

E-mail: [†]{shinnou,koichi}@cl.cs.okayama-u.ac.jp, ^{††}collier@nii.ac.jp

Abstract Simple rule language, rule-based term extraction, bio-medical terms, Disease surveillance system
Bio-medical term extraction is a key technology for a surveillance system of epidemic disease news from the Web. In the previous work we applied statistical learning model to extract terms from the Web site. The previous approach is good at extracting terms with high precision rates; however it is weak at extracting new terms that do not exist in the training data. Since we usually have new disease names a new term extraction approach with high coverage for unknown or low-frequent terms is needed. Recently, Simple rule Language (SRL), a rule-based word extraction language, is freely available. The SRL also has an developing environment called SRL editor. Thus we are constructing rules of bio-medical terms on the several language (such as English, Japanese, Thai and Vietnam) for the multilingual disease surveillance system. In this manuscript we confirm how we construct rules to extract Japanese bio-medical terms from Japanese news articles.

Key words Simple rule language, Rule-based term extraction, Bio-medical terms, Disease surveillance system

1. Introduction

In this manuscript we confirm the possibility to construct

a rule-based term extraction system on bio-medical domain for BioCaster System which is aimed to automatically survey epidemic disease information from the Web documents.

BioCaster system^(注1) extracts disease-related events (i.e., who is influenced by what in which place) utilizing extended name entity extraction approaches from news texts. The found disease events are annotated to Google maps on the basis of the location name, then the user can find overview of what kinds of disease happen in which area. The final goal of this survey system to give an alert to the medical specialists for the risk of the beginning of epidemic disease. For this purpose BioCaster system deal with not only English news but also the other languages such as Chinese, Japanese, Vietnamese, Thai, French, and Spanish because disease news in the beginning stage will be described on the local news sites.

The key issue of the disease surveillance system is extracting bio-medical named entities e.g., disease names, location or area names. In previous work we applied statistical learning approaches into extracting the named entities [1] [4] [5] and showed that Conditional Random Fields performs at about 72% F-score on 18 categories named entities in WHO corpus [5]. The statistical learning approaches are promising however their performance is highly dependent on the quality and quantity of training corpus. In our previous study [5] we construct 500 balanced articles in 7 genre from the news articles; however swine influenza, which is currently widely-concerned term, did not exist in the training corpus nor taxonomy of disease. These unknown names will be not rare because there must be not small cases in which new type of virus or disease can be considered to be expanded. Besides these low frequent names (not only disease names but also location names) must be important to be extracted for finding the beginning stage of disease expansion, but the statistical approaches still miss extracting about 30% of disease names [3].

Therefore we construct a rule-based bio-medical entity extraction system intended for reducing false negative. Recently Simple Rule Language (SRL), a framework that helps to construct internal- and external- term extraction rules is freely distributed. SRL has freely modifiable editor^(注2); thus we expand the capable rules by developing the editor tool and utilizing the editor we are constructing term extraction rules in the above seven languages for building bio-medical terms. In this manuscript we report how we construct the rules of 18 categories i.e., disease name, location names and reveal the preliminary results of term extraction to WHO corpus.

2. SRL Editor

SRL (Simple rule language) is a regular expression based

language developed to perform fact extraction from plain text. It is designed to operate as a two stage process by firstly extracting a set of predefined entity classes (named entity recognition) and then using these entities to fill in template slots using contextual rules.

SRL Editor provides functions such as editor tools of entities and templates, registration of corpus, and visualization of the results of applying rules to the registered corpus. Thus we can see how the rules work well and wrong on the test corpus.

SRL consists of two types of rules such as entity and template rules. Entity rules are used for detecting named entities in text by describing intra-term characteristics such as specific suffix, morphological and character patterns. Template rules are applied after entity rules and are used for completing slots; thus template rules work as the final step of extracting terms using contextual (i.e., outside) information around terms. In the base idea of SRL, entity rules are regarded as kinds of base categories, and then a template rules are more accurate categories i.e., categories of technical terms; for example, a rule to detect the perpetrator of a crime can be defined as like perpetrator(P) :- ... name(person,P) This indicates that the base category (i.e., an entity rule) is “person” and the template rule defines “perpetrator” by narrowing “person”^(注3).

In this paper we are building entity rules for bio-medical terms. Entity rules consist of several convenient functional patterns such as dealing with literals, character-based regular expression matching functions, and registering sets of literals as a category list. In regular expression matching, suffix matching is very convenient in Japanese when we describe compact rules. For example an entity rule for disease terms in Japanese can be defined as regexp “ends (病)” which indicates that a word which contains 病 at the final character would be a name of disease. In category list we can define any sets of literals by registering all of names to a list; then we can define any categories if it is enumerable.

To construct entity rules, we have three strategies: The first strategy is to utilize internal-structure of terms by applying regular expression matching; the second is to enumerate all of the possible strings for target categories; the third is to construct combination rules which consist of lists and regular expressions. We apply the three strategies depending on the characteristics of bio-medical terms.

In the following section we describe how we construct entity rules for the target categories of bio-medical terms.

(注1) : <http://biocaster.nii.ac.jp/>

(注2) : <http://code.google.com/p/srl-editor/>

(注3) : See more details at <http://code.google.com/p/srl-editor/>.

表 1 Categories of bio-medical terms.

Category	Example
ANATOMY	のど (Throat)
BACTERIA	大腸菌 (E. coli bacteria)
CHEMICAL	グルタミン酸 (Glutamic acid)
CONDITION	完治 (Complete cure)
CONTROL	予防接種 (Vaccination)
DISEASE	喘息 (Asthma)
DNA	遺伝子 (Gene)
LOCATION	中国 (China)
NON_HUMAN	鳥 (Bird, Avian)
ORGANIZATION	中学校 (Junior high school)
OUTBREAK	集団感染 (Herd infection)
PERSON	教授 (Professor)
PRODUCT	ワクチン (Vaccine)
PROTEIN	テロメラーゼ (Telomerase)
RNA	リボ核酸 (Ribonucleic)
SYMPTOM	咳 (Cough)
TIME	2009 年 10 月 (October, 2009)
VIRUS	アデノウイルス (Adenovirus)

3. Entity Rules for Japanese Bio-medical Terms

The target categories of bio-medical terms are currently 18 (see in Table 1). The above categories are defined by Kawazoe et al. [2]. They also provide a term-annotated corpus^(注4) which we based on in our construction of entity rules.

All of the biomedical terms consists of (1) single word term or (2) multi-word term. Entity rules of single word terms should be registered to as lists; rules of multi-word terms should be registered as combination of lists and regular expressions. Besides, in Japanese, we use morphological analyzer ChaSen to segment texts; ChaSen outputs high quality segmentation results, however, segmentation errors exist. Thus we made complement rules for absorbing error segmentations.

3.1 Dealing with Single Word Terms

Single word terms or terms that we have no clue to make term-internal patterns for extraction should be registered to term lists in entity rules. In this section how we collect single word terms.

One of the resources for bio-medical terms is the annotated corpus made in previous work. The other resources we elaborate to collect terms are documents in the Web. For example, "Location" names such as country name, and prefecture, city and town name in Japanese. We found in the Wikipedia country name list^(注5), prefecture name list, capi-

(注4) : In the corpus the above 18 categories of bio-medical terms are manually annotated.

(注5) : Country: <http://ja.wikipedia.org/wiki/国の一覧>, Prefecture: <http://ja.wikipedia.org/wiki/都道府県庁所在地> Cities, Towns and vil-

表 2 Word lists for multi-word terms.

List name	Explanation and example
cap	modifier (e.g. 高 (highly))
tie	suffix (e.g. 性、型)
body	word about body (e.g. 肝臓 (liver))
symptoms	word about symptom (e.g. 出血 (hemorrhage))
animal	creature(non-human) name (e.g. 鳥 (avian))
place	country name (e.g. トルコ (Turkey))
virus	virus name (e.g. アデノウイルス (adenovirus))
bacterium	bacterium name (e.g. 大腸菌 (coli bacteria))
disease	disease name (e.g. 喘息 (asthma))
others	others
shoe	～病 (disease), ～感染 (infection) etc.

tal city list, and towns and villages. We have currently 192 country names (with their aliases), 47 prefecture names in Japan, 783 cities, 800 towns and 191 villages.

How to register the single words to SRL is shown in the next section.

3.2 Multi-word Terms

Multi-word terms consists of several morphemes thus we can define the entity rules by describing (1) single content words or modificational morphemes at lists, and (2) combination rules based on (1). The problem issue is how we can construct entity rules for multi-word terms to absorb variations with compact descriptions.

List for single word term as elements of multi-word terms

We construct lists for content words for target categories such as CONDITION and DISEASE. For example, "完治" (complete cure) is a key content word as well as a term of CONDITION. We register the word to a list named as "condition", , whose description on SRL is @condition. After this, @condition can be freely available for pattern matching scheme. Then using @condition SRL can define CONDITION as entity rule below:

- R1 :- name(condition,X) { list(@condition)}

This formula indicates that X is a term in named entity (i.e., category) CONDITION if X matches to a word in @condition list.

A Multi-word terms consist of several modifiers and content words, thus we construct lists depending on their functionality of morpho-semantic characteristics. For DISEASE category, we assumed 11 types of word lists as in Table 2

The key technology of this word lists is we organized morphemes on the basis of the type of modification in compound nouns. To do this we checked about 730 disease names that appear in the term-annotated corpus and disease names in Wikipedia site. The entity rules of combination of these word

ages: <http://ja.wikipedia.org/wiki/全国市町村一覧>

表 3 Disease term whose head is @sidease, sympton or infection.

.. 鳥 インフルエンザ (avian influenza)
Influenza: disease name
Avian influenza : disease name
.. 小児 まひ (infantile paralysis)
Paralysis: symptom (not disease name)
Infantile paralysis : disease name
.. レジオネラ 感染 (legionella infection)
Infection: shoe (not disease name)
Legionella infection : disease name

lists are described in the next section.

Combination rules Combination rules are mainly applied to extraction of DISEASE terms because most of the disease name are compounds and they have some sort of combination rules. In previous section we construct @disease list such as 喘息 (asthma) and インフルエンザ (influenza). Utilizing these lists we can construct entity rules for extracting compound nouns such as 鳥 インフルエンザ (avian influenza) or 高 病原 性 鳥 インフルエンザ (highly pathogen-ic avian influenza).

The entity rule for インフルエンザ and 鳥 インフルエンザ can be described in a compact entity rules by *optional* tag.

- R1 :- name(disease,X) {optoinal(@cap) list(@disease)}

This rule can detect both influenza and avian influenza.

The word consists of the following categories: 高 (cap) 病原 (cap) 性 (tie) 鳥 (avian) インフルエンザ (disease). Then the extraction rule is described on the basis of the sequence of word lists:

- R1 :- name(disease,X) { list(@cap) list(@cap) list(@tie) list(@animal) list(@disease) }

The entity rule indicates that X would be DISEASE if X matches to the sequence of the word lists as braces. Combination of word lists enables the entity rules to have wide coverage of modification patterns.

Most of the disease entities have disease word at the head (i.e., the end of compounds); but some of disease entities have not disease word at the head. The entity rules can describe using the word lists. Table 3 shows the examples of disease names which have disease, sympton and infection word at the head.

Thus we can define the entity rules for disease in the following conditions:

- if the head word belongs to @disease,
- if the previous word of the head word belongs to @disease,
- if the head word belongs to @symptom and the previous word belongs to @body.

Table 4 shows that the number of entity rules for each category. DISEASE is the largest number of all the entity

表 4 Number of entity rules.

CATEGORY	# Entity Rules
ANATOMY	24
BACTERIA	52
CHEMICAL	10
CONDITION	9
CONTROL	19
DISEASE	327
DNA	10
LOCATION	263
ORGANIZATION	237
OUTBREAK	14
PERSON	123
PRODUCT	29
PROTEIN	7
RNA	5
SYMPTOM	15
TIME	204
VIRUS	36
TOTAL	1395

rules.

3.3 Complement Rules for Morphological Analyzer

Because of the error segmentation results of ChaSen, we have sometimes wrong morphological sequences.

- パー キン ソン 病 (Parkinson disease)

However “パー”(Par) and “キン” (kin) do not exit in word lists because these are not words. Thus, we register them as follows:

- R1 :- name(disease,X) { “パー” “キン” “ソ” “ン” list(@shoe)}

The numbers of complement rules are 70.

4. Preliminary Experimental Results

We applied our entity rules to extract bio-medical terms from 200 articles in manually annotated corpus [2]. Table 5 shows genres of articles in the test corpus. Table 6also shows

表 5 Genres of test corpus.

Genres	# Articles
Business	20
Health	99
Politics	20
Science	20
Society	21
Sport	20
Total	200

that statistics of bio-medical terms in the test corpus. We constructed entity rules by checking the above corpus then this experiment is a kind of inside-data evaluation. The aim

表 6 Statistics of bio-medical terms in test corpus.

CATEGORY	TERM
ANATOMY	151
BACTERIA	154
CHEMICAL	95
CONDITION	353
CONTROL	110
DISEASE	508
DNA	12
LOCATION	945
NON_HUMAN	105
ORGANIZATION	855
OUTBREAK	105
PERSON	1436
PRODUCT	60
PROTEIN	29
RNA	3
SYMPTOM	229
TIME	812
VIRUS	147
TOTAL	6229

表 7 Precision and recall rates for test corpus.

CATEGORY	Precision	Recall
ANATOMY	63.09(147/233)	97.35(147/151)
BACTERIA	77.78(145/189)	94.16(145/154)
CHEMICAL	82.43(61/74)	64.21(61/95)
CONDITION	58.79(341/580)	96.6(341/353)
CONTROL	38.96(97/249)	88.18(97/110)
DISEASE	76.27(479/628)	94.29(479/508)
DNA	39.29(11/28)	91.67(11/12)
LOCATION	72.39(750/1036)	79.37(750/945)
NON_HUMAN	92.61(213/230)	94.67(213/225)
ORGANIZATION	75.37(748/1181)	87.49(748/855)
OUTBREAK	63.34(101/134)	96.19(101/105)
PERSON	61.16(899/1470)	62.6(899/1436)
PRODUCT	33.13(53/160)	88.33(53/60)
PROTEIN	72.97(27/37)	93.1(27/29)
RNA	100(3/3)	100(3/3)
SYMPTOM	63.28(193/305)	84.28(193/229)
TIME	67.49(600/889)	73.89(600/812)
VIRUS	58.79(107/182)	72.79(107/147)
Average	65.39(4975/7608)	79.87(4975/6229)

of this experiment is (1) to check the entity rules can detect terms as designed and (2) to see how many the entity rules detect false positive entities. Since the entity rules will extract all of the strings that matches the constructed patterns, it is a key issue not to extract false positive entities. The precision and recall rates for categories are shown in Table 7. According to the recall rates in Table 7 the entity rules work well except for PERSON and CHEMICAL. These two categories are difficult to construct rules because highly ambiguous. For example, some of PERSON's name

are used as LOCATION names; then we have to construct template rules to distinguish contextual information. While precision rates, we believe our entity rules work well because the precision rate is about 76% for DISEASE, which is a key category of our disease surveillance system with high recall rate (about 94%). These results show that the rule-based term extraction is promising because of less false negative. This indicates that new type disease name such as swine influenza will be detected with high possibility.

5. Conclusion

In this manuscript we describe how we construct rule-based term extraction system for bio-medical terms on Simple Rule Language. The key technology to construct entity rules is how we describe intra-term patterns by organizing morpho-semantic functions of morphemes in compounds. The results of preliminary experiment of bio-medical term extraction show that our rule-based term extraction approach is highly promising.

In the future work we compare the above results with those of statistical-learning-approach-based term extraction system.

文 献

- [1] Nigel Collier and Koichi Takeuchi. Comparison of character-level and part of speech features for name recognition in biomedical texts. *Journal of Biomedical Informatics*, (37):423–435, 2004.
- [2] 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.
- [3] Kazuya Okada. *Filtering Articles and Extraction of Epidemic Disease Terms For Disease Surveillance System*. PhD thesis, Master Course Thesis. Graduate School of Natural Science and Technology Okayama University, Japan, 2007. (in Japanese).
- [4] K. Takeuchi and N. Collier. Bio-medical entity extraction using support vector machines. *Journal of Artificial Intelligence in Medicine*, (33, Issue 2):125–137, 2005.
- [5] Koichi Takeuchi, Kazuya Okada, Ai Kawazoe, and Nigel Collier. Towards construction of epidemic disease surveillance system using statistical learning approach. In *Proceedings of The Thirteenth Annual Meeting for The Association of Natural Language Processing*, pages 578–581, 2007. (in Japanese).