Medical Information Extracting System by Bootstrapping of NTTDRDH at NTCIR-10 MedNLP Task

Yuji Nomura, Takashi Suenaga, Daisuke Satoh, Megumi Ohki, Toru Takaki
NTT DATA Corporation
Research and Development Headquarters
Toyosu Center Bldg. Annex, 3-9, Toyosu 3-chome, Koto-ku, Tokyo, Japan
{nomurayu,suenagatk,satoudic,ookimg,takakit}@nttdata.co.jp

ABSTRACT

We participated in a complaint and diagnosis task of MedNLP in NTCIR10. We extracted words of complaint/diagnosis by using a hybrid approach with bootstrapping and pattern matching with a medical term dictionary. It was possible that part of the complaint's or diagnosis's expressions are present in the extracted words. Therefore, our system concatenated the extracted words and their surrounding words by heuristic rules and determined the final complaint's or diagnosis's words. And our system estimated the modality attribute of the extracted complaint/diagnosis by heuristic rules also.

Team Name

NTTDRDH / NTTD

Subtasks

complaint and diagnosis task

Keywords

bootstrapping, information extraction, semi-supervised learning

1. INTRODUCTION

Medical documents have quickly shifted to electronic files from paper, and there are high expectations for technology to process the information in medical documents. If we can extract the medical history of a customer who is going to take out insurance, we can hope to increase the efficiency of examinations.

Therefore, we aim to establish technology which automatically extracts a patient's complaint and diagnosis from a medical history document. We propose a method to extract words of complaint/diagnosis based on bootstrapping[3, 4] which is one of methods for lexical acquisition.

2. RELATED WORK

Bootstrapping is one of the methods to achieve a vocabulary from a small amount of training data. This method alternately repeats extracting the target word(hereafter "instance") and generating the extracting pattern(hereafter "pattern"). This produces a large number of instances from a small number of correct instances.

Bootstrapping can be divided into the following two groups on the basis of the extraction target.

- extracting pairs of words in a particular relationship (binary lexical relations)
- extracting words belonging to a particular category (unary lexical relations)

Currently, Espresso[3] is well known as an effective method for extracting pairs of words in a particular relationship. Bootstrapping has a problem, called semantic drift which transits to unrelated instances, if a polysemous instance has been extracted as the iteration proceeds. Espresso has a function that can suppress semantic drift. In particular, this suppression has been achieved by introducing a confidence score function using PMI(pointwise mutual information). Espresso is intended to extract pairs of words in a binary relation such as has-a and is-a. But, our target words, which are complaint/diagnosis in medical history documents do not necessarily exist as pairs. Even if a pair of words exists, the kinds of word differ in every sentence. Therefore, Espresso can not exhaustively extract all words of complaint/diagnosis.

Thelen et al.[4] proposed Basilisk as a method to extract words that belong to a particular category. This method extracts all patterns that match the correct words in the corpus. They introduced their own score function based on heuristics to score the extracted patterns and the instances extracted by the patterns.

Extracting complaint/diagnosis in a medical history documents is equivalent to extracting words belonging to a particular category. But we propose a method based on Espresso as with Komachi et al.[1] in order to focus on the suppression of the semantic drift.

3. PROPOSED APPROACH

There are various expressions used in complaint/diagnosis in medical history documents. Therefore, we can not register all variation of complaint/ diagnosis terms into a dictionary in advance.

On the other hand, if we extract the target word by machine learning like bootstrapping, it is difficult to cover all extracting patterns in detail with a limited number of medical history documents. And, sometimes a sentence contains only complaint/diagnosis word. In this case, it is not possible to extract the word by examining the patterns around the target word.

Then, we extract words of complaint/diagnosis by using hybrid approach with bootstrapping and pattern matching with a medical term dictionary. After extracting words, we

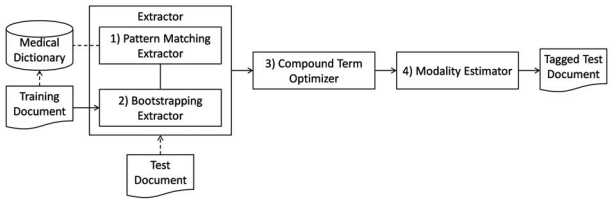

Figure 1: System Architecture

Table 1: Extract Template Definition

TemplateName	[head] definition	[tail] definition
Extract Template1	⟨morph⟩ ⟨morph⟩	⟨morph⟩ ⟨morph⟩
	⟨morph⟩ ⟨morph⟩ ⟨morph⟩	⟨morph⟩ ⟨morph⟩ ⟨morph⟩
	⟨phrase⟩	⟨phrase⟩
	⟨phrase⟩ ⟨phrase⟩	⟨phrase⟩ ⟨phrase⟩
Extract Template2	⟨morph⟩ ⟨morph⟩	⟨morph⟩ ⟨morph⟩
	⟨morph⟩ ⟨morph⟩ ⟨morph⟩	⟨morph⟩ ⟨morph⟩ ⟨morph⟩
	⟨phrase⟩	⟨phrase⟩
	⟨phrase⟩ ⟨phrase⟩	⟨phrase⟩ ⟨phrase⟩
		⟨morph⟩ ⟨phrase⟩

estimate the modality attribute of the extracted words by using information on parts around the words.

Our system is divided into four main modules as shown in Figure 1.

- Pattern Matching Extractor
  This module extracts complaint/diagnosis by pattern matching with a medical dictionary.
- Bootstrapping Extractor
  This module extracts complaint/diagnosis by bootstrapping.
- Compound Term Optimizer
  This module optimizes a compound term as complaint/diagnosis.
- Modality Estimator
  This module estimates the modality attribute of extracted complaint/diagnosis.

As advance preparations, we built a medical dictionary based on public data and training data. Then, the system extracted candidate complaint/diagnosis's words by Pattern Matching Extractor and Bootstrapping Extractor. Pattern Matching Extractor extracts words by pattern matching with the medical dictionary we built. Bootstrapping Extractor extracts words by bootstrapping which learns training data.

In the extracted candidates, it is possibile that part of the complaint/diagnosis's expressions are extracted. Therefore, Compound Term Optimizer concatenates the extracted words and their surrounding words by heuristic rules and determines the final complaint/diagnosis's words.

Finally, Modality Estimator estimates the modality attribute of the extracted complaint/diagnosis by heuristic rules.

3.1 Extract Information

We aim to reduce omissions in extraction in pattern matching with the dictionary by using achieved patterns in bootstrapping.

3.1.1 Pattern Matching

We built a medical term dictionary based on master data of injuries and diseases (in Japanese, 傷病名マスター) provided by an information service of medical fees[1], and words of complaint/diagnosis appearing in training data. In complaint/diagnosis, common words are sometimes included. For example, 「糖尿病」and「１型糖尿病」, 「インフルエンザ」and「インフルエンザＡ型」. Therefore, we perform pattern matching in order of length of word in order to extract a longer word preferentially.

In particular, we perform pattern matching for every sentence of the target documents. For each word of the medical term dictionary, the following processes are performed in order of length of the word.

1. The system checks whether the word appears in the sentence.
2. If the word appears in the sentence, the system extracts the start and end position in the sentence of the part which was matched with the word. If the start and end position do not overlap with all words of the extracted lists, the system extracts the word and adds the positions to the extracted lists.

3.1.2 Bootstrapping

Our method is based on the Espresso algorithm. Espresso extracts pair of words that have a binary relation. But, in this case, we have to extract single words belonging to complaint/diagnosis category. Therefore, the pmi for degree of confidence is computed by the following equation:

$$pmi(i,p) = \log \frac{|i,p|}{|i,*||*,p|} \quad (1)$$

where $|i,*|$ is frequency of instance i, $|*,p|$ is frequency of pattern p, $|i,p|$ is co-occurrence frequency of instance i and pattern p.

To extract a single-relation word, we use the following extracting template to generate a pattern from an instance.

[head] ⟨instance⟩ [tail]

[head] and [tail] intend expression pattern of each position. We defined two extracting templates as a character string to include in each pattern as shown in Table 1. We hypothesized that characteristic information appears after the word of complaint and diagnosis, Therefore, in Extract Template2, we added ⟨morph⟩ ⟨phrase⟩ as a character string to include in [tail] to Extract Template1.

In the case of extracting a single-relation word, semantic drift tends to occur often. One of the reasons for this is because the pattern generated from a single word tends to be simpler than the one generated from a pair of words.

Therefore, we focused on the expression characteristic of complaint/diagnosis and developed a degree of coincidence of composition information to degree of confidence for controlling semantic drift. In particular, we multiplied pmi score by a degree of composition's coincidence of the extracted candidate instance and the correct instances. The correct instances are the source of the pattern that extracts

[1] http://www.iryohoken.go.jp/shinryohoshu/downloadMenu/

Table 2: Concat Rule for Compound Term

Target Pos	Condition	Concat Judge
名詞-数	post morph: 型, 期	Yes
	other	No
記号-アルファベット	post morph: 型, A-Z	Yes
	pre morph: A-Z	Yes
	other	No
名詞-非自立	all	No
名詞-副詞可能		
名詞-接尾-副詞可能		
名詞 (other) 接頭詞-名詞接続	target morph: #, ところ, ため, とも, %, 年, 月, 日, 疑い, 来院,, ,①-⑩	No
接頭詞-数接続	other	Yes

the candidate instance. The degree of composition's coincidence is computed based on Levenshtein distance by using the following characteristics.

- part of speech of all the words
- surface and a part of speech of ending of the words

3.1.3 Hybrid Approach

We extracted the final complaint/diagnosis by using hybrid approach with bootstrapping and pattern matching with medical term dictionary.

3.2 Optimize Compound Term

In the case of extracting words of complaint/diagnosis, we have to extract not only formal disease names but also 'Disease name + Verbal noun' like 'ＡＩＤＳ発症'. In extracting by pattern matching with a dictionary, it is not realistic to register all possible terms into the dictionary in advance. Only the name of the disease portion may be extracted also by bootstrapping. Therefore, we concatenated the extracted words and their surrounding words by rules. We made a rule that determines whether or not to concatenate by a heuristic rule based on the idea of generating a noun phrase of maximal length. The concat rules are shown in Table 2.

3.3 Estimate Modality Attribute

We estimated the modality attribute of the extracted complaint/ diagnosis by pattern matching with the words around the word of complaint/diagnosis. This is the process of pattern matching. In addition, we made various word lists as shown in Table 3.

1. If the word which matches *Family Word List* ahead of the extracted complaint/diagnosis is contained, a modality attribute is determined as *family*.
2. The following processes are carried out on the word located posterior to the extracted complaint/diagnosis in the same phrase.
   (a) If the word which matches *Suspicion Word List* is contained, a modality attribute is determined as *suspicion*.
   (b) If the word which matches *Negation Word List* or *Negation Neighbor Word List* is contained, a modality attribute is determined as *negation*.
   (c) If the word which matches *Positive Word List* or *Positive Neighbor Word List* is contained, a modality attribute is determined as *positive*.

Table 3: Modality Attribute Rule

List Name	Word List
Family Word List	息子, 娘, 親, 母, 父, 兄, 弟, 姉, 妹
Negation Word List	なし, 無し, なく, ない, せず, なかった, 認められず, 認めず, 検出されず, ふれず, 見られず, みられず, 得られず, 異なる, 消失, 否定, 陰性, ０％, 鑑別疾患, 改善, 離脱, 予防, 寛解, 罹患する
Suspicion Word List	疑い, 考え、, 考えてよい, 考えられた, 可能性, 危険性, 疑う, 疑われ, 疑った, 疑わせ, 考慮され, 否定できな, 考えにく
Positive Word List	ある, あり, あった, 認め、, 認めた, 認める, 認められ, 見られる, みられる, 見られた, みられた, 出現, 診断され, 訴え, による, 原因, 伴う, 伴った, 伴い, ともなう, ともなった, ともない, なり
Negation Neighbor Word List	(–)
Positive Neighbor Word List	改善, 陰性, 診断, (space)

3. The following processing is carried out on the dependency phrase of the extracted complaint/diagnosis, and the phrases behind the phrase.
   (a) If the word which matches *Suspicion Word List* is contained, a modality attribute is determined as *suspicion*.
   (b) If the word which matches *Negation Word List* is contained, a modality attribute is determined as *naegation*.
   (c) If the word which matches *Positive Word List* is contained, a modality attribute is determined as *positive*.
4. If modality attribute is not determined, a modality attribute is determined as *positive*.

4. EXPERIMENTS AND RESULTS

We participated in the complaint and diagnosis task of MedNLP in NTCIR10[2].

4.1 Submitted Runs

We submitted three runs for complaint and diagnosis subtask as follows.

- bs_1
  This is a hybrid approach by bootstrapping using Extract Template1 and pattern matching based on a dictionary.
- bs_2
  This is a hybrid approach by bootstrapping using Extract Template2 and pattern matching based on a dictionary.
- dic
  This is baseline method using only pattern matching based on a dictionary.

4.2 Experimental Results

Table 4 shows the official results on test sets of complaint and diagnosis subtask. Table 5 shows the experimental results on training sets. In the experiment, we evaluated in 2-fold cross validation.

Table 4: Complaint and Diagnosis Subtask Formal Run Results

Methods	Precision	Recall	F-measure	Accuracy
bs_1	79.44	77.38	78.40	94.56
bs_2	78.91	78.14	78.52	94.57
dic	80.00	76.19	78.05	94.43

Table 5: Experimental Results

Methods	Precision	Recall	F-measure
bs_1	85.73	74.61	79.79
bs_2	85.02	75.61	80.04
dic	83.71	73.54	78.29

4.3 Result Analysis

Proposed method bs_2's f-measure scores the highest in both experiments in training data and formal run. But the difference between bs_2 and dic which extracts based on pattern matching with a dictionary, is only 0.47 points, and no big effect from bootstrapping was acquired.

The f-measure of bs_2 used Extract Template2 was slightly higher than the one of bs_1 that used Extract Template1. This means that it is effective to actively use the features of the expression which is located behind the words of complaint/diagnosis.

Since we added words of complaint/diagnosis in the training data into the dictionary used in the formal run, in formal run there were more words in the dictionary as compared to the dictionary used in the experiments. As a result, in all methods, forma run's recall is higher than that of the experiments.

On the other hand, in all methods, formal run's precision is less than the experiments'. This may be due to the following two possibilities.

- Difference of context
  Even though the same word is used, in some cases it should not be extracted by the context. Therefore, it is suspected that the error of such extraction increased with the increase in number of words in the dictionary.

- Difference in the use scene of bootstrapping
  In general, bootstrapping is applied to extract words from a large corpus with a small amount of training data. But the conditions in this task are different from the general scene of bootstrapping. Because while there is a large amount of training data, the corpus amount is small in this task. Therefore, increasing the number of pieces of training data above a certain amount did not lead to the generation of the appropriate pattern, and rather, it is suspected that it generated incorrect patterns.

5. CONCLUSION AND FUTURE WORK

We proposed a method to extract words of complaint/diagnosis by using a hybrid approach with bootstrapping and pattern matching with a medical term dictionary.

As mentioned above, even though the words in the dictionary that have been registered in advance as complaint/diagnosis appear in the document, in some cases they should not be extracted by the context. To deal with such cases, further work is needed to explicitly exclude them from the candidate words for extraction by learning negative example pattern from the training data.

Moreover, further work is needed to suppress the generation of incorrect patterns in bootstrapping. It is important not to select instances which would generate a generic pattern as correct instances with the role of the training data. In selecting instances, we will apply the result of learning the pattern negative examples.

6. REFERENCES

[1] M. Komachi and H. Suzuki. Improving Semi-supervised Acquisition of Semantic Knowledge from Query Logs. *Transactions of the Japanese Society for Artificial Intelligence*, pages 217–225, 2008.

[2] M. Morita, Y. Kano, T. Ohkuma, M. Miyabe, and E. Aramaki. Overview of the NTCIR-10 MedNLP Task. In *NTCIR-10 Proceedings*, 2013.

[3] P. Pantel and M. Pennacchiotti. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120, 2006.

[4] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 214–221, 2002.