Preliminary Report of III&CYUT for NTCIR-11 MedNLP-2

Liang-Pu Chen
IDEAS, Institute for Information Industry, Taiwan
eit@iii.org.tw

Hsiang Lun Lin
IDEAS, Institute for Information Industry, Taiwan
hsianglunlin@iii.org.tw

Yan Shen Lai
IDEAS, Institute for Information Industry, Taiwan
larslai@iii.org.tw

Ping-Che Yang
IDEAS, Institute for Information Industry, Taiwan
maciaclark@iii.org.tw

ABSTRACT

We construct a supervised learning system to participate MedNLP2 task in NTCIR-11 that find the keyword out correctly at right position and normalize to identify unique id in ICD10 [4]. In our system, We pick part-of-speech tagging (POS) [1] as feature to train machine learning models based on Conditional Random Fields (CRF) [3] for named entities extraction, then construct a hierarchical classifier to determine ICD code of the terms.

Keywords

natural language processing, medical informatics, named entity recognition, machine learning, conditional random field

Team Name

III&CYUT

1. INTRODUCTION

Recently, more and more medical records are written in electronic format in place of paper. Therefore, NLP techniques is required for extracting data from electronic medical records automatically. This paper describes the system to participate MedNLP2 task in NTCIR-11 that find the keyword out correctly at right position and normalize to identify unique id in ICD10. We have participated two subtasks as follows:

- Task 1) extraction of complaint and diagnosis Task (extract complaint and diagnosis from the text)
- Task 2) normalization of complaint and diagnosis Task (give icd-10 code on complaint and diagnosis)

Many thing must be considered in these tasks, similarity of ICD10 disease names in each ICD category, difference of the same disease name between medical records. Though there is a huge difference between Chinese and Japanese text, we decide to take our experience of the Chinese NER to be implemented on Japanese. Furthermore, we must normalize different term into a single unique icd10 id. There are lots of detail that have to be overcome.

2. MATERIALS

2.1 Corpus

Our materials are provided by NTCIR-11 MedNLP task: the Dummy Patients' Medical Reports (D-Rep) and the Questions from the past State Examinations extracted from the actual past state examinations (Q-Rep).

D-rep is constructed from 'dummy' medical reports that doctors have written for their 'dummy' patients. Each medical report typically contains the chief complaint, patient' s disease history, diagnosis, treatments, clinical course, and the outcome.Q-rep, this corpus is from the actual past state examinations. The question parts and graphic parts are eliminated.

2.2 Conditional Random Field

For named entity recognition, we have employed the popular approaches that conditional random field (CRF) based term identification. CRF is a class of statistical modelling method and a type of discriminative undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data. In our system, we employ CRF++[1] toolkit.

2.3 MeCab & kuromoji POS tag toolkit

To make a CRF training models, we try to extract features from a word, such as part-of-speech are required. In order to obtain these features, we have employed MeCab[2] and kuromji[3] which are open source application for Japanese morphological analysis system.

2.4 Dictionary Resources

For Japenese support, we exploit MEDIS Byomei Master[4] database, extracting name of diseases and ICD10 codes. Then, we build an ICD code tree based on its hierarchical categories, nodes contains ICD code and corresponding disease name.

3. METHOD

We have two process in our system. The first is training, and the next is analysis. The picture as below is the structure of training model.

[1]http://crfpp.googlecode.com/svn/trunk/doc/index.html
[2]http://mecab.googlecode.com/
[3]http://www.atilika.org/
[4]http://www2.medis.or.jp/stdcd/byomei/

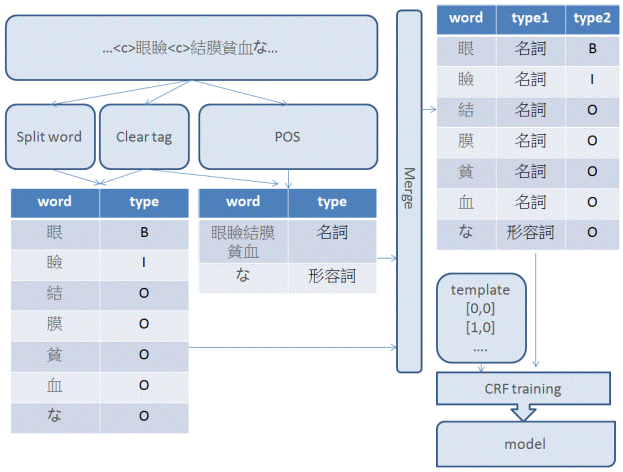

Figure 1: Structure of training model.

In training process, system will split our resource into wold table and find their speech by POS system. The system will merge them to become one table and use our template to train CRF to become our training model.

The next process is analysis. The picture as below is the structure of analysis model.

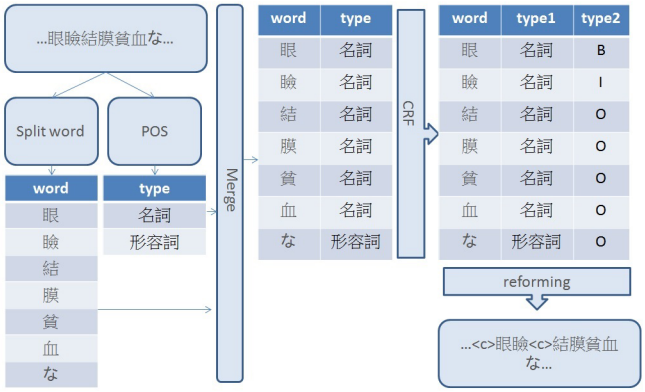

Figure 2: Structure of analysis model.

In analysis process, the system will split the context into word table and analysis the speech by POS system. Use CRF system to get the keyword and reform to new context which have keyword tag.

In process, we have two major steps, format conversion and symptom recognition.

3.1 Format Conversion

First of all, we have converted a corpus in XML format into another formatted which has IOB2[2] and POS formatted both style. In this process, we have four steps have to do.

IOB2 tag is base on keyword detective system which can tag the keyword vocabulary and detect the begin of word. The begin of word will be tagged as "B" tag and other words in the keyword will be tagged as " I" tag, and other words which are not keyword will be tagged as "O" tag. Here have an example as below.

Table 1: Structure of IOB2

Token	Speech	IOB2
4	名詞	O
mm	名詞	O
の	助詞	O
結	名詞	B
節	名詞	I
病	名詞	I
変	名詞	I
が	助詞	O

3.2 Symptom and Diagnosis Recognition

The second step, We try to approach the Task 1. In this task, we have to recognize the terms that means symptoms or diagnosis in documents. Thus, we have employed CRF machine learning model to find out all of symptoms and diagnosis in documents.

We create new feature for POS information. The features for machine learning consist of POS information and the dictionary feature extracting by dictionary. We compare with POS formatted and dictionary. We combine the previous word or next word to be a keyword in the POS formatted sequence. Search this keyword in the dictionary. We will tag it as Begin if this keyword is the first word in the sequence, tag it as End if this keyword is the last word in the sequence, tag it as N if we can not find this keyword in the dictionary, tag it as Y-name if we find this keyword in the dictionary and this keyword is a symptoms, tag it as Y-status if we find this keyword in the dictionary and this keyword is a diagnosis.

3.3 Hierarchical Classifier

In order to approach the task 2, we normalize all of the terms that we find out by previous step and mapping to ICD10 database. Since ICD code has hierarchical categories, we introduce it into our model to improve performance of classifier. We split ICD code into characters (main and sub-categories), build a tree to represent the structure. For example, a leaf of ICD10:A014 can be reach through the path: (ROOT)-A-0-1-4. When a new input bring into system, we determine its ICD code by these steps:

- 1) Begin tree tracing from root node.
- 2) Measure the distance between input and subtrees of current node, move to the nearest root of subtree.
- 3) Repeat step 2 until reach leaf, identify the ICD code of leaf as result.

In step 2, we collect name of diseases(contained by leaves) as an article for each subtree, calculate cosine TF-IDF similarity for distance measuring.

4. OFFICIAL RESULTS

In the MedNLP task, we follow the evaluation method of CoNLL-2000 shared task[5], use the accuracy, recall, precision, f-score to evaluate the result of <c> and <t> tagging.

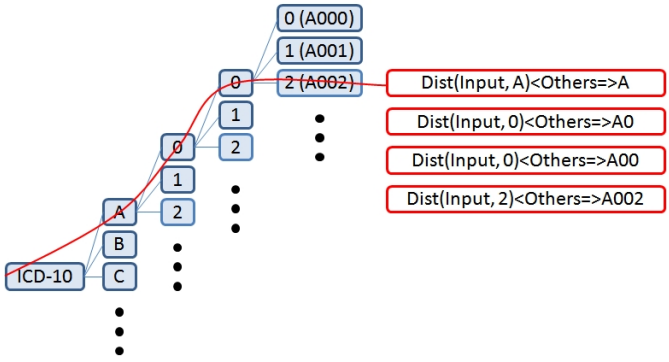

Figure 3: ICD10 hierarchical classifier

Table 2: node definition of ICD10 tree classifier

node tag	node content
A	Union of leaves in the subtree which rooted at A.
A0	Union of leaves in the subtree which rooted at A0.
A00	Union of leaves in the subtree which rooted at A00.
A000(leaf)	Disease name: アジアコレラ、真性コレラ

The result that official released are evaluated at three levels. The accuracy and recall is base on that how many c tag or t tag that our system detect and how much in the rate of correct. In f-score is the average value of accuracy and recall.

We submitted three runs as follows:

- Run 01 : The table combine two dictionaries and the the table data is base on vocabulary.
- Run 02 : The table combine one dictionary and the the table data is base on word.
- Run 03 : The table combine two dictionaries and the the table data is base on word.

Table 3: Official evaluation result with Run-01

Tag	Precision	Recall	F1-score
Positive	62.68	48.62	54.76
Family	37.21	76.19	50.00
Negation	51.86	51.57	51.71
Suspicion	6.56	7.27	6.90

5. CONCLUSIONS

In this time, we try to use three methods to solve the problem, help to find the correct condition name in Japanese. In this three methods, we combine one or two dictionaries into our training model to wish to rise the accuracy rate when the system analysis the raw context. In this three runs, the Run-01 and Run-02 has the close accuracy rate. Use this system to test the past test data. We had the 80% accuracy rate, in this time, the accuracy rate is lower than 70%. According the research, the type of data and amount of data will affect the accuracy rate. If we can have more information data to become our training model. The system can detect more information when it analysis the raw context. And it will have the higher accuracy rate than now. And the complex rate of training model have the positive effect in accuracy rate. The training model have more information that can make the higher accuracy rate.

Table 4: Official evaluation result with Run-02

Tag	Precision	Recall	F1-score
Positive	62.50	47.87	54.21
Family	37.21	76.19	50.00
Negation	51.28	51.42	51.35
Suspicion	6.67	7.27	6.96

Table 5: Official evaluation result with Run-03

Tag	Precision	Recall	F1-score
Positive	51.13	38.89	44.18
Family	40.32	59.52	48.08
Negation	51.37	50.71	51.04
Suspicion	8.51	7.27	7.84

6. FUTURE WORKS

This is our first time to participate the MedNLP task in Japanese. We have not quite experience on Japanese processing and this system is still immature. As to recognition, we will try to collect much larger corpus to improve recall and considering to add more influential features to improve precision.

In named entities recognition, we have employed POS information and dictionary-based contain matching as features. We thought these information are not quite enough to approach this task well. Therefore, we should try to employ more features for improving this system. Also in this experiment, we only have employed corpus which applying from NTCIR-11 data corpus. We thought we could look for more resource to improve our result.

We can also take other distance features for our hierarchical classifier instead of TF-IDF distance, or make the node present different content, not simply the union of leaves' content.

7. ACKNOWLEDGMENTS

This study was conducted under the "Online and Offline Integrated Smart Commerce Platform (1/4)" of the Institute for Information Industry, which is subsidized by the Ministry of Economic Affairs of the Republic of China.

8. REFERENCES

[1] R. Dearden. Structured prioritized sweeping. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 82–89, 2001.

[2] T. Kudo and Y. Matsumoto. Chunking with support vector machines. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

[3] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[4] W. H. Organization et al. The icd-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. 1992.

[5] E. F. Tjong Kim Sang and S. Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132. Association for Computational Linguistics, 2000.