Technical Report of Uni2014 in NTCIR-11 MedNLP-2 (Extraction and Normalization Task)

Kenta Fukuda
Nihon Unisys, Ltd
kenta.fukuda@unisys.co.jp

Abstract

This paper describes approach and evaluation using CRFs and dictionary matching in Task1 (Extraction of complaint and diagnosis Task) and dictionary matching in Task2 (Normalization of complaint and diagnosis Task).

Team name

Uni2014

Subtasks

・Task1 (Extraction Task)

・Task2 (Normalization Task)

Keywords

Rule-based term extraction, Simple rule language, Conditional Random Fields (CRFs), Dictionary matching

1. INTRODUCTION

Nihon Unisys is promoting information and communication technology (ICT) development based on business results of information platform construction related to medical care and health. It is important to extract the valuable data which is automatically normalized from clinical text data written in Japanese in the platform.

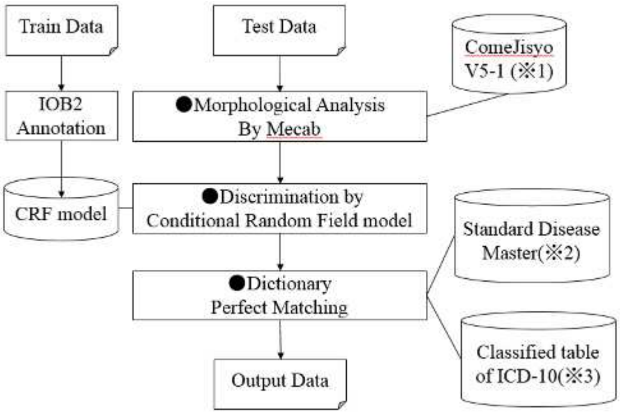
Our team participated in two subtasks, Extraction Task and Normalization Task, of Medical Natural Language Processing (MedNLP) [1,2]. This paper describes approach and evaluation using CRFs and dictionary matching, and review the availability to our medical information system solutions.

2. METHODS

2.1 Task1 (Extraction Task)

The outline methods of our design in Task1 is illustrated Fig 1.

Figure 1: Method design in Task1


Train Data
Test Data
ComeJisyo V5-1 (※1)
IOB2 Annotation
●Morphological Analysis By Mecab
CRF model
●Discrimination by Conditional Random Field model
Standard Disease Master(※2)
●Dictionary Perfect Matching
Classified table of ICD-10(※3)
Output Data

The design is divided into three main processes to convert from test data to output data as shown below.

- Morphological Analysis by Mecab

  We applied a Japanese morphological parser (Mecab) which dictionary includes ComeJisyo version 5.1(※1) to documents and segmented the sentences into tokens with part-of-speech and reading.

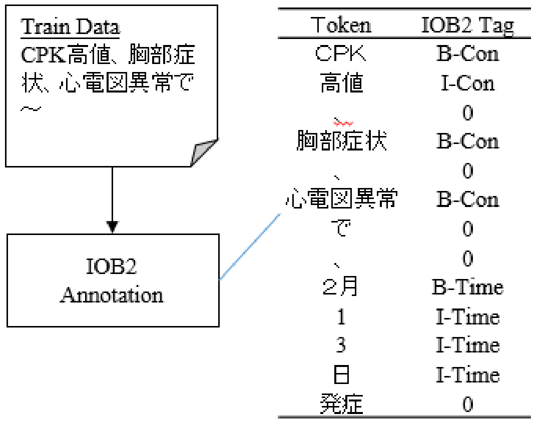
- Discrimination by CRFs model

  CRF++ is a simple, customizable, and open source implementation of CRFs for segmenting/labeling sequential data. CRF++ is designed for generic purpose and will be applied to a variety of NLP tasks, such as Named Entity Recognition, Information Extraction and Text Chunking [3].

  First of all, we converted a corpus in XML format into IOB2 format to apply CRF to the corpus.

  The IOB2 formatted data is a sequence of line, which is a pair of a segment of text and a label of I, O and B. In the case that a segment is just behind a start tag, its label is B-(element name of the beginning tag). Other segments between corresponding start and end tag are labeled I-(element name of the inside tag). Segments outside tags are labeled O (outside). The example sentence in the corpus shown Figure 2 is converted as Table 1. Number of IOB2 tag in sentence is illustrated Table 2.

Figure 2: The example sentence in the corpus



Token	IOB2 Tag
CPK	B-Con
高値	I-Con
	O
胸部症状	B-Con
、	O
心電図異常	B-Con
で	O
、	O
2月	B-Time
1	I-Time
3	I-Time
日	I-Time
発症	O

Table 1:IOB2 Format

IOB2 Tag	I (Inside)	O (Outside)	B (Beginning)
Time	I-Time	O	B-Time
Condition	I-Con	O	B-Con
Condition(Negative)	I-ConN	O	B-ConN
Condition(Suspicion)	I-ConS	O	B-ConS
Condition(Family)	I-ConF	O	B-ConF

Table 2: Number of IOB2 tag

IOB2 Tag	I	B
Time	2205	677
Condition	1046	2093
Condition(Negative)	541	1018
Condition(Suspicion)	58	110
Condition(Family)	8	74

- Dictionary Perfect Matching

  We have used the dictionary of disease name in Japanese which is published by Medical Information System Development Center (MEDIS-DC) in Japan and corresponding with ICD-10. And we also used a classified table of ICD-10 which is published by Ministry of Health, Labour and Welfare in Japan. Their resources are available on the web, shown below.

  ○Standard Disease Master(※2)

  http://www2.medis.or.jp/stdcd/byomei/index.html

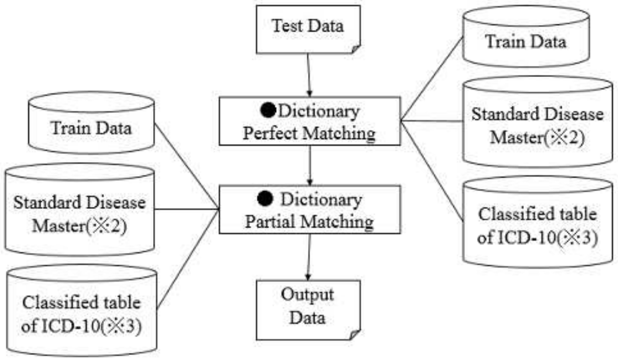
  ○Classified table of ICD-10(※3)

  http://www.mhlw.go.jp/toukei/sippei/

2.2 Task2 (Normalization Task)

The outline methods of our design in Task2 is illustrated Fig 3. The design is divided into two main processes as shown below.

Figure 3: Method design in Task2



- Dictionary Perfect Matching

  We use the dictionary of Standard Disease Master and classified table of ICD-10. If the word match the dictionary word perfectly, we covert the words into ICD-10 code.

- Dictionary Partial Matching

  We use the dictionary of Standard Disease Master and classified table of ICD-10. If the word match the dictionary word partially, we covert the words into ICD-10 code.

3. EVALUATION

3.1 Task1

The results of Task1 are shown in Table 3. We could extracted “Date and time/tense related expressions” and “Symptom and Diagnosis related expressions” in a high accuracy. On the other hand it was difficult to extract the modality types, especially “Suspicion”.

Table 3-1: Evaluation Result of Task1

Tag	Accuracy (%)
Condition	93.53

Table 3-2: Evaluation Result of Task1

IOB2 Tag	Precision	Recall	F
Time	88.14	74.53	74.53
Condition(Only)	81.83	69.38	75.09
Condition(Positive)	70.17	56.34	62.50
Condition(Negative)	51.05	54.61	52.77
Condition(Suspicion)	45.45	12.20	19.23
Condition(Family)	66.67	53.85	59.57

3.2 Task2

The result of Task2 is shown in Table 4. We used “GoldStandard+ICD” as test data which was distributed by MedNLP-2 organizer as correct answer data of Task1.

Table 4: Evaluation Result of Task2

Data	Accuracy (%)
GoldStandard+ICD	69.40

4. DISSCUSIONS AND CONCLUSION

We tried the approach of CRFs and dictionary matching in Task1 and dictionary matching Task2.

In the Task1, it was difficult to extract the modality types in a high accuracy. We think that adding more rules to complement our algorithm is needed to be in a high accuracy. We suggest that it is important to consider the local rule of each hospitals and add these rules to the algorithm to use widely as a function of medical information system, because each hospitals have each rules of recording documents.

In the Task2, we used Standard Disease Master and classified table of ICD-10 as translation dictionaries. It is necessary to improve accuracy that we should make an algorithm to be able to convert unknown word, because our system could not do the words which do not correspond to Standard Disease Master and classified table of ICD-10 partially at all.

In the conclusion, we could learn the technologies to develop computational systems for retrieving medical information from medical text documents, and think that it is very important not only to develop these supporting system but also to consider a purpose to analysis data which are extracted from those documents. Because the requirements of extracting data depend strongly on the purposes of data analysis, we should consider the system from each viewpoints. We heard that next MedNLP’s theme is “Data Mining”, so we continue to challenge the next trial to develop “one stop solution” which is covered from data extraction to data analysis.

5. REFERENCES

[1] Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. 2014. Overview of the NTCIR-11 MedNLP-2 Task. In Proceedings of the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.

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

[3] Kudo, T., Yamamoto, K., Matsumoto, Y. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. IPSJ SIG Technical Report.