HCU at the NTCIR-12 MedNLPDoc Task

Ayumu Hiramae
Hiroshima City University
hiramae@ls.info.hiroshima-cu.ac.jp

Hidetsugu Nanba
Hiroshima City University
nanba@hiroshima-cu.ac.jp

Toshiyuki Takezawa
Hiroshima City University
takezawa@hiroshima-cu.ac.jp

ABSTRACT
Our team, HCU, participated in the Phenotyping Task. We tackled the task by using machine learning techniques and edit distance by DP matching. This paper outlines the methods we used to obtain the result evaluated by the task organizer.

Keywords
conditional random fields (CRF), edit distance, DP matching.

1. INTRODUCTION
We have studied automatic construction of an ontology by extracting technical terms from various texts, such as research patents [1] and cooking recipes [2]. Our goal in NTCIR-12 [3] was to evaluate our method for technical term extraction.

This paper is organized as follows. Section 2 explains the method for extracting diseases from a given medical record and for assigning ICD-10 codes. To investigate the effectiveness of our method, we conducted experiments, as reported in Section 3. We describe related work in Section 4, and Section 5 presents our conclusions.

2. OUR METHOD
Our method consists of the following two steps.

1. Extraction of diseases from a given medical record.

2. Assignment of ICD-10 codes to diseases extracted in Step 1.

In the following sections, we describe these steps.

2.1 Extraction of Diseases from a Given Medical Record
We formalized the task as a sequential labeling problem, and employed conditional random field (CRF) as a machine learning method using the following features; (we use MeCab as a Japanese morphological analysis tool to identify the part of speech):

- a word;

- its part of speech;

- whether the word frequently appears in disease names in ICD-10[1];

- whether the word frequently appears in disease names in the MEDIS database[2].

Examples of words that appear in ICD-10 and MEDIS are shown in Table 1.

Table 1 Examples of words that frequently appear in disease databases

Database	Examples of words	Frequency in each database
ICD-10	性, 症, 炎, 部, 腫	5,181
MEDIS	皮膚 (skin), 足 (foot), 手 (hand), 膝 (knee), 腹部 (abdomen)	1,228

2.2 Assignment of ICD-10 Codes
We calculated the edit distance [4] between a disease name extracted in Step 1 and each disease name in ICD-10, and assigned the ICD code having the highest similarity.

3. EXPERIMENTS
3.1 Extraction of Diseases from a Given Medical Record
We evaluated our method using NTCIR-10 MedNLP [4] and NTCIR-11 MedNLP-2 [5]. The datasets of MedNLP and MedNLP-2 consist of 45 and 81 records, respectively. We conducted 5-fold cross validation. We evaluated our method using precision, recall, and F-measure. Table 2 shows the experimental results. As can be seen, our method obtained high precision scores in both datasets.

Table 2 Experimental results for the extraction of diseases

	precision	recall	F-measure
MedNLP	0.815	0.701	0.754
MedNLP-2	0.919	0.854	0.885

3.2 Assignment of ICD-10 Codes
We submitted one system result. The experimental results at three levels, SURE, MAJOR, and POSSIBLE, are shown in Tables 3, 4, and 5, respectively.

Table 3 Official evaluation result (SURE)

Precision	Recall	F-measure
0.209	0.364	0.266

Table 4 Official evaluation result (MAJOR)

Precision	Recall	F-measure
0.361	0.363	0.362

---

[1] http://www.dis.h.u-tokyo.ac.jp/byomei/icd10/

[2] http://www2.medis.or.jp/master/syoken/

Table 5 Official evaluation result (POSSIBLE)

Precision	Recall	F-measure
0.420	0.230	0.297

The experimental results showed that our method obtained low precision scores, which indicate that edit distance was not effective. In future, we will examine other similarity methods. We will also examine a method that takes account of words that occur in the context of diseases.

3.3 Discussion

There are two typical errors in our system, as described in Sections 3.3.1 and 3.3.2.

3.3.1 Main topics in medical records

In the following sentence, our system assigned a code "C187" to "S 状結腸癌 (sigmoid colon cancer)." However, our system should not assign a code to this string, because this medical record was written in 2005, and most of this record focuses on the operations in 2005. The following sentence is only the explanation for the patient's past disease, which is not the main topic of this record.

> １９９９年４月　S状結腸癌に対してハルトマン手術施行。

> (In April, 1999, Hartmann's operation for sigmoid colon cancer.)

Therefore, we need to detect the main topic (disease or operation) in each record, and assign the corresponding code.

3.3.2 Mismatch between automatically extracted disease names and index terms in ICD-10

In the following sentence, our system identified "腫瘍 (tumor)" as a disease, and assigned a code "D489". However, in this case, the ICD-10 code "D430" for "前頭葉腫瘍 (frontal lobe tumor)" should be assigned.

> 頭部ＣＴで右前頭葉に腫瘍を認める。

> (The tumor is found in the right frontal lobe by head CT.)

ICD-10 codes differ according to the body site in which the symptom occurs. Therefore, we need to target nouns that occur in the context of diseases because they are related phrases such as a body parts.

4. RELATED WORK

Wikification is a related task, in which the procedure of consists of the following two steps: (1) identification of concepts and (2) the linking of concepts to the corresponding Wikipedia pages. In the MedNLP task, these steps correspond to (1)' extraction of diseases from a medical record and (2)' assigning the ICD-10 codes. Illinois Wikifier[3][6] used Integer Linear Programming (ILP) for Wikification. The system maximizes the objective function assigning concepts to Wikipedia titles with relations extracted from both a document and an external knowledge base. If we consider a medical record and the external medical resource in Section 2.1 as an input document and Knowledge Base used in Wikifier, we can apply this Wikifier approach to our task; resolve the problem mentioned in Section 3.3.2.

As another approach for resolving the same problem, we will collect synonyms of diseases by using a statistical machine translation (SMT) technique. Zhou et al. [7] obtained synonyms automatically using the SMT technique. If translations of two terms X and Y are the same, then the terms X and Y are considered paraphrases. Based on this idea, they automatically obtained synonyms from a translation model, which was created from pairs of English and Chinese sentences, and they used it for evaluation of text summarization.

In other Wikification research, McNamee et al. [8] proposed a multilingual named-entity linking method, which links concepts in multiple languages onto English Wikipedia pages. In the future, following McNamee et al., we will conduct multilingual entity linking.

5. CONCLUSION

We developed a CRF-based and edit-distance-based system, and submitted the result to the Phenotyping Task. Our method obtained F-scores of 0.266, 0.362, and 0.297 at SURE, MAJOR, and MAJOR levels, respectively. We will incorporate ideas from Wikification and SMT technique to overcome the MedNLP task.

6. REFERENCES

[1] Nanba, H., Mayumi, S., and Takezawa, T. (2011) Automatic Construction of a Bilingual Thesaurus using Citation Analysis. In Proceedings of the 4th International CIKM Workshop on Patent Information Retrieval (PaIR'11), pp.25-30.

[2] Nanba, H., Doi, Y., Takezawa, T., Sumiya, K., and Tsujita, M. (2014) Construction of a Cooking Ontology from Cooking Recipes and Patents. In Proceedings of Workshop on Smart Technology for Cooking and Eating Activities: CEA2014.

[3] Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T., (2016) Overview of the NTCIR-12 MedNLPDoc Task. In Proceeding of 12th NTCIR Workshop Meeting.

[4] Morita, M., Kono, Y., Ohkuma, T., Miyabe, M., and Aramaki, E. (2013) Overview of the NTCIR-10 MedNLP task. In Proceeding of 10th NTCIR Workshop Meeting, pp.696-701.

[5] Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T., (2014). Overview of the NTCIR-11 MedNLP-2 Task. In Proceeding of 11th NTCIR Workshop Meeting, pp.147-154.

[6] Cheng, X., and Roth, D. (2013) Relational inference for wikification. In Proceedings of EMNLP.

[7] Zhou, L., Lin, C-Y., Munteanu, D.S, and Hovy, E. (2006) ParaEval: Using Paraphrases to Evaluate Summaries Automatically. In proceeding of HLT-NAACL2006, pp.447-454.

[8] McNamee, P., Mayfield, J., Lawrie, D., Oard, D. W., and Doermann, D. S. (2011) Cross-Language Entity Linking. In Proceedings of IJCNLP pp.255-263.

---

[3] https://cogcomp.cs.illinois.edu/page/software_view/Wikifier