NCU IISR System for NTCIR-11 MedNLP-2 Task

Sheng-Wei Chen
Yuan-Ze University, Taoyuan, Taiwan
s1016019@mail.yzu.edu.tw

Po-Ting Lai
National Tsing-Hua University, HsinChu, Taiwan
potinglai@gmail.com

Yi-Lin Tsai
National Tsing-Hua University, HsinChu, Taiwan
s102065514@m102.nthu.edu.tw

Jay Kuan-Chieh Chung
Yuan-Ze University, Taoyuan, Taiwan
s1003341@mail.yzu.edu.tw

Sherry Shih-Huan Hsiao
National Taiwan University Hospital, Taipei, Taiwan
sherryhsiao@ntuh.gov.tw

Richard Tzong-Han Tsai*
National Central University, Taoyuan, Taiwan
thtsai@csie.ncu.edu.tw

ABSTRACT

This paper describes NCU IISR's Japanese ICD-10 Code Linking system for NTCIR-11 MedNLP. Our system uses Conditional Random Fields (CRFs) to label ICD-10 mentions and temporal expressions. We also use CRFs to detect the modalities of the ICD-10 mentions. To resolve the problem of ICD-10 mention normalization, we use the Lucene engine to link mentions to the corresponding ICD-10 database entries. Evaluated on the MedNLP test set, our system achieved f-scores of 79.96% for ICD-10 term recognition, 67.64% for time expression and 69.4% for ICD-10 mention normalization.

Team Name

IISR

Subtasks

Task 1 (Extraction task)

Task 2 (Normalization task)

Keywords

medical informatics, machine learning, information retrieval, named entity recognition

1. INTRODUCTION

Recently, more and more medical records are stored in electronic format, which increases the importance of information processing techniques in medical fields. The NTCIR MedNLP shared task was proposed to provide a platform for developing techniques to retrieve important information from medical documents in Japanese.

For the 2014 MedNLP shared task [1], participants are expected to extract information from medical reports written by physicians and past medical exams. There are three tasks in this competition, and our IISR lab team participated in the following two:

Task 1: Extraction of Complaint and Diagnosis Task (extract complaints and diagnoses from the text)

Task 2: Normalization of Complaint and Diagnosis Task (give icd-10 codes for extracted complaints and diagnoses)

For Task 1, we developed three models to extract three types of named entity mentions, including time, ICD-10 mention and modality. All these models are based on the conditional random fields model.

For Task 2, we use the Lucene search engine[1] to index every ICD-10 mention in the ICD-10 database[2][2] and ICD-10 mentions on the MedNLP2 training set. Since there are some English ICD-10 mentions, we retrieved the ICD-10 codes and their English titles/terms from WHO ICD-10 classifications [3] as English database. We also employ machine translation tools to map them to ICD-10 codes.

The remainder of the paper is organized as follows: In Section 2, we give an overview of our system and describe its implementation, including temporal expression recognition, ICD-10 mention recognition, modality detection, and ICD-10 mention normalization. In Section 3, we detail the official results of our participation in the

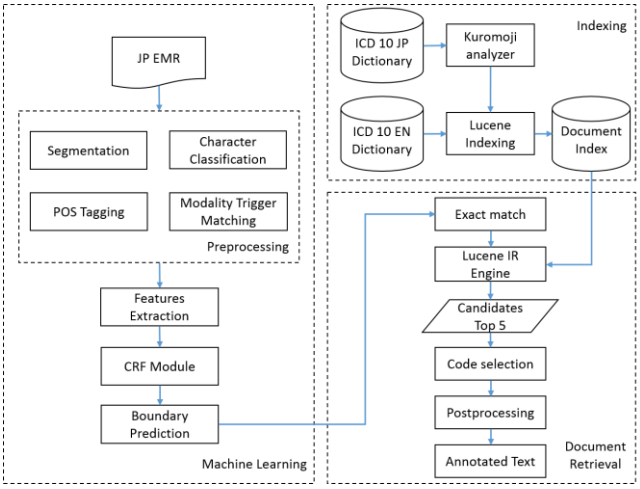

Figure 1. Flowchart of our system for Task 1 and Task 2.

* corresponding author

[1] http://lucene.apache.org/

[2] http://www2.medis.or.jp/stdcd/byomei/

[3] http://apps.who.int/classifications/icd10/browse/2010/en

C-SURF	C-TYPE	M-SURF	M-BASE	M-POS1	M-POS2	M-POS3
呼	B-CJK_UNIFIED_IDEOGRAPHS	B	B-呼吸	B-名詞	B-サ変接続	*
吸	I-CJK_UNIFIED_IDEOGRAPHS	I	I-呼吸	I-名詞	I-サ変接続	*
音	I-CJK_UNIFIED_IDEOGRAPHS	B	B-音	I-名詞	B-接尾	一般
異	I-CJK_UNIFIED_IDEOGRAPHS	B	B-異常	I-名詞	B-形容動詞語幹	*
常	I-CJK_UNIFIED_IDEOGRAPHS	I	I-異常	I-名詞	I-形容動詞語幹	*
な	B-HIRAGANA	B	B-ない	B-形容詞	B-自立	*
し	I-HIRAGANA	I	I-ない	I-形容詞	I-自立	*

Figure 2. An example of the features used in ICD-10 mention recognition.

MedNLP task. Finally, in Section 4, we present our conclusions and indicate the direction of our future work.

2. SYSTEM DESCRIPTION

Figure 1 is a flowchart that illustrates the three stages of our system. First, we recognize temporal expressions and ICD-10 mentions. Next, we classify the modality of the ICD-10 mentions. In the third stage, the ICD-10 mentions are formulated as Lucene queries and then sent to the Lucene IR engine to retrieve the corresponding ICD-10 codes.

2.1 Temporal expression and ICD-10 mention recognition

We consider recognizing temporal expressions and ICD-10 mentions in text as a sequence labeling task [6], and we develop a model based on linear chain conditional random fields [3; 9] to label mentions. We use the Kuromoji[4] system to tokenize sentences and label tokens with morphological tags. This allows us to use morphological features such as C-SURF, C-TYPE[5], M-SURF, M-BASE, M-POS1, M-POS2, M-POS3, etc. [5]. Figure 2 shows an example of the features.

2.2 Modality detection

Modality detection (MD) assigns the modalities of complaint and diagnosis information to ICD-10 codes. The four modalities are "positive", "negation", "suspicion", and "family". We formulate MD as a sequence labeling problem and apply CRFs to solve it. In addition to the features used in Figure 2, we also use a modality keyword list [7] as a feature. Figure 3 shows an example of the features we use in MD. The "なし" is in the negation keyword dictionary and the first character is labeled as "B-negation" and the second one is labeled as "I-negation".

There are two possible conflicts that may occur in MD. The first conflict occurs when a phrase is labeled as "positive", "negation", "suspicion", or "family" but not recognized by our ICD-10 mention recognition model. The second occurs when boundaries are inconsistent between ICD-10 mention and the modality terms. In case of either conflict, the modality keyword matches are discarded.

C-SURF	BIO-Fam.	BIO-Neg.	BIO-Susp.	BIO-Pos.
麻	O	O	O	O
痺	O	O	O	O
な	O	B-negation	O	O
し	O	I-negation	O	O

Figure 3. An example of the features used in modality detection.

2.3 ICD-10 mention normalization

In Task 2, ICD-10 mentions must be normalized to unique ICD-10 codes. The steps of our normalization process are summarized in Figure 4. *Candidate Selection* maps ICD-10 mentions to ICD-10 codes. If the mapped ICD-10 codes are not unique, then we apply *Disambiguation* to select unique ICD-10 codes.

Candidate Selection: In Step 1, we map each ICD-10 mention to an ICD-10 code in the database. If a mention matches exactly with an ICD-10 code, it will be assigned this ICD-10 code. If two or more matching codes are found, the *Disambiguation* process is executed. If no matching codes are found, we submit the mention string to the Lucene search engine and retrieve the top-five ICD-10 codes.

Lucene Engine: We consider each ICD-10 entry in the database a document *d*. The terms of *d* are the tokens of the entry, and the ICD-10 code is *d*'s title. After tokenizing every ICD-10 entry in the database into several tokens, we formulate each as a Lucene query. Different tokens are separated by spaces, which signify "OR" operators in Lucene query format. The rank score is calculated by BM25 [10]:

$$score(Q,d) = \sum_{i}^{n} idf(q_i) \times \frac{tf(q_i,d) \times (k_1+1)}{tf(q_i,d) + k_i \times (1 - b + b \times \frac{|d|}{|d|})}$$

$$idf(q_i) = \log\left(1 + \frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}\right)$$

where $Q = \{q_1, q_2, \ldots, q_n\}$ is the query ICD-10 mention with token length n, d is the ICD-10 mention in the dictionary, tf is the number of times that q_i appears in d, k_1 controls non-linear term frequency normalization (saturation), and b controls to what degree document length normalizes tf values. k_1 is 1.2 and b is 0.75 [4] in our experiment.

In Step 2, we first search for Japanese mentions in the Japanese ICD-10 database. If there is no matching candidate ICD code for a Japanese mention, we will translate[6] it to

[4] http://www.atilika.org/

[5] http://site.icu-project.org/

[6] https://code.google.com/p/java-google-translate-text-to-speech/

English and then search the English database. Conversely, if we find no match for an English mention in the English database, we translate it to Japanese and search for it in the Japanese database.

Disambiguation:

In Step 3, we examine the ranked list returned by the Lucene search engine in Step 2. For each mention, if the list contains ICD-10 codes which were assigned in Step 1 to other mentions in the same document, then we will assign the ICD code to it. The purpose of this step is to catch duplicate codes that doctors have written as different mentions.

In Step 4, we link the mention with the most frequent ICD-10 code in the ranked list.

In Step 5, we count the co-occurrence ICD-10 code in the same document in training set as a list. If the mention is not found any candidate code in previous steps, we will select the code with the highest frequency in the ranking list through the previous and next mention which has assigned a code.

In Step 6, we assign the most frequency ICD-10 code in whole training set.

In Step 7, we choose the top 1 candidate return from the Lucene Engine.

Candidates Selection
Step 1: Exact matching.
Step 2: The top 5 ICD-10 codes of Lucene (index field cross searching include translation).
Disambiguation
Step 3: Choose the unambiguous ICD-10 code.
Step 4: Choose the most frequent ICD-10 code in top 5 codes
Step 5: Choose the ICD-10 code with the highest co-occurrence.
Step 6: Choose the most frequency ICD-10 code in the document.
Step 7: Choose the top1 ICD-10 code.

Figure 4. The pipeline ICD-10 mention normalization processes.

3. Results and Analysis

The training corpus contains 102 documents and test corpus contains 49 documents. Details on these two independent training and test datasets are shown in Table 1.

Table 1. The distribution of datasets

Corpus	Training	Test
Documents	102	49
Sentences	3752	2071
ICD-10 mentions(<c> tag)	3304	2136
Time(<t> tag)	684	369
Positive	2075	1333
Negation	1046	705
Suspicion	108	55
Family	75	43

The performance is evaluated in terms of three metrics: precision (P), recall (R) and F-measure (F) ($\beta = 1$) [8], which are defined as follows:

$$\text{Precision} = \frac{\text{the number of correctly recognized } \textit{items}}{\text{the number of recognized } \textit{items}}$$

$$\text{Recall} = \frac{\text{the number of correctly recognized } \textit{items}}{\text{the actual number of } \textit{items}}$$

$$F_\beta = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision} + \text{Recall})}$$

The *item* can be different from time, mention, modality and ICD10-code in different evaluations.

The Table 2 shows our performance of ICD-10 mention and temporal expression recognition. Our system achieves f-score 79.96% for ICD-10 mention recognition, 67.64% for temporal expression. The Table 3 also shows our performance of modality detection. The Table 4 shows our performance of ICD-10 mention normalization. 69.4% for ICD-10 mention normalization with gold ICD-10 mention boundary. The lower ICD-10 mention normalization performance might be caused by the candidate selection step which could not return the correct ICD-10 codes from the top 5 candidates.

Table 2. The performance of ICD-10 mention & temporal expression recognition on the test set.

Composition	Tag	A	P	R	F
ALL	c	94.63	**89.39**	**72.34**	**79.96**
	t	83.94	56.64	67.64	
ALL w/o M-SURF, M-BASE	c	94.71	89.24	71.82	79.58
	t	-	-	-	-
ALL w/o C-TYPE	c	94.57	89.15	72.14	79.74
	t	-	-	-	-

Table 3. The performance of modality detection on the test set.

Composition	Modality	P	R	F
ALL	Pos.	**83.59**	**69.71**	**76.02**
	Neg.	79.47	67.81	73.17
	Susp.	78.12	45.45	57.47
	Fam.	90.62	69.05	78.38
ALL w/o M-SURF, M-BASE	Pos.	83.57	69.26	75.75
	Neg.	**79.73**	**67.81**	**73.29**
	Susp.	74.19	41.82	53.49
	Fam.	**93.55**	**69.05**	**79.45**
ALL w/o C-TYPE	Pos.	82.81	69.18	75.39
	Neg.	79.80	67.52	73.15
	Susp.	**77.14**	**49.09**	**60.00**
	Fam.	90.62	69.05	78.38

Table 4. The performance of ICD-10 mention normalization on the test set.

System	Composition	F
NE + ICD	none	39.6
	Add training set	**56.6**
Gold Standard + ICD	none	46.8
	Add training set	**69.4**

4. Conclusion

We have described our Japanese ICD code linking system, which participates in the MedNLP task 1 and task 2. The recognition and modality detection are based on the Conditional Random Fields. And the ICD-10 mention normalization uses the Lucene engine and training set ICD-10 mentions to enhance our systems on candidate selection.

We use the F1-measure to evaluate our system. And there are still many error cases caused by candidate selection step. We believe the problems are due to the limitations of the dictionaries and lack of a suitable ICD-10 term mapping strategy. In our future work, we will apply a robust candidate selection strategy to overcome the problem.

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 NTCIR-11* (2014).

[2] Center, M.I.S.D., 2012. *Hyojun Byomei Handobukku 2012 [Standard Disease Name Handbook 2012] (In Japanese)*. Shakai Hoken Kenkyujo, Inc.

[3] Lafferty, J.D., Mccallum, A., and Pereira, F.C.N., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning* (2001), Morgan Kaufmann Publishers Inc., 282-289.

[4] Manning, C.D., Raghavan, P., and Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge university press Cambridge.

[5] Miura, Y., Ohkuma, T., Tomoko, O., Masuichi, H., Yamada, E., Aramaki, E., and Ohe, K., 2013. UT-FX at NTCIR-10 MedNLP: Incorporating Medical Knowledge to Enhance Medical Information Extraction. In *Proceedings of the NTCIR-10* (2013), 728-731.

[6] Nguyen, N. and Guo, Y., 2007. Comparisons of sequence labeling algorithms and extensions. In *Proceedings of the 24th International Conference on Machine Learning* (Corvalis, Oregon2007), ACM, 681-688.

[7] Nomura, Y., Suenaga, T., Satoh, D., Ohki, M., and Takaki, T., 2013. Medical Information Extracting System by Bootstrapping of NTTDRDH at NTCIR-10 MedNLP Task. In *Proceedings of the NTCIR-10* (2013), 732-735.

[8] Rijsbergen, C.J.V., 1979. *Information Retrieval*. Buttersworth.

[9] Sha, F. and Pereira, F., 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* (2003), Association for Computational Linguistics, 134-141.

[10] Sparck Jones, K., Walker, S., and Robertson, S.E., 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1&2. *Information Processing & Management 36*, 6, 779-840.