UT-FX at NTCIR-10 MedNLP: Incorporating Medical Knowledge to Enhance Medical Information Extraction

Yasuhide Miura
Fuji Xerox Co., Ltd., Japan
yasuhide.miura@fujixerox.co.jp

Tomoko Ohkuma
Fuji Xerox Co., Ltd., Japan
Ohkuma.Tomoko@fujixerox.co.jp

Hiroshi Masuichi
Fuji Xerox Co., Ltd., Japan
hiroshi.masuichi@fujixerox.co.jp

Emiko Yamada Shinohara
The University of Tokyo Hospital, Japan
emiko-tky@umin.net

Eiji Aramaki
The University of Tokyo, Japan
JST PRESTO, Japan
eiji.aramaki@gmail.com

Kazuhiko Ohe
The University of Tokyo Hospital, Japan
The University of Tokyo, Japan
kohe@hcc.h.u-tokyo.ac.jp

ABSTRACT

The UT-FX team participated in the de-identification subtask and the complaint and diagnosis subtask of the NTCIR-10 MedNLP pilot task. This report describes our approach to solving the two subtasks.

Team Name

UT-FX

Subtasks

De-identification task
Complaint and diagnosis task

Keywords

natural language processing, named entity recognition, machine learning, medical informatics

1. INTRODUCTION

The UT-FX team participated in two subtasks, the de-identification (DEID) task and the complaint and diagnosis (CD) task, of the NTCIR-10 "Medical Natural Language Processing (MedNLP)" pilot task. DEID task is a task to identify personal information in medical reports. CD task is a task to identify medical information in medical reports. Both tasks target same Japanese medical reports that are provided by the task organizers. The detail of the two tasks can be found in the overview paper of the MedNLP task[7].

The remainder of this paper is organized as follows. Section 2 explains our approaches to the two tasks and describes the detail of our system. Section 3 describes experiments we performed to evaluate the proposed system. Section 4 describes the application of our system to the test data. Section 5 concludes the paper.

2. SYSTEM DESCRIPTION

We prepared a modular system for the two tasks. Figure 1 portraits the architecture of our system. We used the same system for both tasks with different system compositions.

2.1 Basic Architecture

2.1.1 Named Entity Recognition

The basic architecture of our system is a machine learning based named entity recognizer. We assumed the personal information in DEID task and the complaint and diagnosis information in CD task as named entities. Named entity recognition (NER) is well studied in the field of natural language processing. NER can be interpreted as a sequential labeling problem. We utilized linear-chain conditional random fields (CRF)[6], one of widely used methods to handle the problem, with character-level node. Character-level processing is chosen since Japanese text is unsegmented text and a character-level NER is known to achieve the state-of-the-art accuracy[3].

NER is known as a knowledge-intensive task and the use of external knowledge often boost the performance of it. Kazama and Torisawa[5] exploited Wikipedia category labels to enhance NER. Various knowledge resources (e.g. dictionary, terminology, ontology) are available in medical fields. We decided to exploit three medical knowledge resources, MedDRA/J[1], MEDIS Byomei Master[2], and MEDIS Shojo Shoken Master ⟨Shintai Shoken Hen⟩[3], to enhance our system. Additionally to these knowledge resources, we also introduced named entities that are defined on a different corpus. We trained the BASELINE composition of our system (detail will be described in Section 2.3) on the updated version of the discharge summary corpus (DS Corpus) mentioned in Aramaki et al.[2]. DS Corpus includes date/time, symptom, and disease named entities which we regarded as external knowledge in the MedNLP task.

Table 1 lists all features that are used in our system. For all features, sliding window features illustrated in figure 2 are considered. All features except for "NE-C" derive information from character, morpheme, or external knowledge. Therefore several preprocesses are done prior to the feature extraction. A morphological analysis and assignments of the resulting morphemes to character nodes are done to extract "M-*" features. A BIO-style match of the three knowledge resource similar to Kazama and Torisawa[5] is applied to ex-

[1]http://www.pmrj.jp/jmo/php/indexe.php
[2]http://www2.medis.or.jp/stdcd/byomei/index.html (In Japanese)
[3]http://www2.medis.or.jp/master/syoken/ (In Japanese)

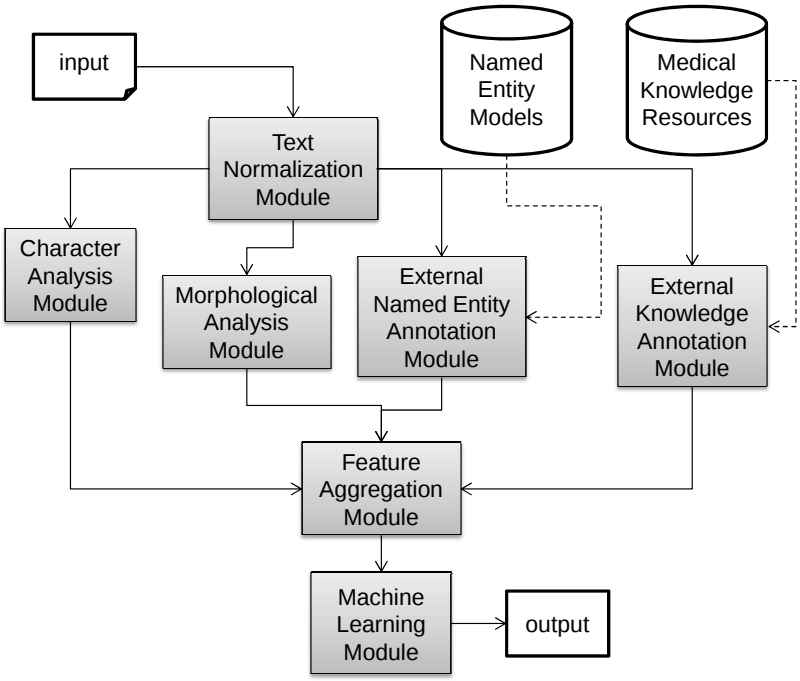

Figure 1: The architecture of UT-FX system.

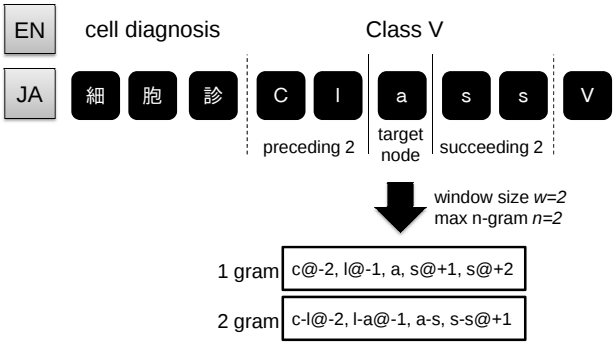

Figure 2: An example of sliding window features of "C-SURF" with window size $w = 2$ and max n-gram $n = 2$. A number following "@" represents the position from the target node.

tract "MEDDRA", "MEDIS-BM", and "MEDIS-SSM" features. The DS Corpus named entities are recognized and the BIO-style match of them are performed to extract "NE-DT" and "NE-SD" features. "NE-C" feature is the only exception, that uses information of c tags which is an objective of CD task. This special feature is used in a modality detection process which will be described in Section 2.1.2.

2.1.2 Modality Detection

Modalities of complaint and diagnosis information (c tags) are annotated in the MedNLP sample data. Three modalities, respectively "negation", "suspicion", and "family" are considered in the annotations. Negation and suspicion are well known phenomena that are widely studied in the natural language processing community. We took a simple approach to detect all three modalities at once. We treated c tags as nested named entities of c tag with no modality

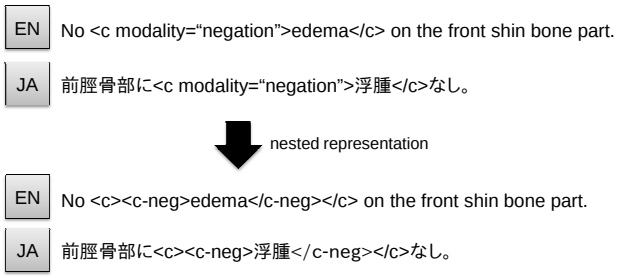

Figure 3: An example of a nested name entity representation.

and c tag with modality. Figure 3 shows an example of a nested representation. Once modalities are converted to nested representations, a 2-stage NER approach is taken to recognize them. The approach is similar to the outside-in layering technique taken by Alex et al.[1]. In the first stage, outside c tags are recognized. In the second stage, inside modality tags (⟨c-neg⟩ in the figure) are recognized using the information of outside c tags (⟨c⟩ in the figure). The c tags information are encoded as "NE-C" feature in similar manner to the other named entity features.

2.2 Modules

This section briefly describes our system in an implementation perspective. All modules in the system are implemented in Java to realize high portability.

Text Normalization Module

Three simple text normalization processes are applied to an input text as a first step. Firstly, a Unicode normalization in form NFKC[4] is applied. Secondly, all upper case characters

[4]http://unicode.org/reports/tr15/

Table 1: The list of features used in our system. "POS" in the table denotes "part-of-speech".

Feature	Brief Description
C-SURF	The surface form of a character.
C-TYPE	The type of a character. The Unicode block[i] is used for the type category.
M-SURF	The surface form of a morpheme.
M-BASE	The base form of a morpheme.
M-POS1	The POS layer 1 of a morpheme.
M-POS2	The POS layer 2 of a morpheme.
M-POS3	The POS layer 3 of a morpheme.
M-CJ-FORM	The conjugation form of a morpheme.
M-CJ-TYPE	The conjugation type of a morpheme.
MEDDRA	The matching with MedDRA/J entries.
MEDIS-BM	The matching with MEDIS Byomei Master entries.
MEDIS-SSM	The matching with MEDIS Shojo Shoken Master ⟨Shintai Shoken Hen⟩ entries.
NE-DT	The matching with recognized DS Corpus date/time named entities.
NE-SD	The matching with recognized DS Corpus symptom and DS Corpus disease named entities.
NE-C	The matching with c tags of the MedNLP task.

[i] http://www.unicode.org/charts/

are converted to lower case ones based on the definition of Unicode Standard version 4.0. Thirdly, all half-width characters are converted to full-width characters using ICU[5].

Character Analysis Module

Unicode blocks that the characters of a text belong to are extracted as character types.

Morphological Analysis Module

A morphological analysis is applied to a text using Kuromoji[6] with mode set to "Search". Assignments of resulting morphemes to corresponding characters are also done in this module.

External Knowledge Annotation Module

The entries in the three medical knowledge resources (MedDRA/J, MEDIS Byomei Master, and MEDIS Shojo Shoken Master ⟨Shintai Shoken Hen⟩) are matched to a text. For each resource, assignments of BIO-style tags (e.g. "B-MEDIS-BM", "I-MEDIS-BM") to each character are also done in this module.

External Named Entity Annotation Module

DS Corpus trained named entity recognizers and c tag named entity recognizer are applied to a text. For each named entity recognizer, assignments of BIO-style tags to each character are also done in this module.

Feature Aggregation Module

Features are aggregated based on a system composition setting and are encoded to the input format of the machine learning module. Sliding window features are set here with parameters window size w and max gram size n. A simple frequency based feature filtering is also available to ignore sparse features with frequency threshold t.

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

Table 2: The list of system compositions.

Composition	Features
BASELINE	{C-SURF, C-TYPE, M-SURF, M-BASE, M-POS1, M-POS2, M-POS3, M-CJ-FORM, M-CJ-TYPE}
DATETIME	BASELINE ∪ {NE-DT}
SYMPDIS	BASELINE ∪ {NE-SD}
MEDDIC	BASELINE ∪ {MEDDRA, MEDIS-BM, MEDIS-SSM}
FULL	BASELINE ∪ {NE-SD, MEDDRA, MEDIS-BM, MEDIS-SSM}

Table 3: The 5-fold cross validation results of DEID task.

Composition	Tag	*Precision*	*Recall*	F_1*Score*
BASELINE	a	86.67%	69.64%	77.23
	h	98.51%	88.00%	92.96
	t	90.42%	85.07%	87.66
DATETIME	a	79.55%	62.50%	70.00
	h	98.53%	89.33%	93.71
	t	91.52%	85.07%	88.18

Machine Learning Module

CRF is applied to aggregated features. For the implementation of CRF, MALLET[7] is used with default parameters.

2.3 System Compositions

We prepared five system compositions for the MedNLP task. Table 2 lists all compositions and their feature sets. BASELINE is a composition that only uses the features based on character and morpheme. DATETIME and SYMPDIS add one named entity feature to BASELINE. MEDDIC adds the external medical knowledge resource features to BASELINE. FULL is the union of SYMPDIS and MEDDIC.

3. EXPERIMENTS

We performed several experiments to evaluate our system.

3.1 De-identification Task

We considered only three tags, a (age), h (hospital), and t (time), which appeared more than 50 times in the sample data for the experiment of DEID task.

A performance of our system was evaluated using all 2,244 sentences of the sample data. 5-fold cross validations were ran on two system compositions: BASELINE and DATETIME. The parameters of the feature aggregation module were set to $w = 2$, $n = 2$, and $t = 2$. Table 3 shows the micro average 5-fold cross validation values of the two compositions. A statistical significance of the two compositions were tested by approximate randomization test (a testing method same as MUC-4[4]) with iteration number set to 10,000. For precision, recall, and F_1score, the two compositions were not statistically significant with $p \leq 0.05$.

[7]http://mallet.cs.umass.edu/

Table 4: The 5-fold cross validation results of CD task. The underlined values represent statistically significant improvements.

Composition	*Precision*	*Recall*	F_1*Score*
BASELINE	87.87%	81.43%	84.53
SYMPDIS	87.46%	84.18%	85.79
MEDDIC	88.57%	83.45%	85.94
FULL	88.39%	84.76%	86.54

3.2 Complaint and Diagnosis Task

3.2.1 Named Entity Recognition

A performance of our system was evaluated on *c* tag NER using all 2,244 sentences of the sample data. 5-fold cross validations were ran on four system compositions: BASELINE, SYMPDIS, MEDDIC, and FULL. Same values as the DEID experiments were used for the parameters of the feature aggregation module. Table 4 shows the micro average 5-fold cross validation values of the four compositions. Statistical significances between four compositions were tested by three pairs: SYMPDIS–BASELINE, MEDDIC–BASELINE, and FULL–MEDDIC. Statistically significant improvement with $p \leq 0.05$ were achieved in, the recall and the F_1 score of SYMPDIS, the precision, the recall, and the F_1 score of MEDDIC, and the recall of FULL.

3.2.2 Modality Detection

A performance of our system was evaluated on *c* tag modality detection using all 2,244 sentences of the sample data. In the second stage of our modality detection method, the information of *c* tags are necessary. We skipped the first stage and utilized the gold *c* tags in the second stage for this experiment. The gold tags are used to ignore the effect of the NER performance of the first stage. Four compositions same as the NER experiments were evaluated (Table 5). The statistical significances of SYMPDIS, MEDDIC, and FULL were tested against BASELINE in this experiment. Statistical significances were not observed in any precision, recall, or F_1 score with $p \leq 0.05$.

4. TEST RUNS

4.1 De-identification Task

We submitted BASELINE and DATETIME compositions to DEID task. BASELINE and DATETIME correspond to "A1" and "A2" of the overview paper[7] respectively. In overall F_1 score metric, BASELINE scored better than DATETIME and ranked sixth out of the fifteen submitted systems with the score of 87.98.

4.2 Complaint and Diagnosis Task

We submitted BASELINE, MEDDIC, and FULL compositions to CD task. BASELINE, MEDDIC, and FULL correspond to "A3", "A1", and "A2" of the overview paper[7] respectively. In overall F_1 score metric of the 2-way result, MEDDIC scored better than BASELINE and FULL and ranked first out of the twenty-two submitted systems with the score of 85.93.

The best composition MEDDIC in the test run differed from the best composition FULL of the cross validation result. However, the result is not so surprising since the F_1 score difference of FULL to MEDDIC was not statistically significant as described in Section 3.2.1.

Table 5: The 5-fold cross validation results of CD task modality detection with the gold *c* tags.

Composition	Modality	*Precision*	*Recall*	F_1*Score*
BASELINE	none	87.46%	94.98%	91.06
	negation	87.50%	76.39%	81.57
	suspicion	61.11%	30.56%	40.74
	family	78.57%	34.38%	47.83
SYMPDIS	none	87.54%	94.67%	90.97
	negation	87.13%	76.59%	81.52
	suspicion	57.50%	31.94%	41.07
	family	78.57%	34.38%	47.83
MEDDIC	none	87.99%	94.22%	91.00
	negation	84.13%	76.79%	80.29
	suspicion	69.70%	31.94%	43.81
	family	76.47%	40.62%	53.06
FULL	none	87.81%	93.76%	90.69
	negation	84.65%	76.59%	80.42
	suspicion	55.81%	33.33%	41.74
	family	75.00%	37.50%	50.00

5. CONCLUSIONS

We presented a system that utilizes external medical knowledge into a state-of-the-art named entity recognizer. In CD task, the introduction of external medical knowledge improved *c* tags recognition by about 2.03 in overall F_1 score of the test run 2-way result and contributed to realize the first ranking system. The result suggests the promising future of a natural language processing in medical fields, where numerous knowledge resources are available.

6. REFERENCES

[1] B. Alex, B. Haddow, and C. Grover. Recognising nested named entities in biomedical text. In Proceedings of the Workshop on BioNLP 2007, pages 65–72, 2007.

[2] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Mashuichi, and K. Ohe. TEXT2TABLE: Medical text summarization system based on named entity recognition and modality identification. In Proceedings of the BioNLP 2009 Workshop, pages 185–192, 2009.

[3] M. Asahara and Y. Matsumoto. Japanese named entity extraction with redundant morphological analysis. In Proceedings of HLT-NAACL 2003, pages 8–15, 2003.

[4] N. Chinchor. The statistical significance of the MUC-4 results. In Proceedings of MUC-4, pages 30–50, 1992.

[5] J. Kazama and K. Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. In Proceedings of EMNLP-CoNLL 2007, pages 698–707, 2007.

[6] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001, pages 282–289, 2001.

[7] M. Morita, Y. Kano, T. Ohkuma, M. Miyabe, and E. Aramaki. Overview of the NTCIR-10 MedNLP task. In Proceedings of NTCIR-10, 2013. To appear.