Incorporating Knowledge Resources 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, Japan
emiko-tky@umin.net

Eiji Aramaki
Kyoto University, 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

This paper describes a method to extract medical information from texts. The method targets to extract complaints and diagnoses from electronic health record texts. Complaints and diagnoses are fundamental information and can be used for more complex medical tasks. The method utilizes several medical knowledge resources to enhance the performance of extraction. With an evaluation using NTCIR-10 MedNLP data, our method marked 86.53 in F_1 score with a cross validation. The score is comparable to top scoring teams in NTCIR-10 MedNLP task. The approach taken to incorporate knowledge resources has a high generality. It is not restricted to the resources presented in this paper and can be applied to various other resources.

1 Introduction

Spread of electronic health record (EHR) brought a large amount of unstructured medical data that can be processed electronically. The data include valuable information about patients health. An automatic extraction of medical information from them is beneficial since manual analyses of them by medical experts are often difficult because of their quantity. This paper describes a method that enables such automatic extraction.

Various kinds of information have been targeted for an extraction from EHR texts. NLP shared tasks of Informatics for Integrating Biology and the Bedside (i2b2)[1] designed challenges to extract kinds of medical information such as: smoking status, obesity, medication, medical problem, medical test, and treatment. Medical Records tracks of Text Retrieval Conference (TREC)[2] modeled an extraction of cohorts that are effective for medical researches. Natural Language Processing (MedNLP) task of NTCIR[3] aimed to extract patient complaints and diagnoses from EHR texts. The method described in this paper targets to extract complaints and diagnoses, the same kind of information that NTCIR-10 MedNLP intended. Complaints and diagnoses are fundamental information and can be useful for complex medical tasks. Example of such tasks are: an assignment of disease codes and a detection of adverse effects in medications. Figure 1 shows an example of a diagnosis description in an EHR text.

EN No <c modality="negation">edema</c> on the front shin bone part.

JA 前脛骨部に<c modality="negation">浮腫</c>なし。

Figure 1: An example of a diagnosis description in an EHR text of NTCIR-10 MedNLP data. In NTCIR-10 MedNLP, *c* tag is used to denote a complaint or a diagnosis.

The extraction of complaints and diagnoses is known to achieve a moderate performance (78.86 in F_1 score) by applying a simple conditional random field (CRF) based named entity recognition (NER) method (Imachi et al., 2013). Our method utilizes medical knowledge to a CRF based NER method to enhance an extraction performance. Our contribution in this paper is that we show a substantial increase in the extraction performance of complaints and diagnoses by incorporating several medical knowledge resources. The paper

[1] https://www.i2b2.org/

[2] http://trec.nist.gov/

[3] http://ntcir.nii.ac.jp/

International Joint Conference on Natural Language Processing Workshop on Natural Language Processing for Medical and Healthcare Fields, pages 1–6, Nagoya, Japan, 14-18 October 2013.

also discusses the detailed effects of the individual knowledge resources and the generality of the method.

The outline of this paper is as follows. Section 2 explains the detail of the method. Section 3 describes an experiment we performed for an evaluation. Section 4 notes the related works. Section 5 discusses the result of the experiment and the generality of the method. Section 6 concludes the paper.

2 Method

A method that we prepared for a medical information extraction is basically a machine learning based named entity recognizer. The method assumes that information to be extracted can be expressed as named entities. NER can be interpreted as a sequential labeling problem. We utilized linear-chain CRF (Lafferty et al., 2001), 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 (Asahara and Matsumoto, 2003).

NER is known as a knowledge-intensive task and the use of external knowledge often boosts the performance of it (Ratinov and Roth, 2009). Various knowledge resources (e.g. dictionary, terminology, ontology) are available in medical fields. We decided to exploit three publicly available medical terminologies, MedDRA/J[4] (Brown et al., 1999), MEDIS Byomei Master[5] (Medical Information System Development Center, 2012), and MEDIS Shojo Shoken Master ⟨Shintai Shoken Hen⟩[6].

Additionally to these terminologies, we also utilized information obtained from an external corpus in a medical domain. We introduced named entities that are defined on the updated version of the discharge summary corpus (DS Corpus) mentioned in Aramaki et al. (2009). DS corpus contains *symptom* named entities and *disease* named entities that are similar to complaints and diagnoses in NTCIR-10 MedNLP task. BASELINE composition of our method (detail will be described in Section 3.1) was trained on DS Corpus to realize a DS Corpus named entity recognizer.

[4]http://www.pmrj.jp/jmo/php/indexe.php

[5]http://www2.medis.or.jp/stdcd/byomei/index.html (In Japanese)

[6]http://www2.medis.or.jp/master/syoken/ (In Japanese)

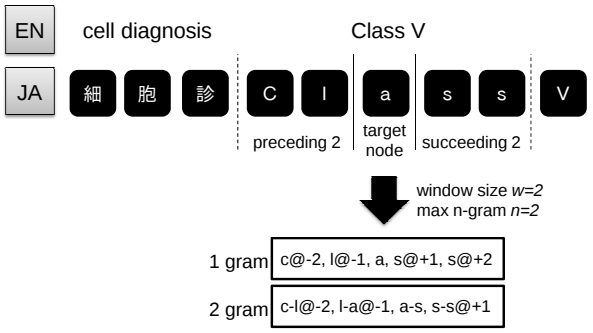

Figure 2: An example of sliding window features of "C-SURF" with window size $w = 2$ and max n-gram $n = 2$. A number following "@" represents the position from the target node.

Table 1 lists all features that are used in our method. For all features, sliding window features illustrated in figure 2 are considered. All features 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 terminologies similar to Kazama and Torisawa (2007) is applied to extract "K-MEDDRA", "K-MEDIS-BM", and "K-MEDIS-SSM" features. The DS Corpus named entities are recognized and the BIO-style matches of them are performed to extract "K-NE-SD" feature.

2.1 Implementation

This section briefly describes the method in an implementation perspective. Figure 3 portraits the architecture of the method.

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[7] is applied. Secondly, all upper case characters 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[8].

Character Analysis Module

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

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

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

Feature	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 part-of-speech layer 1 of a morpheme.
M-POS2	The part-of-speech layer 2 of a morpheme.
M-POS3	The part-of-speech layer 3 of a morpheme.
M-CJ-FORM	The conjugation form of a morpheme.
M-CJ-TYPE	The conjugation type of a morpheme.
K-MEDDRA	The BIO-style matching result of a character with MedDRA/J entries.
K-MEDIS-BM	The BIO-style matching result of a character with MEDIS Byomei Master entries.
K-MEDIS-SSM	The BIO-style matching result of a character with MEDIS Shojo Shoken Master ⟨Shintai Shoken Hen⟩ entries.
K-NE-SD	The BIO-style matching result of a character with recognized DS Corpus symptom named entities and DS Corpus disease named entities.

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

Table 1: The list of features used in the method.

Morphological Analysis Module

A morphological analysis is applied to a text using Kuromoji[9] with mode set to "Search". Assignments of resulting morphemes to corresponding characters are also done in this module.

External Named Entity Annotation Module

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

External Terminology Annotation Module

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

Feature Aggregation Module

Features are aggregated based on a feature composition and are encoded to the input format of the machine learning module. Sliding window features are set here with the parameters of 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.

Machine Learning Module

CRF is applied to aggregated features. For the implementation of CRF, MALLET[10] is used with default parameters.

3 Experiment

3.1 Feature Compositions

We prepared six feature compositions of the method. Table 2 lists all compositions and their feature sets. BASELINE is a composition that we prepared as a baseline of the method. It only consists of the features based on character and morpheme. DSNE adds the named entity feature to BASELINE. MEDDRA and MEDIS add one terminology feature to BASELINE. MEDDIC adds all terminology features to BASELINE. FULL uses all defined features.

3.2 Evaluation Data

A performance of our method was evaluated on the training portion of NTCIR-10 MedNLP data. The data consist of 2,244 sentences with 1,922 complaint or diagnosis (*c* tag) annotations. Modality information included in some of *c* tags is not considered in this experiment. The detail of the data can be found in the overview paper of NTCIR-10 MedNLP (Morita et al., 2013).

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

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

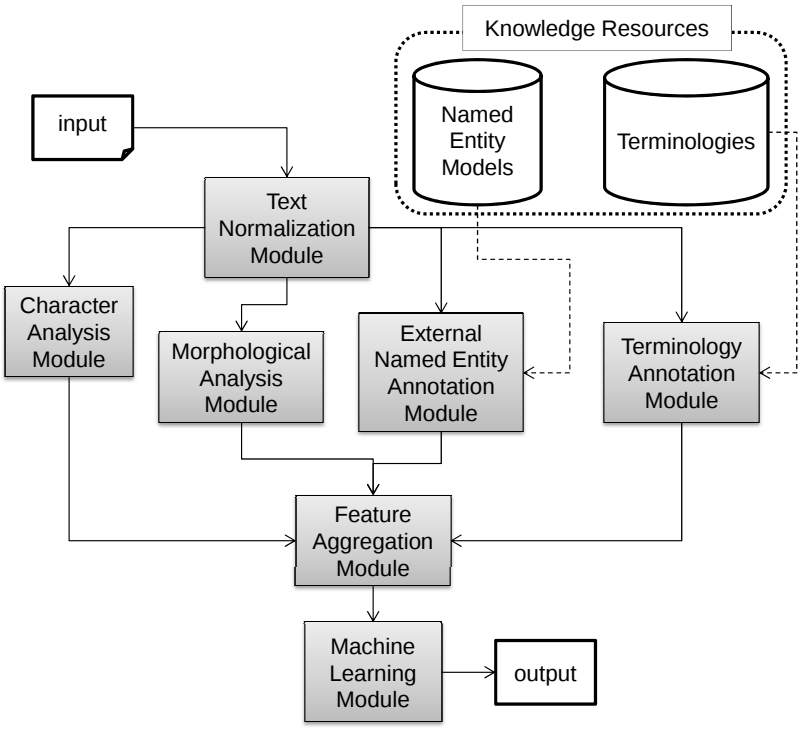

Figure 3: The architecture of the method.

Composition	Feature Sets
BASELINE	{C-SURF, C-TYPE, M-SURF, M-BASE, M-POS1, M-POS2, M-POS3, M-CJ-FORM, M-CJ-TYPE}
DSNE	BASELINE ∪ {K-NE-SD}
MEDDRA	BASELINE ∪ {K-MEDDRA}
MEDIS	BASELINE ∪ {K-MEDIS-BM, K-MEDIS-SSM}
MEDDIC	MEDDRA ∪ MEDIS
FULL	DSNE ∪ MEDDIC

Table 2: The list of feature compositions.

Composition	Precision	Recall	F_1 score
BASELINE	87.87%	81.43%	84.53
DSNE	87.46%	84.18%	85.79
MEDDRA	88.88%	82.78%	85.72
MEDIS	89.40%	82.52%	85.82
MEDDIC	88.57%	83.45%	85.94
FULL	88.39%	84.76%	86.53

Table 3: The 5-fold cross validation results of the method. The underlined values represent statistically significant improvements.

3.3 Extraction Performance

We measured precisions, recalls, and F_1 scores of *c* tag extraction as extraction performances. 5-fold cross validations were ran on six system compositions: BASELINE, DSNE, MEDDRA, MEDIS, MEDDIC, and FULL. 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 six compositions.

A statistical significance of two compositions was tested by a randomization test described in Noreen (1989) with iteration number set to 10,000. Statistical significances between six compositions were tested by five pairs: DSNE–BASELINE, MEDDRA–BASELINE, MEDIS–BASELINE, MEDDIC–BASELINE, and FULL–MEDDIC. Statistically significant improvements with $p \leq 0.05$ were achieved in, the recall and the F_1 score of DSNE, the precision, the recall, and the F_1 score of MEDDRA, the precision and the F_1 score of MEDIS, the precision, the recall, and the F_1 score of MEDDIC, and the recall of FULL.

4 Related Works

NER is well studied in the field of natural language processing. A number of design issues in NER are discussed in Ratinov and Roth (2009). This section explains NER methods that have close relationship with our method.

A character-level processing of NER is investigated in some literatures. Asahara and Matsumoto (2003) showed that a state-of-the-art Japanese

Terminology	# of Terms
MedDRA/J	922
MEDIS BM & SSM	1,041
MedDRA/J ∩ MEDIS BM & SSM	421

Table 4: The number of terms that are present in NTCIR-10 MedNLP data for each terminology. MEDIS BM & SSM is the union of the two MEDIS terminologies that we used.

NER can be realized with character level processing. Klein et al. (2003) demonstrated the effectiveness of using character substrings in an English NER.

The effectiveness of using dictionaries or gazetteers is shown in previous works. Florian et al. (2003) used location, person, and organization gazetteers in their NER framework and reported an error reduction in an extraction performance. Cohen and Sarawagi (2004) exploited a state, a city, a person, and a company dictionaries to improve NER. Jonnalagadda et al. (2013) used various medical resources in their NER system and showed an increase in an extraction performance of medical concepts. Automatic constructions of a dictionary/gazetteer are also examined. Kazama and Torisawa (2007) and Toral and Muñoz (2006) exploited Wikipedia to construct a dictionary/gazetteer that is useful for NER.

5 Discussion

5.1 Effects of Knowledge Resources

The use of terminology resulted to high precision recognizers. The best result in precision of 89.40% was obatained by only using the MEDIS terminologies, but its recall was the only one that did not show the statistically significant improvement against the baseline. The use of MedDRA terminology was similar to the use of MEDIS terminologies with a slightly higher recall and a slightly lower precision. Regardless of this similarity, terms that are present in NTCIR-10 MedNLP data are somewhat different between the two kinds of terminologies (Table 4). The percentages of terms that are not unique are about 40.4% and 45.7% for MedDRA/J and MEDIS BM & SSM respectively. We assume that even though the two kinds of terminologies are rather different in term presence, both kinds included similar information that is essential for NER.

The introduction of the external named entities (DSNE) resulted to a different result in certain extent compared to the terminology utilizations. The recall marked the second highest score of 84.18% but the precision was lower, although not statistically significant, than the baseline. We assume that symptom named entities and disease name entities in DS Corpus can be a clue to recognize complaints and diagnoses (high recall) but differences between them degraded the certainty of recognition (low precision).

5.2 Generality of Knowledge Resource Incorporation

The approach we took for the incorporation of terminology has a high generality. The approach requires only entries of a terminology. More rich contents like glosses or synonyms are not required. This characteristic makes the incorporation applicable to almost any kind of terminology.

The technique to introduce external named entities also has a high generality. The technique encodes named entity results as binary features for each entity type. This encoding can be done to almost any type of named entity. However, as mentioned in Section 5.1, the introduction of external named entity showed the defect in precision. This may be undesirable in some practical uses.

6 Conclusion

We presented a method that utilizes external medical knowledge into a state-of-the-art named entity recognizer. An evaluation using NTCIR-10 MedNLP data showed that the introduction of the medical knowledge resources improves the complaints and diagnoses extraction performance by about 2.00 in F_1 score. The best F_1 score 86.53 obtained in our method is comparable to top scoring results in Complaint and Diagnosis subtask of NTCIR-10 MedNLP.

The presented knowledge resource incorporation method has high generality, and its application is not restricted to the resources described in this paper. For example, a drug terminology can be incorporated to a medication extraction. This high generality suggests the promising future of a natural language processing in medical fields, where numerous knowledge resources are available.

References

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. TEXT2TABLE: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192.

Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of HLT-NAACL 2003*, pages 8–15.

Elliot G. Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug Safety*, 20(2):109–117.

William W. Cohen and Sunita Sarawagi. 2004. Exploiting dictionaries in named entity extraction: Combining semi-Markov extraction processes and data integration methods. In *Proceedings of KDD 2004*, pages 89–98.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of CoNLL 2003*, pages 168–171.

Hiroto Imachi, Mizuki Morita, and Eiji Aramaki. 2013. NTCIR-10 MedNLP task baseline system. In *Proceedings of NTCIR-10*, pages 710–712.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, Hongfang Liu, and Graciela Gonzalez. 2013. Evaluating the use of empirically constructed lexical resources for named entity recognition. In *Proceedings of CSCT 2013*, pages 23–33.

Junichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of EMNLP-CoNLL 2007*, pages 698–707.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of CoNLL-2003*, pages 180–183.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*, pages 282–289.

Medical Information System Development Center, editor. 2012. *Hyojun Byomei Handobukku 2012 [Standard Disease Name Handbook 2012] (In Japanese)*. Shakai Hoken Kenkyujo, Inc.

Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. 2013. Overview of the NTCIR-10 MedNLP task. In *Proceedings of NTCIR-10*, pages 696–701.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. John Wiely and Sons, Inc.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL-2009*, pages 147–155.

Antonio Toral and Rafael Muñoz. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, pages 56–61.