NTCIR-10 MedNLP Task Baseline System

Open source system at MedNLP

Hiroto Imachi
The University of Tokyo
Tokyo, Japan
hiroto_imachi@mist.i.u-tokyo.ac.jp

Mizuki Morita
The University of Tokyo
Tokyo, Japan
morita.mizuki@gmail.com

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

ABSTRACT

Natural language processing (NLP) technology that handles clinical, medical and health records has been drawn much attention, because such kinds of records potentially could be rich clinical resources. This paper describes an NLP system that extracts two kinds of information from clinical documents in Japanese, which was developed as a baseline system in the NTCIR-10 MedNLP Pilot Task. Since our system consists of only open source tools and resources, it can be freely used by anyone. The experimental results showed reasonably good performances in both of two subtasks in the NTCIR-10 MedNLP Pilot Task; (1) de-identification task (precision: 86.10%, recall:74.54%, F-measure 79.9%) and (2) complaint and diagnosis task (precision: 87.37%, recall: 71.86%, F-measure 78.86%). These results have demonstrated the basic feasibility of our simple system.

Keywords

natural language processing (NLP), medical informatics, freeware, open source, named entity recognition (NER), machine learning, conditional random field (CRF)

Team Name

cks01

Subtask

(1) De-identification task

(2) Complaint and diagnosis task

1. INTRODUCTION

Developing an automatic system for analyzing clinical records is becoming more and more important, because the amount of electronic medical records has been increasing in recent years. However, current methods for analyzing them are far from well established.

This paper describes the system to process clinical records in Japanese. Our system focuses on two categories of information retrieval: (1) privacy information and (2) patient clinical information. Since these tasks are similar to the task of named entity recognition (NER), our system has employed existing NER methods and tools. Note that our system consists of only open source tools and resources, because it has aimed to be freely distributed.

The experimental results showed reasonably high performance in the both of de-identification task and complaint and diagnosis task as follows:

De-identification task: precision: 86.10%, recall:74.54%, F-measure 79.90%

Complaint and diagnosis task: precision 87.37%, recall: 71.86%, F-measure 78.86%

By comparing our results with those from other groups in the NTCIR-10 MedNLP Pilot Task, we would demonstrate that our baseline system without extra resources could achieve moderate results.

The remainder of this paper is organized as follows: Section 2 explains resources and tools of our system; Section 3 describes experiments we performed to evaluate the proposed system; Section 4 reported the results; Section 5 concludes the paper.

2. MATERIALS

2.1 Corpus

In the NTCIR-10 MedNLP Pilot Task, we have worked with a corpus consists of dummy clinical texts (medical history summaries) in Japanese written by physicians. The corpus is composed of 50 patient summaries. Each summary contains chief complaint, present illness, past history and so on.

The corpus is tagged with corresponding element names. They are categorized as patients' personal information or medical information. Personal information consist of age, sex, time, location and hospital, and they are tagged with <a>, <x>, <t>, <l> and <h>, respectively. Patient medical information means complainant and diagnosis, which are tagged with <c>. Some of them are marked with modalities (negation, suspicion and so on), which are of importance for clinical information [3].

An example of a tagged sentence in this format is shown in Table 1.

Table 1: Exapmle of Tagged Sentence

<t> 昨日から </t><c > 悪寒 </c>、および <c > 発熱 </c> の症状。 (symptom of chills and fever from yesterday)

The detailed information of each tag is shown in the overview paper of the NTCIR-10 MedNLP Pilot Task [9].

2.2 CRF++

Among various methods for term extraction, one of the popular approaches is conditional random field (CRF) based term identification. CRF is a discriminative graphical model that has been widely used in NLP starting from part-of-speech tagging [4], named entity recognition [8], parsing [12], information extraction [11] and many other tasks. Especially term identification is one of the most successful applications of CRF. Since our system has utilized an open source application CRF++[1], it could be freely distributed.

2.3 MeCab

To realize a CRF based term identification, features from a word, such as part-of-speech and pronunciation, are required. In order to obtain these features, we have employed MeCab[2], which is an open source application for Japanese morphological analysis system.

2.4 Dictionary Resources

We have used a dictionary of disease name in Japanese [6] that is published by Medical Information System Development Center (MEDIS-DC) in Japan. This resource is available on the web[3].

3. METHODS

3.1 Format Conversion (IOB2 Representation)

First of all, we converted a corpus in XML format into another format, 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 or B. In the case that a segment is just behind an start tag, its label is B-(element name of the tag). Other segments between corresponding start and end tag are labeled I-(element name of the tag). Segments outside tags are labeled O. The example sentence in the corpus shown in table 1 is converted as Table 2.

3.2 Conditional Random Fields (CRF) Training

We have trained CRF model with standard parameters and a standard template provided by the tool CRF++. The features for machine learning consist of basic features reported in previous Japanese medical NER studies [1, 2]. The example of features are presented in Table 3. Among the features

[1]http://crfpp.googlecode.com/
[2]http://mecab.googlecode.com
[3]http://www.medis.or.jp/

Table 2: IOB2 Format Example

token	IOB2 tag
昨日 (yesterday)	B-t
から	I-t
悪寒 (chills)	B-c
、	O
および (and)	O
発熱 (fever)	B-c
の	O
症状 (symptom)	O
。	O

in Table3, only the “Dictionary feature” comes from the disease name dictionary mentioned above. The others including “POS (part-of-speech)” and “Pronunciation” come from the output of morphological analysis tool MeCab.

Table 3: Features

Features	example
Token	*insulin*
Last character	*n*
POS (part-of-speech)	*noun*
Pronunciation	*i-n-syu-ri-n*
Character type	*alphabet*
Dictionary feature	*Out-of-Dictionary*

3.3 Test and Evaluation

We have run the CRF for the *test set*. Evaluation criteria followed those of CoNLL 2000 Shared Task [13] (precision, recall, F-measure and accuracy).

4. RESULTS

4.1 De-identification Task

The results are shown in Table 4. The row of <l> is empty because <l> tag appears only once in the training data (the *sample set*), estimation of <l> tag was completely failed. Another tag <l> is in the similar situation. These scarce data cannot be identified correctly without external knowledge.

High accuracy of <h> tag came from frequent fixed phrases “当院 (this hospital)” and “近院 (nearest hospital)”.

Words with <a> and <t> tags have more various representations than those with <h> tag but frequently appearing expressions such as “歳 (year old)” and “まで (until)” make these taggings easier. Failure cases in these tags occurred at, for example, “１日２回 (twice a day)”. This segment does not represent date time, but the word “日 (day)” led our system into error. This type of failure might not be solved merely by enlarging training data set.

4.2 Complaint and Diagnosis Task

Accuracy for <c> tag was relatively low due to large proportion of unknown words in the test case. Utilizing external knowledge, such as International Classification of Primary Care (ICPC) and International Statistical Classification of Diseases and Related Health Problems (ICD), might improve the problem partially.

Performances in modality detection are even lower than those in simple term detection. Although various methods are so far proposed [5, 7, 10, 14], it is still a difficult task for the current NLP techniques.

Table 4: Evaluation Result

tag	Precision	Recall	$F_{\beta=1}$
<a>: age	88.89	75.00	81.36
<t>: time	83.08	76.60	79.70
<h>: hospital	96.15	65.79	78.12
<l>: location	0.00	0.00	0.00
<x>: sex	100.00	50.00	66.67
<c>	87.37	71.86	78.86
<c> positive	62.58	62.08	62.33
<c> family	100	59.09	74.29
<c>negation	78.47	66.40	71.93
<c> suspicion	60.00	20.00	30.00

5. CONCLUSION

We developed a baseline system for the first shared task on information retrieval from clinical documents in Japanese. Although our system is fairly simple, requiring the open software and free resources, the experimental results showed moderate performances. The error analysis and some ideas for improvement were also shown. We hope that many people will use our system.

6. ACKNOWLEDGMENTS

Part of this research has been supported by the JST PRESTO program and a grant-in-aid for scientific research (A) of Japan Society for the Promotion of Science.

7. REFERENCES

[1] E. Aramaki, T. Imai, K. Miyo, and K. Ohe. Automatic deidentification by using sentence features and label consistency, 2006.

[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 Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2003) Workshop on BioNLP*, pages 185–192, 2009.

[3] W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 5:301–10, 2001.

[4] R. Dearden. Structured prioritised sweeping. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 82–89, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[5] P. L. Elkin, S. H. Brown, B. A. Bauer, C. S. Husser, W. Carruth, L. R, Bergstrom, and D. L. Wahner-Roedler. *A controlled trial of automated classification of negation from clinical notes.* BMC Medical Informatics and Decision Making 5:13, 2005.

[6] K. Hatano and K. Ohe. Information retrieval system for japanese standard disease-code master using xml web service. In *American Medical Informatics Association (AMIA) Symposium*, pages 597–602, 2003.

[7] Y. Huang and H. J. Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304–311, 2007.

[8] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[9] M. Morita, Y. Kano, T. Ohkuma, M. Miyabe, and E. Aramaki. Overview of the ntcir-10 mednlp task. In *Proceedings of the NTCIR 10*, 2013.

[10] P. G. Mutalik, A. Deshpande, and P. M. Nadkarni. Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the umls. *Journal of the American Medical Informatics Association*, 8(6):598–609, 2001.

[11] D. Pinto, A. McCallum, X. Lee, and W. Croft. Table extraction using conditional random fields. In *Proceedings of the 26th ACM SIGIR*, pages 235–242, 2003.

[12] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2003) Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 213–220, 2003.

[13] E. F. Tjong Kim Sang and S. Buchholz. Introduction to the conll-2000 shared task: Chunking. In C. Cardie, W. Daelemans, C. Nedellec, and E. Tjong Kim Sang, editors, *Proceedings of CoNLL-2000 and LLL-2000*, pages 127–132. Lisbon, Portugal, 2000.

[14] V. Vincze, G. Szarvas, R. Farkas, G. Mora, and J. Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(11), 2008.