ABSTRACT

For MedNLP complaint and diagnosis subtask we tried a simple dictionary-matching method using MeCab, and achieved 61.10% F-score in official evaluation. Our method can evaluate the coverage of terminology data in a simple and inexpensive way.

Team Name
NIPH

Subtasks
Complaint and diagnosis (Japanese)

Keywords
1. INTRODUCTION

Recently, the volume of medical records written in electronic format is increasing. As a result, information processing technique in medical fields is in demand. Medical records contain information in natural language such as description of the medical status of patients, which needs to be processed with NLP techniques. However, few NLP resources such as ontology and electronic dictionary are publicly available in the Japanese medical community, in contrast to bio- or genome-science community. Computerized medical records systems have such resources built-in, but they differ from system to system and not for use outside the system.

In such circumstances some doctors develop their own terminology lists and dictionaries privately, which are customized to their own purposes, and as such, they may have a limited scope of coverage. Our attempt here is to use one of such resources against the NTCIR-MedNLP task [2] data to evaluate the coverage of such customized data.

We take one of such dictionaries, developed by Dr. Keijiro Torigoe and used for developing a diagnosis support system National Institute of Public Health [3]. The dictionaries, called Disease Master Data and Symptom Master Data, consist of records of about 1900 disease names and 800 symptom names respectively, with English translations and short descriptions[1]. We try to find complaint information by simply matching the text with these Master data, with the method described in the following section.

2. METHODS

The Japanese names of diseases and symptoms are extracted from the Master Data. In addition to these, we added the names of diseases, syndromes, symptoms, deficiencies, and side effects extracted from a laboratory test manual [4] to increase the volume of candidate names. We used Japanese morphological analyzer MeCab [1] as a dictionary matcher. That is, we built a user dictionary from these data and analyzed the text with MeCab with the user dictionary. The approach taken here is very simple so that everyone can employ with minimal training. We hope that, if this simple approach is successful, non-experts in NLP can test their own resources casually and in the end help enrich public NLP resources.

2.1 Data

The dictionary that we built for MeCab is based on the following three datasets. The numbers of dictionary entries taken from each dataset are shown in parentheses.

- Disease Master Data: List of names of diseases and syndromes, compiled by Dr. Keijiro Torigoe (1912).
- Symptom Master Data: List of names of symptoms and findings, compiled by Dr. Keijiro Torigoe (861).
- Lab Test Data: Names of diseases, syndromes, symptoms, deficiencies, and side effects extracted from laboratory test manual (2492).

XML character entity references in these master data, representing Greek characters, are mapped to corresponding characters in JIS Zenkaku. Alphanumeric characters in the records are also mapped to corresponding characters in Zenkaku. With these master datasets combined as a user dictionary, MeCab ver 0.995 with IPA dictionary was used for matching. All the extract names were treated as common noun, with smaller costs than those in the system dictionary, to give higher preferences. All the remaining fields for the grammatical attributes were set to the default value for common nouns. The costs for those names from the Disease Master Data and the Symptom Master Data were set to 100, and the costs for names from the Lab Test Data were set to 200. These are significantly smaller than most of words in the system (IPA) dictionary, which are assigned the value in the range of 4000–6000.

To signify that the words are from the user dictionary, the "reading" field and the "pronunciation" field of the MeCab record are utilized. For each entry, the reading field was set to the name of the dataset, and the pronunciation field

[1] The data are available from I’ROM Holdings Co. Ltd. (http://www.irom-hd.co.jp/) under BSD license.
Table 1: Result of the Analysis on the Sample Text: *Complete match* denotes the <<e>> elements that match an entry or a series of entries in the user dictionary; *Partial match* denotes the elements that include an entry as its part; *No match* denotes other elements.

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete match</td>
<td>697</td>
<td>36.3</td>
</tr>
<tr>
<td>Partial match</td>
<td>352</td>
<td>18.3</td>
</tr>
<tr>
<td>No Match</td>
<td>873</td>
<td>45.4</td>
</tr>
</tbody>
</table>

3. ADDITION OF THE ATTRIBUTE

A <<e>> element in the NTCIR MedNLP may have an attribute named *modality*. The value of the *modality* attribute may be *family* (indicating the symptom or disease was found in one of the patient’s family member, not in the patient him/herself), *negation* (indicating the symptom or disease was NOT found), or *suspicion* (indicating that the doctor is suspicious about the finding). A heuristic rule-based method was used for determining the value of *modality* attribute.

- **family**: If a word designating a family member such as father or grandfather was found in a sentence, all the elements following the word within a sentence are given *family* value to their *modality* attribute;
- **negation**: If a word designating negation or suspicion was found after the elements and within a window bounded by punctuation marks, symbols, particles, or verb *suru* the elements were given *negation* or *suspicion* to their *modality* attribute.

The rules were obtained from manual observation of the sample text and implemented in Perl as a postprocessor.

4. RESULTS

There were 1004 <<e>> elements found by our algorithm. Of those found, 127 were from Disease Master Data, 276 were from Symptom Master Data, 147 were from Lab Test Data, and 454 were from the sample text data. The formal results of the current approach were 58.67% in precision, 63.74% in recall, 61.10% in F-score, and 93.50% in accuracy.

5. CONCLUSIONS

We have tested the coverage of Master Data used in National Institute of Public Health against NTCIR MedNLP data, using Japanese morphological analyzer MeCab. We can conclude that our method can evaluate the coverage of terminology data in a simple and inexpensive way.

The coverage of the current Master Data was reasonable in terms of official score, but according to the results of the preliminary test, the coverage of the Master Data was indicate that ontological inference may be necessary to increase the coverage. On the other hand, the results may indicate that if a dictionary with ontological structure could be obtained, a simple dictionary matching would produce useful results.

6. REFERENCES