ABSTRACT
We have been developing an English-Japanese thesaurus of medical terms for 20 years. The thesaurus is compatible with MeSH (Medical Subject Headings) developed by National Library of Medicine, USA and contains approximately 30 thousand headings with 200 thousand synonyms (consisting of the names of anatomical concepts, biological organisms, chemical compounds, methods, disease and symptoms). In this study, we aimed to extract medical terms as many as possible from the text data by a simple longest-matching Perl script. After changing the given UTF-8 text to EUC format, the matching process required only 2 minutes including loading of a 10 MB dictionary into memory space with a desktop computer (Apple Mac Pro). From the 0.1 MB text document, 2,509 terms (including English spellings) were tagged and visualized in a color HTML format. Particularly focusing on the names of disease and symptoms, 893 terms were found with several mistakes and misusages. However, this process has a limitation in assigning ambiguous abbreviations and misspelled words. The simple longest-matching strategy may be useful as a preprocessing of medical reports.

1. INTRODUCTION
The Life Science Dictionary (LSD) project, founded in 1993, is a research project by us to develop a systematic database for life science (of course, including medical terms) and tools for the convenience of life scientists [1]. Our services are designed to provide and encourage access within the scientific community to the most up-to-date comprehensive information on English-Japanese translation dictionary of life science terms. In keeping with the users’ expectations, we have been enriching and refining the database according to a medical thesaurus compatible with MeSH (Medical Subject Headings) developed by National Library of Medicine, USA) thesaurus. Recent versions of LSD contains approximately 30 thousand headings with 200 thousand English and Japanese synonyms, consisting of the names of anatomical concepts, biological organisms, chemical compounds, methods, disease and symptoms.

One of the practical applications of thesaurus is text mining. For example, adverse drug events can be enriched and refined the database records to a medical thesaurus compatible with MeSH (Medical Subject Headings, biological, disease, molecule, method, and knowledge classified and marked by one of the following categories according to the.

2. METHODS
2.1 Dictionary
A tagger dictionary was made from LSD database as an EUC text file, which contains approximately 200,000 rows and 4 columns: (1) synomy terms, (2) subject heading terms, (3) categories, and (4) project heading ID (from MeSH). For the category of terms, all terms were classified and marked by one of the following categories according to the MeSH tree: anatomy, biological, disease, molecular, method, and knowledge (Fig. 1).

Fig. 1 Contents of tagger dictionary

2.2 Perl scripts
To take full advantage of the LSD in which many phrases have been registered, “the longest matches first” principle was adopted in the matching process. For this purpose, the tagger dictionary was sorted in the descending order by byte lengths, and term matching was performed for each of the dictionary entries in this order. For the sake of the speed of text matching in Perl language, both the text and the dictionary were first converted to EUC, encoding, and they were treated as byte strings in the matching process. Also, all two-byte roman characters were converted to corresponding ASCII characters, and multi-byte character strings were converted to ASCII strings, which were then converted to appropriate ASCII character(s) as far as possible.

For better readability of the resulting data as well as for the case of any secondary use, a standard HTML format was used as the output in which unique “class” attribute was assigned to each word of the category (Fig. 2A). This allows the user to extract more color and format the output data. The text mining tool is to be also used for the extraction of the desired set of terms.

To our surprise, there were many misspellings and typos, even in Japanese terms, in the test document (Table 4). In addition, there were also a few cases of misspellings and typos, even in Japanese terms, in the test document. For the test set containing 1,121 sentences, tagging process including UTF8-to-EUC conversion, 120 seconds were required with the Perl script by an Apple Mac Pro machine (1.8 GHz Quad-Core Intel Xeon, 16 GB memory). The speed of tagging seemed to be simply proportional to the length of the source text.

3. RESULTS
3.1 Speed
For the test set containing 1,121 sentences, tagging process including UTF8-to-EUC conversion, 120 seconds were required with the Perl script by an Apple Mac Pro machine (1.8 GHz Quad-Core Intel Xeon, 16 GB memory). The speed of tagging seemed to be simply proportional to the length of the source text.

3.2 Overall result
From the 0.1 MB text document, 2,509 terms (including English spellings) were tagged and visualized. The most abundant category was the names of disease and symptoms, and 893 terms were found (Table 1).

Table 1 Number of tagged terms

<table>
<thead>
<tr>
<th>Category</th>
<th>Tagged</th>
<th>Subcategory</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomy</td>
<td>459</td>
<td></td>
<td>Clinical test</td>
</tr>
<tr>
<td>Biological</td>
<td>891</td>
<td></td>
<td>Drug name</td>
</tr>
<tr>
<td>Biological (or Symptom)</td>
<td>895</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molecular (or Drug)</td>
<td>993</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method (or Index)</td>
<td>622</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other knowledge</td>
<td>185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2,569</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.3. Misspelling and typographical issue
In addition to many correctly-tagged terms, several patterns of misspelled or incorrect tags were found. The mostly misspelled terms were English abbreviations (Table 2). Especially, in the description of clinical test data, a variety of abbreviations were used, which cannot be matched. The meanings of 2- or 3-word abbreviations are ambiguous, we had omitted most of the abbreviations.

Table 3 Examples of partly-tagged words

Table 4 List of misspellings

3.4. Misspelling and typographical issue
In our experience, there were many misspellings and typographical errors, even in Japanese terms, in the text document (Table 4). Precise text matching did not use incorrect spellings that medical doctor can recognize their meanings.

Table 5 List of misspelled terms

4. DISCUSSION
With our tagging dictionary and scripts, most of medical terms were easily marked and visualized as an HTML document. From the 0.1 MB text document, 2,569 terms (including English spellings) were tagged and visualized in a color HTML format. Particularly focusing on the names of disease and symptoms, as much as 893 terms were found. Additional ‘mouse-over heading’ and web reference enables easy reviewing of the tagged terms.

Through this task, we have learnt the potential of thesaurus and scripts in finding medical terms from given Japanese texts. However, this process has a limitation in assigning ambiguous abbreviations and misspelled words. Moreover, there is an insurmountable difficulty to accomplish a ‘perfect matching’ with a fixed text dictionary, since improvement of thesaurus is a laborious work. The simple tagging strategy may be useful as a preprocessing of medical reports. Combination of natural text processing with this tool will be convenient for the practical use.

REFERENCES

You can see our tagged data at http://mol.pharm.kyoto-u.ac.jp/lsdp.html