Finding Every Medical Terms by Life Science Dictionary for MedNLP

Unable to attend
All correspondence should be addressed to Shuji Kaneko at skaneko@pharm.kyoto-u.ac.jp

Shuji Kaneko[1], Nobuyuki Fujita[2], Hiroshi Ohtake[3]
[1]Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan
[2]National Institute of Technology and Evaluation, Tokyo, Japan
[3]Center for Arts and Sciences, Fukui Prefectural University, Fukui, Japan

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 test 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 test 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, 893 terms were found with several mistakes and missings. 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 and 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 records to a medical thesaurus compatible with MeSH (Medical Subject Headings developed by National Library of Medicine, USA) thesaurus. Recent version 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 rapidly extracted by finding the causal relationship of drug treatment and related symptoms recorded in medical records. Favorably, our thesaurus contains a wide range of medical concepts as mentioned. In addition, we have previously developed a series of gloss-embedding Perl scripts for medical English texts [2]. In this study, therefore we aimed to tag every medical term (Japanese and English) as many as possible to evaluate the robustness of thesaurus and tagging program.

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) synonym strings, (2) subject heading strings, (3) category of term, (4) subject 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, molecule, method, and knowledge **(Fig. 1)**.

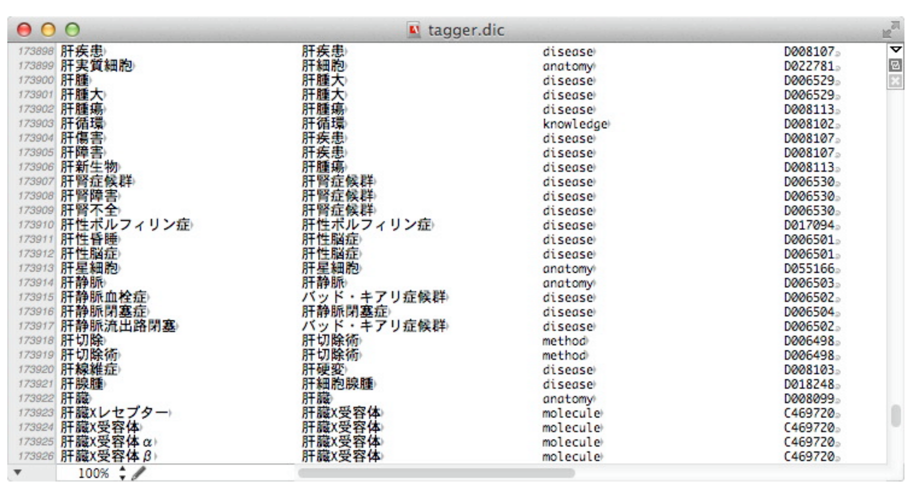

Fig. 1 Contents of tagger dictionary

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 of byte lengths, and text 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 characters unique in Unicode were converted to appropriate ASCII character(s) as far as possible. For better readability of the resulting data as well as for the ease of any secondary use, a standard HTML format was used as the output in which unique "class" attribute was assigned to each of the category **(Fig. 2A)**. This allows the users to customize text coloring even after the output of the data. We also added a 'mouse-over heading' feature, in which the embedded subject heading of the term will be displayed when the cursor was placed over the tagged term **(Fig. 2B)**. In addition, by clicking the tagged part, the user can confirm the thesaurus entry in our WebLSD online dictionary system.

(A)

(B)

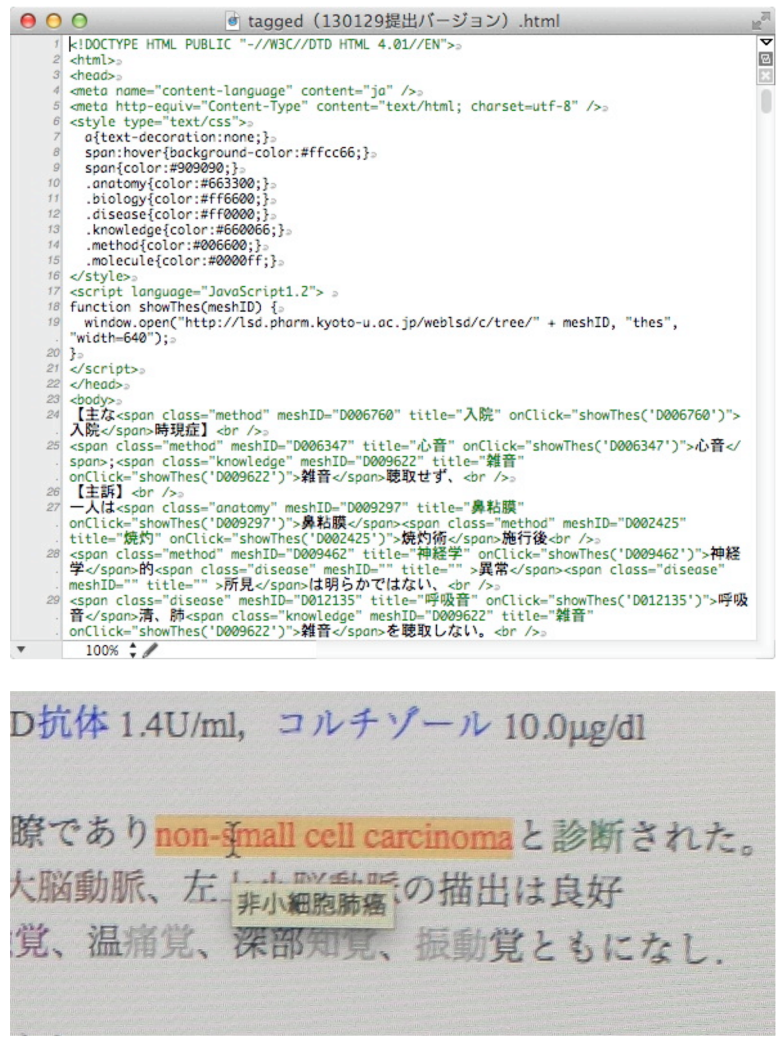

Fig. 2 HTML output (A) and mouse-over heading function (B)

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

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 our Perl script by an Apple Mac Pro machine (3.2 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 test document, 2,569 terms (including English spellings) were tagged and isolated. The most abundant category was the names of disease and symptoms, and 893 terms were found **(Table 1)**.

Table 1 Number of tagged terms

Category	Tagged
Anatomy	439
Biological	35
Disease (or Symptom)	893
Molecule (or Drug)	395
Method (or Index)	622
Other knowledge	185
Total	2,569

Table 2 List of missed abbreviations

Subcategory	Examples
Clinical test	T-Chol, Hb, Plt, eosino, BP, MPO, PaCO2, ALT, Cre, T-Bil, ZTT, APTT, etc.
Drug name	DIC（ダカルバジン）
	CLDM（クリンダマイシン）
	PIPC（ピペラシリン）
	PAPM/BP（パニペネム･ベタミプロン合剤）

3.3. Missed or incorrect tags

In addition to many correctly-tagged terms, several patterns of missed or incorrect tags were found. The mostly missed terms were English abbreviations **(Table 2)**. Especially, in the description of clinical test data, a variety of abbreviations were used, which cannot be marked. Since the meanings of 2- or 3-word abbreviations are ambiguous, we had omitted most of the abbreviations from tagger dictionary. However, if we know the part of document is apparently indicating clinical data, we can make a specific tagger dictionary for clinical tests. Similarly, some of the drug names were written in acronyms or non-universal abbreviations.

The most typical pattern of incorrect tag was 'partly-tagged' term **(Table 3)**. In these cases, part of unit concepts were registered in the dictionary, however, the combination of two or more concepts is common particularly in the names of disease and symptom, which were not completely covered in our thesaurus.

Table 3 Examples of partly-tagged words

Partial	Compounded	More complex case
温痛覚	Murphy徴候	眼球の黄染
顔面紅斑	心音不整	前頚部の腫脹
日光過敏	眼球結膜黄染	胆嚢軽度腫大
剥離爪	肺MAC症	下肺にはhoney comb

3.4. Misspelling and typographical issue

To our surprise, there were many misspellings and typographical errors, even in Japanese terms, in the test document **(Table 4)**. Precise text matching did not tag incorrect spellings that medical doctor can recognize their meanings.

Table 4 List of misspellings

In the text (Wrong)	Correct
predonisolone	prednisolone
theophyline	theophylline
Mycobacterium abcessus	Mycobacterium abscessus
Enterococcus fecalis	Enterococcus faecalis
Klebsiella pneumonoae	Klebsiella pneumoniae
コルトコフ音	コロトコフ音
グルドパ	グルトパ（Grtpa）
クオンテェンフェロン	クオンティフェロン

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 test 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 our 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

1. Kaneko S, Fujita N, Ugawa Y, Kawamoto T, Takeuchi H, Takekoshi M, Ohtake H. 2003. Life Science Dictionary: a versatile electronic database of medical and biological terms. "Dictionaries and Language Learning: How can Dictionaries Help Human & Machine Learning", Asialex, pp. 434-439.
2. Ohtake H, Kawamoto T, Takekoshi M, Kunimura M, Morren B, Takeuchi H, Ugawa Y, Fujita N, Kaneko S. 2003. Development of a genre-specific electronic dictionary and automatic gloss-embedding system. "Dictionaries and Language Learning: How can Dictionaries Help Human & Machine Learning", Asialex, pp.445-449.