Finding Specific Medical Terms Using the Life Science Dictionary for MedNLP

Shuji Kaneko
Graduate School of Pharmaceutical Sciences, Kyoto University
Kyoto, Japan
skaneko@pharm.kyoto-u.ac.jp

Nobuyuki Fujita
National Institute of Technology and Evaluation
Tokyo, Japan
fujitan@nifty.com

Hiroshi Ohtake
Center for Arts and Sciences
Fukui Prefectural University
Fukui, Japan
ohtake@fpu.ac.jp

ABSTRACT

We have been developing an English-Japanese thesaurus of medical terms for the past 20 years. The thesaurus is compatible with MeSH (Medical Subject Headings, developed by the National Library of Medicine, USA) and contains approximately 30,000 headings with 200,000 synonyms (consisting of the names of anatomical concepts, biological organisms, chemical compounds, methods, diseases and symptoms). In this study, we aimed to extract as many medical terms as possible from the test data by using a simple longest-matching Perl script. After changing a given UTF-8 text to EUC format, the matching process required only 2 minutes including the loading of a 10 MB dictionary into a 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 displayed in a color HTML format. In the case of names of diseases and symptoms, it was found that 893 terms had a number of formal errors and omissions. Furthermore, the matching process was found to have certain limitations in matching ambiguous abbreviations and misspelled words. In spite of this drawback, however, the simple longest-matching strategy may prove to be useful in the preprocessing of medical reports.

Keywords

Life Science Dictionary, Medical Thesaurus, MeSH

Team Name

LSDP (standing for Life Science Dictionary Project)

Subtask

Free Task (finding specific medical terms)

1. INTRODUCTION

The Life Science Dictionary (LSD) project, initiated in 1993, is a research project that aims to develop a systematic database for life sciences that provides medical terms and tools for the benefit of life scientists [1]. Our services are designed to allow the scientific community access to the most up-to-date and comprehensive information relating to the English-Japanese translation dictionary of life science terms. In line with the users' expectations, we have been enriching and refining the database records to a medical thesaurus compatible that is compatible with MeSH (Medical Subject Headings) thesaurus developed by the US National Library of Medicine. The most recent version of LSD contains approximately 30,000 headings with 200,000 English and Japanese synonyms, consisting of the names of anatomical concepts, biological organisms, chemical compounds, methods, diseases and symptoms.

One of the practical applications of the 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. In this regard, our thesaurus contains a wide range of medical concepts as mentioned above. 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 as many medical term (Japanese and English) as possible to evaluate the robustness of our thesaurus and tagging program.

2. METHODS

2.1 Dictionary

A tagger dictionary was made from LSD database as an EUC text file, which contains 4 columns and approximately 200,000 rows: (1) synonym strings, (2) subject heading strings, (3) term categories, (4) subject heading ID (from MeSH). As for the term categories, all terms were classified and marked as belonging to one of the following categories of the MeSH system: anatomy, biological, disease, molecule, method, and knowledge (Fig. 1).

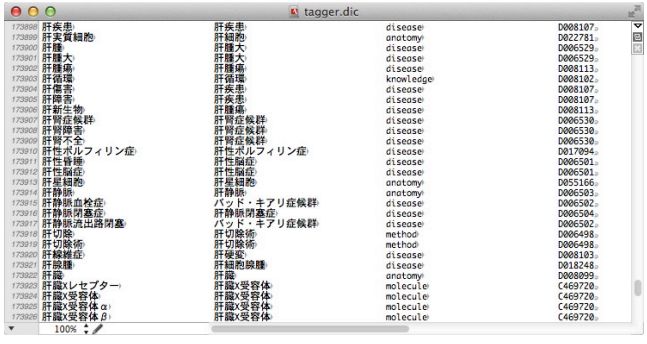


Fig. 1. Contents of tagger dictionary

2.2 Perl scripts

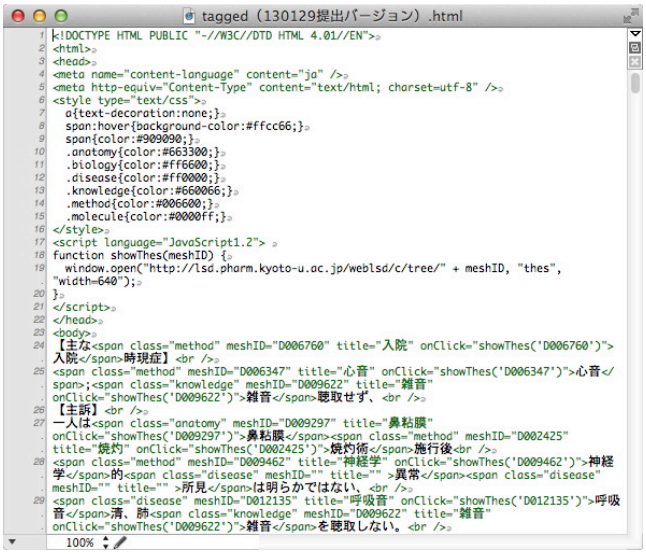
To take full advantage of the LSD, in which many phrases have been registered, "the longest matches first" principle was adopted in the matching process. For this purpose, the tagger dictionary

was sorted in the descending order of byte lengths, and text matching was performed for each of the dictionary entries in this order.

To enhance the speed of text matching in Perl language, both the text and the dictionary were first converted to EUC encoding, and were then treated as byte strings in the matching process. Furthermore, all two-byte roman characters were converted to corresponding ASCII characters, and multi-byte characters unique in Unicode were converted to appropriate ASCII characters as far as possible.

To improve the readability of the resulting data and to ease the difficulty of any secondary use, a standard HTML format was used as the output, in which a unique "class" attribute was assigned to each category (Fig. 2A). This allows 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 is 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

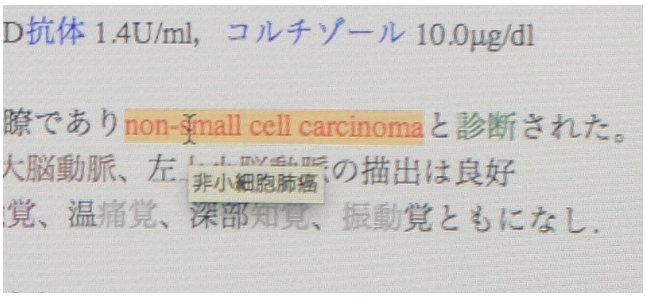


Fig. 2 HTML output (A) and mouse-over heading function (B)

3. RESULTS

3.1 Speed

For the test set containing 1,121 sentences, with a tagging process including UTF8-to-EUC conversion, 120 seconds were required with our Perl script on an Apple Mac Pro computer (3.2GHz Quad-Core Intel Xeon, 16GB memory). The speed of tagging seemed to correspond directly 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 found to be the names of diseases and symptoms, in which 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

3.3 Missed or incorrect tags

In addition to many correctly-tagged terms, several patterns of missed or incorrect tags were found.

The majority of missed terms were English abbreviations (Table 2). In particular, in the description of clinical test data, a variety of abbreviations were used that cannot be marked. Since the meanings of 2- or 3-word abbreviations are ambiguous, we omitted most of the abbreviations from the tagger dictionary. However, if we know that this section of a document is clearly 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.

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）

The most typical pattern of incorrect tag was 'partly-tagged' term (Table 3). In these cases, although part of unit concepts was registered in the dictionary, the combination of two or more concepts is common, particularly in the names of diseases and symptoms, and this was 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. Precise text matching did not tag incorrect spellings where medical doctor can recognize their meanings.

Table 4. List of misspellings

In the text	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 medical terms were easily marked and could be displayed as an HTML document. From the 0.1 MB test document, 2,569 terms (including English) that focused primarily on the names of diseases and symptoms, as many as 893 terms were found. The additional ‘mouse-over headings’ and web references could enable easy reviewing of the tagged terms.

Through this research, we have been able to assess the potential of our thesaurus and scripts in matching medical terms appearing in any given Japanese texts. However, this matching process has been found to have certain limitations in matching ambiguous abbreviations and misspelled words. Moreover, the difficulty of achieving a ‘perfect matching’ with a fixed text dictionary may well turn out to be insurmountable, since the constant attempt to improve a thesaurus is a considerable undertaking. Nevertheless, the simple tagging strategy may prove to be useful in the preprocessing of medical reports. In particular, the use of natural text processing in conjunction with this newly developed matching device should have a number of practical applications in the field of life sciences.

5. 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.