Evidence-Based Treatment
of Medical Guideline

Pingfang Tian[1,2], Zhonghua Zhu[1,2(✉)], and Zhisheng Huang[3]

[1] College of Computer Science and Technology, Wuhan University of Science
and Technology, Wuhan 430065, People's Republic of China
[2] Hubei Province Key Laboratory of Intelligent Information Processing
and Realtime Industrial System, Wuhan 430065, People's Republic of China
`webartisan@yeah.net`
[3] Department of Computer Science, Vrije University of Amsterdam,
1081 HV Amsterdam, The Netherlands

Abstract. Medical guidelines are recommendations on the appropriate treatment and care of people with specific diseases and conditions. Evidence-based medical guidelines are the document or recommendations which have been annotated with their relevant medical evidences, namely research findings from medical publications. We have observed the fact that there exist significant amount of medical guidelines have not yet annotated with relevant medical evidences, which becomes even more serious in the Chinese medical guidelines. In this paper, we propose an approach of evidence process of medical guidelines, such that we can find relevant evidences for those non-evidence-based medical guidelines. We develop a system called Link2Pubmed, which can retrieve the text which is described with a natural language and get the corresponding medical evidences. We use the word segmentation and part-of-speech tagging tools in natural language processing (NLP) to extract the keywords, and then translate them into corresponding English concepts in SNOMED CT, a well-known medical ontology. This system is an attempt to solve the existing problems in Chinese medical guidelines, which lack the annotations of relevant evidences.

Keywords: Medical guideline · NLP · SNOMED CT · PubMed · Link2Pubmed

1 Introduction

Clinical guidelines are the summary of the clinical practice work experience, the analysis of randomized controlled clinical study report and the consensus after discussion of recognized experts with high academic level. The guidelines have characteristics of openness and universality which can be used as a reference when clinical doctors deal with specified file [1] and clinical problems.

Clinical guideline is the bridge connecting the clinical evidence and clinical practice, reflects the status of the best clinical evidence. Due to its formulation process under specific racial, geographical, economic level, people's value and cultural factors,

using clinical research evidences in the native region is the most valuable. But influenced by medical conditions and the scientific research, the current Chinese medical communities are lack of a large number of clinical research evidences, especially the high quality randomized controlled study. So when making Chinese clinical guidelines, we often have to refer and use high quality clinical evidence abroad. In recent years, China has a certain amount of high quality clinical and randomized controlled clinical study [2]. However, the existing Chinese clinical guidelines are lack of the support of high quality evidences, which tend to be recognized by Chinese experts after discussion of consensus of high academic level, regarding as the recommendations.

For a clinical guideline with high quality, the most basic and most important issue is to make them evidence-based ones, including comprehensive collecting evidence and evidence for scientific and accurate evaluation.

Various academic groups in China encounter a lot of problems in the process of writing clinical diagnosis and treatment guidelines, the restricted factor is that the high quality randomized controlled clinical study is too little in China. Of course, we can't give up developing clinical guidelines because of a lack of the Chinese high level clinical research. Considering circumstances influenced by some factors, we can choose high-quality randomized controlled clinical studies in those countries which are also suitable for China. So, many Chinese scholars retrieve evidences on medical website abroad [3] (e.g. NGC [4], Cochrane Library [5, 6]) by manual, which is time consuming.

In this paper, we propose an approach of evidence process of medical guidelines, such that we can find relevant evidences for those non-evidence-based medical guidelines. We develop a system called Link2Pubmed, which can retrieve the text which is described with a natural language and get the corresponding medical evidences. We use the word segmentation and part-of-speech tagging tools in natural language processing (NLP) [7, 8] to extract the keywords, and then translate them into corresponding English concepts in SNOMED CT [9, 10], a well-known medical ontology. This system is an attempt to solve the existing problems in Chinese medical guidelines, which lack relevant evidences. Our initial experiments show that our approach is efficient for the target.

The rest of this paper is organized as follows: Sect. 2 presents the general idea of Link2Pubmed. Section 3 shows the system processing flow of Link2Pubmed in detail. Section 4 presents the design and implementation of the system module. Section 5 shows the implementation details, system test and evaluation. Section 6 discusses the related work, future work, and makes the conclusions.
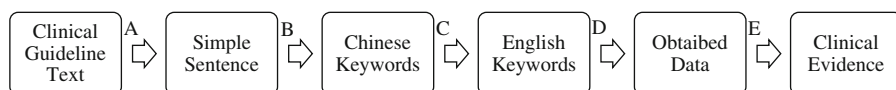
2 Link2Pubmed System

For the situation that many Chinese clinical guidelines lack annotated with relevant medical evidences, this paper proposes a solution to evidence-based treatment of existing medical guidelines, namely according to the fact guidelines described to find corresponding medical evidence. Due to the current status of lacking evidence in Chinese clinical guidelines, our approach is to retrieve evidences on PubMed, and converts the data obtained from PubMed into ones with required formats, and finally present them in the user interface of the system.

PubMed [11] is a free search engine, providing search and a summary of biological medicine. Its database source is MEDLINE. Its central theme is medicine, but also includes other medical field related to it, such as nursing or other health disciplines. It also provides support for the related biomedical information on the quite comprehensive, such as biochemistry and cell biology. Journal articles in free PubMed information service do not include the full text, but may provide links to the provider. MEDLINE collected articles from 1966 until now, including medical, nursing, veterinary medicine, health care system and literature of the preclinical science article. These data from more than 70 countries and regions, more than 4800 biomedical journals, in recent years, the data involved in more than 30 languages, dating back to 1966 years of data involved in more than 40 languages, around 90 % of English literature, 70 % ∼ 80 % of literature with the author to write English abstract. So the retrieved medical evidence on PubMed can guarantee finding enough medical evidence.

However, retrieval on PubMed needs to provide medical subject headings which will be retrieved, so for the guidelines described in natural language text, they must be transformed from Chinese natural language description guideline to medical subject headings which can be retrieved on PubMed. This paper designs and implements the system, Link2Pubmed. From the perspective of guideline text described in Chinese natural language, this paper extract keywords from guideline text, then converted them into corresponding medical subject headings, then go on PubMed search. In the third section of the article tells the whole system design process.

3   Link2Pubmed System Processing Flow

The process of Link2Pubmed system design is shown in Fig. 1. The figure described the whole system of the Link2Pubmed data processing flowchart.

Fig. 1. Link2Pubmed system flowchart

In the whole system, there are 6 kinds of forms of data: Clinical Guidelines Text, Simple Sentence, Chinese Keywords, English Keywords, Obtained Data, Clinical Evidence. There are 5 kinds of options between the 6 kinds of forms data, which are:

A. Semi-automatic process for clinical guidelines. Through semi-automatic processing, Clinical guideline texts generate the corresponding simple sentence patterns of text;
B. Extracting keywords. For simple sentence text, keyword extraction can get the Chinese keywords;
C. Keywords conversion from Chinese to English. In order to retrieve the keywords extracted from guideline text in PubMed search, we must transform Chinese keywords into the corresponding English medical concepts.

D. PubMed retrieval. Retrieving in PubMed using translated English keywords can get the corresponding medical evidence.
E. Data formatting. The data retrieved from the PubMed contains large amounts of information, the information expressed in XML format, in order to make the user more intuitive to see the data; we must manipulate the data format in another way.

3.1 Semi-Automatic Process for Clinical Guidelines

Because of the limitation of the current natural language processing technology, it can't implement word segmentation and part-of-speech tagging [12, 13] with a whole guideline or a large section of a guideline. This is because with the increase of processing text length, the consumed time that the word segmentation tool spend to deal with text will have exponentially times growth, at the same time, the accuracy of the result of word segmentation and part-of-speech tagging will be greatly reduced. So when to extract keywords in the guideline, the object to deal with should be a simple sentence, rather than a large section of the text. So, transforming guideline text into simple sentences after word segmentation and part-of-speech tagging is a good solution. But medical guidelines which are described by natural language can't simply use punctuation marks to divide the text to achieve the purpose of transforming guideline text into a simple sentence because of the syntactic structure, grammatical structure and other logical relationship of natural language. There is no such tools can smartly implement text segmentation with a description of natural language. So we can only use a semi-automatic method, namely using artificial methods in natural language processing, analyzing the text structure and the logical relationship between sentences guide and for text segmentation.

3.2 Extracting Keywords

For the simple sentence after semi-automatic processing, keywords extracted from them involved in symptoms, disease, medicines and medical terms, etc. We can know what the meaning of the sentence with these words is. In this design, the word segmentation tool is ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) [14] which comes from the Chinese academy of sciences. The main features include: Chinese word segmentation, the part of speech tagging, named entity recognition; new word recognition, etc., at the same time, it also supports custom glossary, which greatly help to extract keywords and phrases in guidelines like symptoms, disease, medicines and medical terms, etc.

3.3 Keywords Conversion from Chinese to English

After getting the keywords, we can't retrieve them before translation. Using some medical dictionary for translation, it can be done for key words or phrases expressed meaning transformation in both English and Chinese, but it often doesn't present well

for these words and phrases have corresponding medical concepts. SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms), is a system organized, an advantageous set of the medical term for computer processing, covering most aspects of Clinical information, such as disease, can see, the operation and microorganism, drugs, etc. Using the term set can be coordinated between different disciplines and specialties, and the location of the care for clinical data indexing, storage, retrieval, and aggregation. This greatly improves the accuracy of the translation.

3.4   PubMed Retrieval

PubMed offers open data query link, the user can send the request directly to call this interface. What we need to do is using English keywords to generate a request URL, and then sends an HTTP request to the PubMed, it will return the relevant data and information.

3.5   Data Formatting

Information retrieved from a PubMed is expressed in XML format which store huge amounts of information and only a few of this information is really needed to users, which allows the user effectively get useful information. In order to present the retrieved data more clearly to the user, the system will output data in a form that a user looks more intuitive after parsing XML.

4   System Module Design and Implementation

Based on the system processing flow, Link2Pubmed system module design is shown in Fig. 2:

4.1   Guideline Semi-Automatic Processing

To deal with the text of the guideline described in Chinese, before and after the Chinese syntactic structure and logic is relatively complex, so in this system the semi-automatic method is adopted to process guideline text. But in the process of dealing with guideline text, the following points must comply with:

First, must make the pledge that the meaning of the sentence is as same with the original text. Namely, in the process of processing, convert guideline text equivalent into simple sentences. This is one of the most important in the process of guide semi-automatic processing.

Second, pay attention to the logical relation of the sentence with the other before and after it. In the guideline described in Chinese, due to the logic relationship with the sentence before and after it, sometimes a few words before and after the same aspect of description are presenting the same truth, so we should combine several sentences into one sentence, describing a fact.
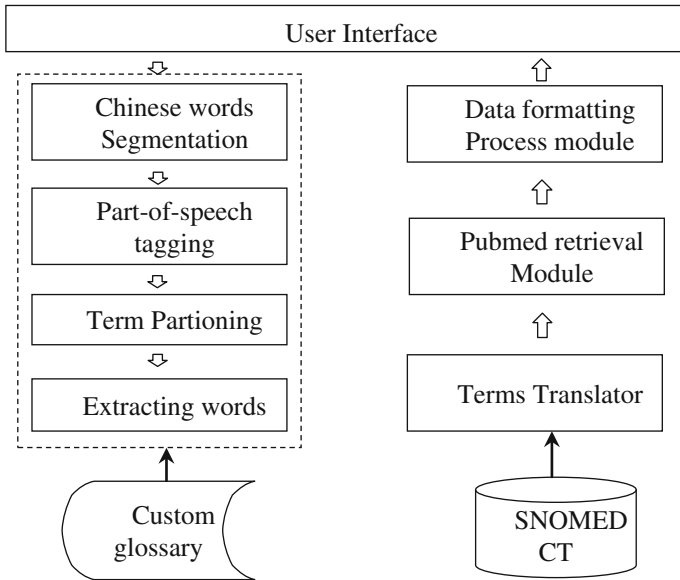
Fig. 2. Link2Pubmed System architecture diagram

Third, pay attention to the hierarchical relationships in a single sentence. In the Chinese guideline, a sentence can have more clauses, describing various aspects of a fact, the amount should be subdivided each clause, also should be a sentence to describe a fact.

The sentences after semi-automatic processing should be clear, eventually a sentence to describe a fact which has no superfluous information and doesn't miss any information.

Figure 3 is a text taken from Chinese antimicrobial treatment guidelines, which describes what rules should pay attention to when using the penicillin. In this text, the sentences are combinations of complex sentences, which describe the various aspects of a fact in one sentence, just the fact that it can be subdivided into several facts, namely one sentence describes only fact. Figure 4 is the result of semi-automatic processing, the original of the complex sentence is broken down into several simple sentences, and each simple sentence describes only one truth.

1. 无论采用何用种给药途径，用青霉素类药物前必须详细询问患者有无青霉素类过敏史、其他药物过敏史及过敏性疾病史，并须先做青霉素皮肤试验。

2. 过敏性休克一旦发生，必须就地抢救，并立即给病人注射肾上腺素，并给予吸氧、应用升压药、肾上腺皮质激素等抗休克治疗。

3. 全身应用大剂量青霉素可引起腱反射增强、肌肉痉挛、抽搐、昏迷等中枢神经系统反应（青霉素脑病），此反应易出现于老年和肾功能减退患者。

Fig. 3. Segment before semi-automatic processing

用青霉素药物前，必须询问患者有无青霉素类过敏史。
用青霉素药物前，必须询问患者有无其他药物过敏史。
用青霉素药物前，必须询问患者过敏性疾病史。
用青霉素药物前，必须先做青霉素皮肤试验。
过敏性休克一旦发生，必须就地抢救。
过敏性休克一旦发生，应立即给病人注射肾上腺素。
过敏性休克一旦发生，应立即给病人吸氧。
过敏性休克一旦发生，应用升压药抗休克治疗。
过敏性休克一旦发生，应用肾上腺皮质激素抗休克治疗。
全身应用大剂量青霉素可以起腱反射增强。
全身应用大剂量青霉素可以起肌肉痉挛。
全身应用大剂量青霉素可以起抽搐。
全身应用大剂量青霉素可以起昏迷。

Fig. 4. Segment after semi-automatic processing

4.2 Keyword Extraction Module

Custom glossary. To extract reasonable keywords from a sentence, the first thing needs to know is which words should be extracted. ICTCLAS is a tool that can customize the word in the glossary, and in this glossary we can specific a part of speech for a specified custom word, for example, "penicillin" can specify part of speech of "YP", namely "drug"; We can also specify the "fever" as "ZZ", namely the symptoms. Custom glossary is a large library, and all of the keywords to be extracted are included in the library. The repository contains all kinds of drugs, symptoms, disease, and qualitative values and other medical terms. The complement of the custom glossary determines the keyword extraction accuracy.

Word segmentation and part-of-speech tagging. Using ICTCLAS, sentences can be divided into individual words, separated by spaces, at the same time for each term labeling part of speech, so that you can pick out the word with an annotation according to the specific part of speech, which are the keywords to be extracted [15]. ICTCLAS is a good tool, in this operation, what we need to do is to load the custom glossary and invoke the ICTCLAS segmentation and part-of-speech tagging interface.

Word segmentation and extraction. After word segmentation and part-of-speech tagging, sentence is still in the form of a string, but we can use the characteristics of the string after marking to implement segmentation. After word segmentation and part-of-speech tagging, we can get a combination of words and part of speech, and there is a space separated between words and words, this is really the whole sentence can be divided into individual words, choose the custom part of speech of words, which is in the end we want to extract.

Translation Module. We use SNOMED CT, the well-known medical ontology, for the translation support in the translation module. The main advantage to use a medical

ontology for the translation process is that it provides a standard terminology set for the medical domain. We obtain the standard set of the medical concepts with their English labels and use the Google translation to obtain their corresponding Chinese terms [16].

We design a local user translation dictionary with both English and Chinese terms, which consist of concept ID in SNOMED CT, type of the term, such as YP, which stands for drug, and others, the English term, and the Chinese term, like this:

61651006|YP|Cefamandole|头孢孟多
61651006|YP|Cefamandole|头孢孟多(物质)
61651006|YP|Cefamandole|头孢孟多(产品)
61651006|YP|Cefamandole|羟苄四唑头孢菌素
61651006|YP|Cefamandole (substance) |头孢孟多
61651006|YP|Cefamandole (substance) |头孢孟多(物质)
61651006|YP|Cefamandole (substance) |头孢孟多(产品)
61651006|YP|Cefamandole (substance) |羟苄四唑头孢菌素
61651006|YP|Cefamandole (product) |头孢孟多
61651006|YP|Cefamandole (product) |头孢孟多(物质)
61651006|YP|Cefamandole (product) |头孢孟多(产品)
61651006|YP|Cefamandole (product) |羟苄四唑头孢菌素
61651006|YP|Cephamandole|头孢孟多
61651006|YP|Cephamandole|头孢孟多(物质)
61651006|YP|Cephamandole|头孢孟多(产品)
61651006|YP|Cephamandole|羟苄四唑头孢菌素
61862008|YP|Methicillin|耐甲氧西林
61862008|YP|Methicillin|耐甲氧西林(物质)
61862008|YP|Methicillin|甲氧西林(产品)

Although there might exist multiple matching in the translation dictionary, we would always prefer a translated term which is shortened to others. For those which cannot be translated by using the local translation dictionary, we use the Baidu translator to do the complement work.

4.3   PubMed Retrieval Module

Translating module will output English keywords. If we want to get the PubMed IDs (PMIDs) for articles about breast cancer. The query string can be written as follows: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=breast +cancer. If you want to know more about Entrez Programming, you can log on http://www.ncbi.nlm.nih.gov/books/NBK25500/ for more information.

4.4   Data Formatting Module

The XML parser is used to convert the parse XML data into a particular data structure. It only stores some fundamental information such as title, authors, PMID, journal,

abstract, published time. Link2Pubmed use a structure array to store these properties which have been formatted.

5   Implementation, Test and Evaluation

5.1   System Implementation

After starting Link2Pubmed, we will see the interface as shown in Fig. 5. The system interface is very simple, only one input box and one output. But when we input a Chinese sentence after semi-automatic processing in the input box, it will obtain the related medical evidence after retrieval.

Through semi-automatic processing, we get the Chinese guideline text composed of some simple sentences. For example, there is now a simple sentence after semi-automatic processing: "如果患者是妊娠期患者,那么应避免使用甲硝唑" (If the patient is a patient with pregnancy, so avoid using metronidazole.). Input this simple sentence in the "Source" Column of Link2Pubmed system. Through the analysis of the Link2Pubmed, it can extract two Chinese keywords: "妊娠" (pregnancy) and "甲硝唑" (metronidazole), and it can convert to the corresponding English medical terms "Pregnancy" and "Metronidazole". After retrieval, we can get the results shown in Fig. 5. In Fig. 5, we can see some medical evidence in the results section lists, and each piece of evidence is given some basic information such as title, author, journal name
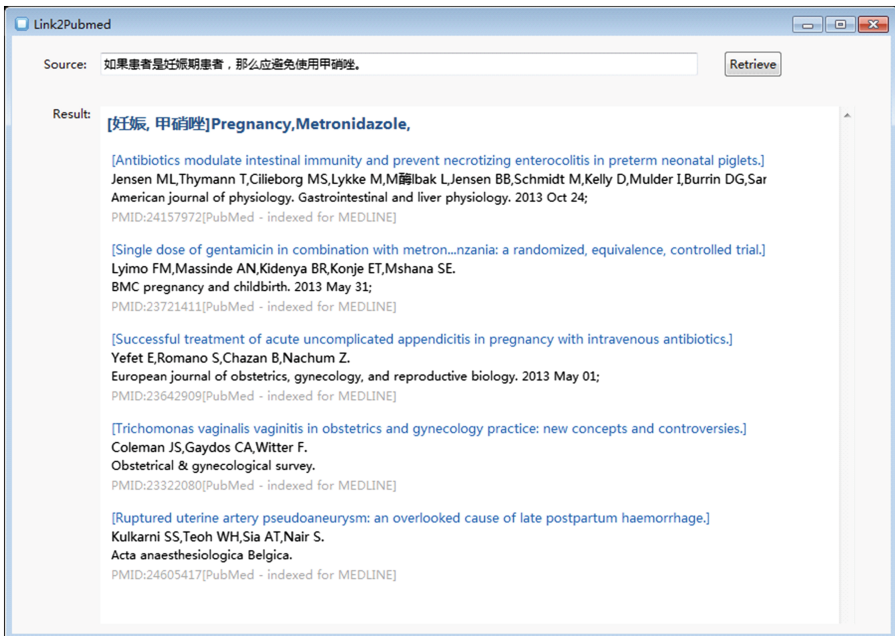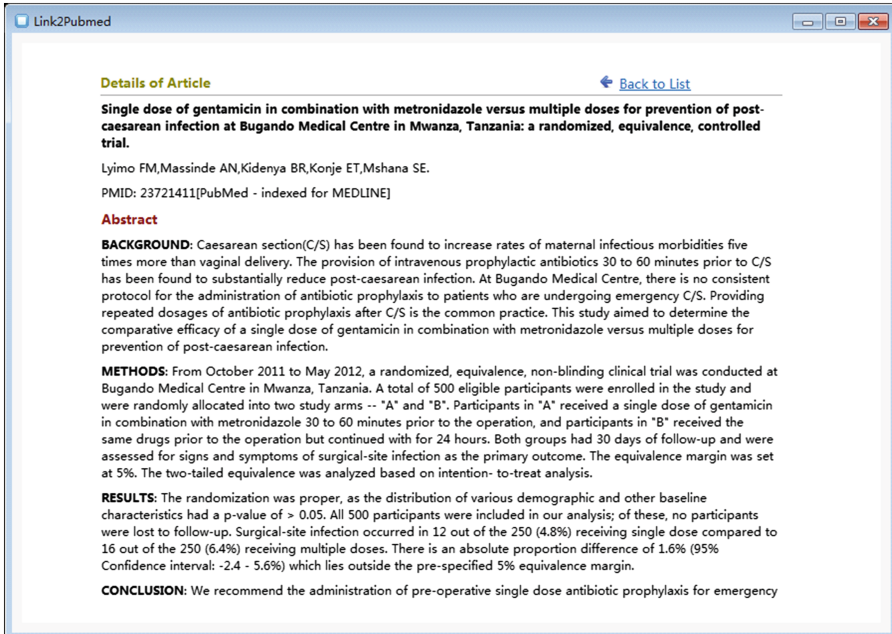

Fig. 5. Link2Pubmed Retrieval Result

Fig. 6. Evidence Details

and PMID. Click on the title of each piece of evidence to view the details of this evidence, as shown in Fig. 6. In Fig. 6, it shows the abstract of the evidence in detail.

5.2   Experiment and Evaluation

We use the Chinese clinical guidelines for rational use of antibiotics as the test data for the experiment. We select 100 statements after semi-automatic processing as test data of Link2Pubmed randomly. We can define a stipulated as follows: KA1 is the amount of keywords in the sentence, KA2 is the amount of keywords which be found, KA3 is the amount of keywords in the sentence which exist in custom glossary, FP1 is the percentage of keywords being found actually (KA2/KA1), FP2 is the percentage of keywords being found in Link2Pubmed with custom glossary (KA2/KA3), TP is the accuracy completion percentage of translation, and IA shows whether or not the evidence is available. We calculate the average value of FP1, FP2, TP and IA of each test sentence Meanwhile, we set two groups of data which are respectively with uncompleted custom glossary and completed custom glossary. The results are shown in Fig. 7.

We can see the item FP1 in Fig. 7, as long as the keywords of test sentences are already defined in the custom glossary the words will be found out in Link2Pubmed, but in fact the custom glossary is uncompleted, the number of words didn't contain all medical vocabulary, which makes that the actual keyword extracting percentage is 76.7 %, not 100 %. At the same time, we should pay attention to that the average IA, which means whether the evidence is available, is only 0.4. This is directly related
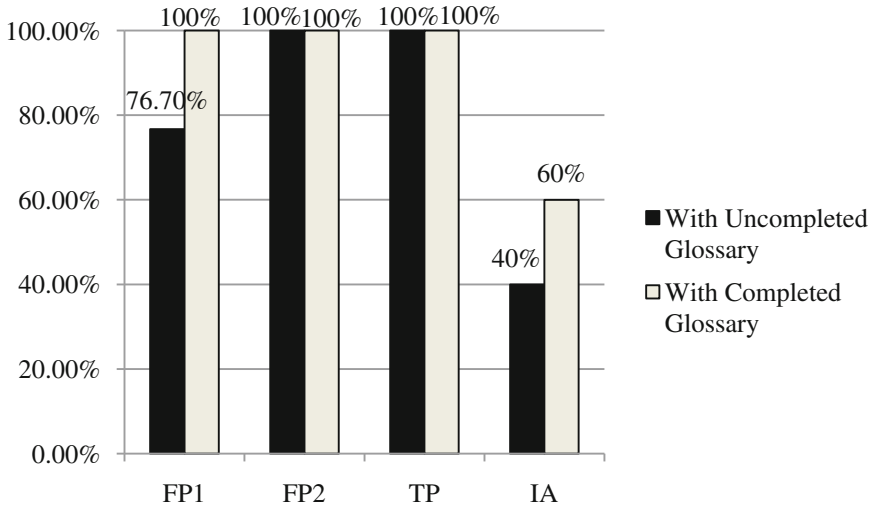
Fig. 7. Retrieval Result

to the accuracy and completeness of keyword extracting. With the completed custom glossary of test data, the actual keyword extracting percentage is 100 %, which also improves the availability of the evidences in a degree, this mainly makes up for the factors that the availability of evidences is not high caused by incomplete keyword exacting. One another factor determining the availability of evidences is the accuracy of keyword translation; on the other hand, there is indeed no corresponding evidence in PubMed. However, the system can be improved if a high-quality custom glossary is provided.

6   Conclusion and Future Work

In order to solve the problem that the Chinese clinical evidence is insufficient, Link2Pubmed provides an approach to process the existing medical guidelines for evidence-based treatment. Link2Pubmed can retrieve guideline text, and automatically extract keywords, and convert them to the corresponding medical concepts, and retrieve on PubMed accurately and effectively, so as to get a lot of medical evidences.

Link2PubMed provides a tool for Chinese researchers for medical evidence retrieval. Here are some future works we are going to improve the system:

1. The keywords are not able to be maximized extracted. Keyword extraction mainly depends on the custom glossary. So, a glossary which covers more medical terms will make the extraction of keywords more completely.
2. The keywords between English ones and Chinese ones are not totally transform-able. Some extracted Chinese keywords may not find their corresponding concepts in SNOMED CT. Under that situation, we have to use some additional dictionaries to translate them into corresponding English terms. The cost of guidelines'

semi-automatic processing is high. Depending on the current level of natural language processing, we cannot reasonably implement the guideline text conversion, but with the development of natural language, the quality will be improved.

One of the main features of Link2Pubmed is that we use the medical ontology such as SNOMED CT in our system. As we have discussed before, the main advantage to use SNOMED CT in our Link2Pubmed system is that it provides a standard terminology set in the medical domain. That is useful not only in the fragmentation and POS processing in the natural language processing tool, but also useful for the translation process. That shows that the Semantic Web technology is useful for the evidence process of medical guidelines.

References

1. Xiao-wei, J., Zhu-ming, J.: The present status of Chinese clinical practice guidelines. Chin. J. Clin. Nutr. **18**(6), 327–329 (2010)
2. Jiang, Z.M., Wilmore, D.W., Wang, X.R., et al.: Randomized clinical trial of intravenous soybean oil alone versus soybean oil plus fish oil emulsion after gastrointestinal cancer surgery. Br. J. Surg. **97**(6), 804–809 (2010)
3. He, L., Du, X.L.: Retrieval research of evidence based clinical practice guidelines. J. Clin. Rehabilitative Tissue Eng. Res. **11**(40), 8173–8177 (2007)
4. Mast, K.R., Salama, M., Silverman, G.K., Arnold, R.M.: End-of-life content in treatment guidelines for life-limiting diseases. J. Palliat. Med. **7**(6), 754–773 (2004)
5. Kennedy, S., Bergqvist, A., Chapron, C., D'Hogghe, T., et al.: ESHRE guideline for the diagnosis and treatment of endometriosis. Hum. Reprod. **20**(6), 2698–2704 (2005)
6. Burton, M.J., Couch, M.E.: Rosenfeld. R.M.: Extracts from the cochrane library homeopathic medicines for adverse effects of cancer treatments. Otolaryngol. Head Neck Surg. **141**(2), 162–165 (2009)
7. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceeding of the 25th International Conference on Machine, pp. 160–167 (2008)
8. Chopra, A., Prashar, A., Sain, C.: Natural language processing. Int. J. Enhancements Emerg. Eng. Res. **1**(4), 131–134 (2013)
9. Elkin, P.L., Brown, S.H.: Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin. Proc. **81**(6), 741–748 (2006)
10. Wasserman, H., Wang, J.: An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu. Symp. Proc. **2003**, 699–703 (2003)

11. 沈锡宾, 吕小东, 郝秀原, 孙静, 汪谋岳, 郭利劭. PubMed Central简介及其对期刊的评估和收录. 中国科技期刊研究 **17**(5), 866–868 (2006) (XiBin, S., XiaoDong, L., XiuYuan, H., Jing, S., MouYue, W., LiShao, G.: PubMed central introduction and a review and record of journal. Chin. J. Sci. Tech. Periodicals **17**(5), 866–868 (2006))

12. Jiang, W., Huang, L., Liu, Q., Lü, Y.: A cascaded linear model for joint chinese word segmengation and part-of-speech tagging. In: Proceeding of ACL, pp. 897–904 (2008)

13. Jiang, W., Mi, H., Liu, Q.: Word lattice reranking for Chinese word segmentation and part-of-speech tagging. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp. 385–392 (2008)

14. Zhang, H., Yu, H., Xiong, D., Liu, Q.: HHMM-based Chinese lexical analyzer ICTCLAS. SIGHAN '03 Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp. 184–187 (2003)

15. 夏天, 樊孝忠, 刘林. 利用JNI实现ICTCLAS系统的Java调用. 计算机应用 **24**(Z2), 177–178 (2004). (Xia, T., Fan, X., Liu, L.: Java calls using JNI ICTCLAS system. Comput. Appl. **24**(Z2), 177–178 (2004)

16. Ceusters, W., Smith, B., Kumar, A.: Dhaen, A.: Ontology-based error detection in SNOMED-CT. In: Proceedings of MEDINFO 482–486 (2004)