Performance Evaluation for Strategy Based on Auto-Adapting Users in Cross Language Information Retrieval

Zhongjian Wang

Harbin University of Commerce, Harbin 150028, China
Harbin Far East Institute of Technology, Harbin 150025, China
wangzhj@hrbcu.edu.cn

Abstract

With the different query the different retrieval results will be got. According to the viewpoint of each user has different query need, a method that can adapt user automatically are proposed in information retrieval. To detect the information demands of users from multilanguage electronic text, the method analyzed the keywords used on information retrieval and gathered those keywords, summarized the relationship between keywords, selected accuracy word to word translation, deal with ambiguity of words to build reference information by learning and feedback process. The retrieval system is able to auto-adapt retrieval demand of different users by renewing reference information. Evaluation experiment results indicate the trend of adapting user and the availability of the method.

Keywords: *Cross language, Information retrieval, Auto-adapting user*

1. Introduction

With the developing and popularity of Internet and the fast increasing of electronic text, the research of natural language processing has got a widespread concern by many researchers, Such as NLP technology is used in the protection of privacy in E-communication [1], the research of word sense disambiguation [2] uses Information Gain to calculate the weight of different position's context to construct the feature vectors, and get an improved Bayesian model. Especially the application of information retrieval is more and more extensive. Although many information retrieval systems have been developed, they almost use single language and the Boolean matching strategy by means of keyword search techniques for the simplicity to implement. There is also more information retrieval system for multilanguage information retrieval, and they are ability to search same information by different language, but still have some problems.

A multilanguage electronic information retrieval tool for sorting out a large quantity of electronic text information and for finding desirous information for users by Internet become increasingly important, and the text information that user want to find are expressed by different language, usually users can not find requisite information by search queries in single language.

For single language information retrieval method, a thesaurus was constructed by extracting interrelated words of request keywords from search target text gather with statistic method. The thesaurus is used to information retrieval system [3]. The method was proposed for extracting the information needs of WWW search system users, by analyzing the search keywords list logged by the system. Paper [4] presents an intelligent information retrieval system; the system is based on automatic thesaurus construction and used to document

clustering and classification. Paper[5] proposed a called after the sequential searching model, the algorithm calculate the relationship between each search keywords logged, and group them into several groups, which is supposed to represent the information needs of users. Paper[6] proposed a method to realize an interactive guidance mechanism for document retrieval system, developed a user-interface which presents users the visualized map of topics at each stage of retrieval process and extracted automatically topic words by frequency analysis, measured the relationship among topic words by their co-occurrence.

For Multilanguage information retrieval system method, paper [7] proposed a theoretically grounded alternative, which uses sense disambiguation based upon context terms within the source text, introduced the concept of translation probabilities incorporating a context term and extends. Experiment results show availability of disambiguation. Paper [8] presents an approach that computes translation probabilities for a given query by using only a bilingual dictionary and a monolingual corpus in the target language. The method combines term association measures with an iterative machine learning approach based on expectation maximization. Their approach considers only pairs of translation candidates and is therefore less sensitive to data sparseness issues than approaches using higher n-grams. The learned translation probabilities are used as query term weights and integrated into a vector-space retrieval system. It used in English-German cross-lingual retrieval show improvements over a baseline using dictionary lookup without term weighting. Many researches focus on the web information retrieval field and utilize their own web crawlers to crawl, index, and analyze contents of the pages and network structure of the web [9]. But we are only interested in the retrieval results of natural language contents; each has its own research challenges and problems.

In this paper, we extend request keywords gather by a thesaurus and get translation keywords by a dictionary. The method deals with ambiguity by user feedback process and learning process. Along with the processing for search record, get the information in target text and search results text, extract common words in correct results and erroneous results, also use them as reference information for following retrieval. The system is consisted of translation processing, retrieval processing, and feedback processing and learning processing. The following we will introduce each part of the method and confirm the usefulness of the method by experiments.

2. Outline of the Method

In general, a variety of keywords are used to retrieve information on the same topic. These keywords are different according to each user with various viewpoints. When search text is multilanguage, retrieval system must find the information of user's longing, witch are described by different language.

There are two problems for a keyword retrieval information system:

The first problem is that various keywords are used to retrieval information for the same retrieval topic. Users use differ keywords because they has differ viewpoint. So that, retrieval system must gather and filtrate those keywords, correlative synonyms to find all demand information of users.

The second problem is that keywords of some language must be mapping to a keyword gather by different language.

When different request keywords are enlarged with synonyms by a thesaurus, the keyword gather will be a large quantity with different language. That is important how to select synonyms, how to filtrate important keywords that has high priority of express retrieval demand of users and how to select translation words by dealing with ambiguity.

The flowchart of the method is shown in Figure 1.

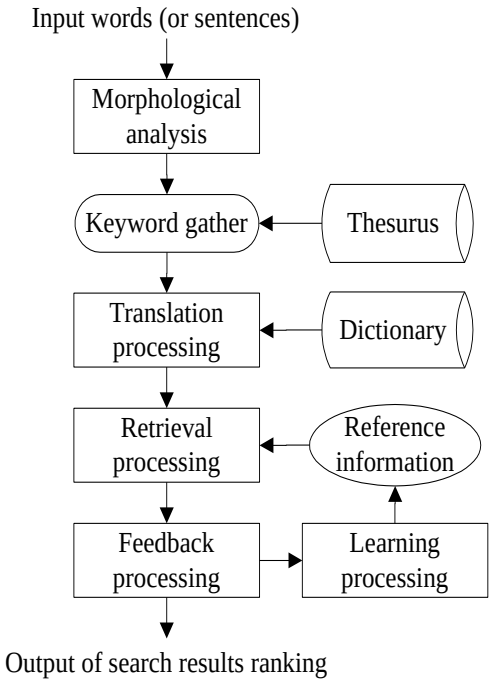

Figure 1. Outline of the method

At first, input a sentence, a phrase or a word. The system processes the input by morphological and syntactical analysis, and then system extracts common words from input that have been processed. The system gets all synonyms of common words and translation words. The system searches all sentences including the extracted words and including translation words from different language. The system ranks all the retrieved results by means of their order of priority for all sentences or phrases with different language.

2.1. Morphological Analysis

We use the ChaSen as tool of Morphological for a Japanese sentence [10] and the ICTCLAS as tool of word segmentation for Chinese sentences [11]. The word is gathered and deletes those function words by referring to function word list. The function word is those words that don't express fact meaning, Such as particle words in Japanese and auxiliary words in Chinese.

2.2. Translation Processing

Mapping the words gather of input sentences to words of differ language by Japanese to Chinese (Chinese to Japanese) and Japanese to English (Chinese to English) dictionary.

Sometimes the words of queries are untranslatable, because there are not in the dictionary. Such as an apparent word is a new word or a compound word. The mostly untranslatable words of query generally in dictionaries are new words, compound words, people's name, the name of organization and special terms. Paper [12] proposed a method of cross language information retrieval based dictionary. The key problem is the translation of word to word.

The main problem and difficulty are how to solve the translation ambiguity in the method of dictionary-based cross language information retrieval [13-15]. To solve this problem, we use feedback process and learning process to reduce the ambiguity of word translation. Because user selects desired search results then the selection procedure is a supervised learning. This supervised learning is carried out automatically when user selects search results.

2.3. Retrieval Processing

Retrieval processing uses the keywords gather that consist of synonyms and translation words as retrieval query. In order to get all retrieval result that is relational demand of user, we extend keyword gather by using similarity word and synonym of keywords as query. Otherwise reference information that got by feedback processing and learning processing is used on searching.

2.4. Feedback Processing

For the results of information retrieval, the search engine gives generally a list with short summary. User selects the desired results. This process is a supervised learning.

Retrieval results are ranked at priority by likelihood evaluation function, LEF is shown as formulation (1) that is calculated by number of common words in correct search results and erroneous search results. User selects the desired search results and unused search results then the system extracts the common words in selected results. Reference information is consist of word pairs obtained from retrieval results that user selects.

At beginning, reference information is empty. The system gains reference information of word translation by repeating process of feedback processing. That reference information can not only deal with ambiguity, but also can filter retrieval results and let the retrieval results adapt users.

2.5. Learning Processing

To improve the adaptability of method for user, statistic method is used. The system calculates the appearance frequency of the common words in correct search results and erroneous results that was returned from feedback processing. By morphological analyzing, the system gets a sequence of words of search result sentences which is correct or erroneous judged by user.

Following Figure 2 is an illustration of reference information. Learning processing gets reference information by calculating appearance frequency of common words, then retrieval processing uses reference information to improve search effect.

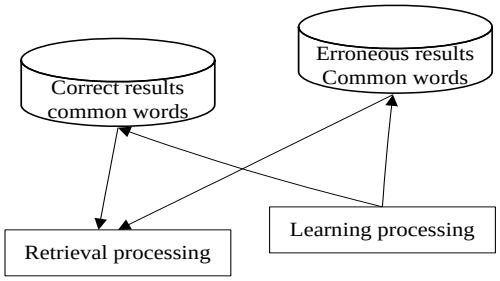

Figure 2. Reference information

Ranking of retrieval result is according to the LEF, called value of likelihood evaluation function; calculate by formulation (1):

$$LEF = \alpha \cdot \frac{\sum(CH)}{\sum(S)} - \beta \cdot \frac{\sum(EH)}{\sum(S)} \quad (1)$$

Here α and β are coefficients, CH, EH and S are the number of correct words in retrieval text, the number of erroneous words in retrieval text and the length of the character string respectively. The coefficient of LEF is decided by greedy method in preliminary experiments.

2.6. Preliminary Experiment

To decide the optimum coefficient of LEF, we carried out a preliminary experiment. We use three kinds of text: engineering text, economics text and literature text, every kind of the text 500 sentences about 46800 words. We suppose some parameter pairs, input queries and search results from the experiment data. We use greedy method; repeat procedure of the input, search and evaluate by supposed parameters, the main results of preliminary experiment are shown as table 1.

Table 1. Preliminary Experiments of Optimum Coefficients

α	1	0	1	5	10	1	1	1	2
β	1	1	0	1	1	5	10	2	1
Precision	73.95	80.76	86.0	88.78	84.1	82.14	87.63	87.63	85.62
recall	60.05	79.24	84.7	96.22	92.9	84.86	79.37	77.37	82.38

According the experimental results, α=5 and β=1 are decided.

3. Experiments and Evaluation

To evaluation availability of the method, we carry out retrieval experiments.

3.1. Data Collection

We collect three fields of text as data of experiments from Web. The engineering text contains 196,085 words, the economics text contains 185,915 words and the literature text contains 212,816. Total of the data are 594,816 words. The engineering text contains the text of electronics, communication engineering, machine engineering and nuclear industry. The

economics text contains the text of economic system, economic policy and economic theory. The literature text contains history novel, modern novel and war novel etc.

We also gathered experiment data of 5800 Japanese paragraphs from Internet, and used morphology tool of ChaSen to segment the text to words. The segmented text is used on learning processing. Then system extends the queries gather by thesaurus, map those queries to Chinese and English queries by the dictionary. The extended queries gather is used on retrieval target text.

We can input Chinese queries and Japanese queries. But in this paper, we not use English text; only carry out two kinds of languages: Chinese text and Japanese text. The retrieval results are ranked at priority of value of LEF. That Japanese text is consisted of four fields: language, law, program and basis.

To get search reference information, user feedback the results of the system search with correct judgment or erroneous judgment. The system extracts common words from correct results and erroneous results.

3.2. Experiment Procedure

At the beginning of experiment, the reference information is empty. We construct a file of batch that contains queries as input. At first the system use the thesaurus to extend the input queries, then translates query words by the dictionaries and searches the text that contains queries. For the results, user selects desirable results. At the feedback process, the selected and unselected results are returned to the system, the learning processing deal with the results. The results are segmented into words; the relationship is established with high frequency words and saved as reference information. Next search will make use of the renewed reference information to deal with word ambiguity. The reference information includes queries and registered common words that extract from selected search results and with high appearance frequency. With the learning processing, the reference information are renewed, the frequency of correct words and the frequency of erroneous words in search results are renewed. Otherwise, we define a time parameter that denotes the order of search time. The time parameter of queries and words show the use time that the queries and words are used in information retrieval. If the LEF of two results is equal then the rank of priority is decided by the time parameter.

3.3. Experiment Evaluation Method

For the evaluation of experiment results, we use the following formulas; precision and recall are calculated by (2) (3). The precision is the percentage which the correct search results occupy the total search results. The recall is the percentage which the correct search results occupy the total correct results in search object text.

$$precision = \frac{NCRR}{TNRR} \times 100\% \tag{2}$$

$$recall = \frac{NCRR}{TNCROT} \times 100\% \tag{3}$$

Here CNRR is the number of correct retrieval results, TNRR is the number of total retrieval results and TNCROT is the total number of correct result in the object text.

The correct retrieval results are the results that the paragraphs contain queries and user desires. CNRR is the number of correct retrieval paragraph that contain queries.

3.4. Experiment Results

We random select some results to calculate the precision and recall, the calculation results as follow Table 2, 3, 4 show.

Table 2. Retrieval Result of before Feedback

query	Retrieval Results	Correct Results	Error Results	Precision [%]
Language text	101	74	27	73.3
Law text	156	75	81	48.1
Program text	140	57	83	40.7
Basis text	92	16	76	17.4
Average				45.4

Table 2 shows the results of first times before feedback processing.

Table 3. Retrieval Result of After Feedback

Query	Retrieval Results	Correct/Error Results	Precision /Recall[%]
Language	125	101/24	81.0/98.0
Law	171	124/47	72.3/93.4
Program	158	124/34	78.5/95.5
Basis	35	19/16	54.3/100.0
Average			75.3/95.7

Table 3 and Table 4 show the results of first times and second times search after feedback respectively. The results show the Precision and the Recall are improved.

Table 4. Retrieval Result of After Second Feedback

Query	Retrieval Results	Correct/Error Results	Precision /Recall[%]
Language	128	110/18	85.9/100.0
Law	187	156/7	83.4/96.3
Program	165	132/11	80.0/89.4
Basis	69	54/2	78.3/99.5
Average			82.3/96.3

4. Discussion

In this paper, the proposed method deals with only words translation ambiguity. The unknown word translation problem is not in considering scope. About unknown words are those words that can not be found in dictionaries.

The experiments are carried out repeatedly, the system would adapt users by feedback processing and learning processing. After the experiments, reference information has registered; the search results will be improved.

According to the experiment results that show in Table 2, 3, 4, the precision and the recall are improved after feedback processing. By the feedback, words ambiguity is disposed to a certain degree and search results of adapting user are obtained. The common words are useful to filter search results.

Through feedback processing and learning processing, the reference information is enriched; retrieval results are more adapting users besides cross information retrieval of unknown words.

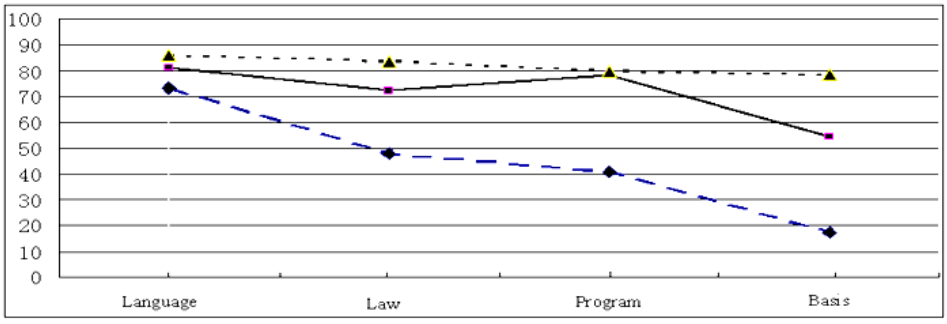

Figure 3. Change of Precision

Figure 3 shows the change of precision. Three curves delegate the precision of the before feedback, after feedback and after second feedback respectively. The three curves indicate the performance of proposed method. The relevant data are shown as Table 2, 3, 4.

When experiment is carried out after feedback, the reference information is established, the search precisions are improved clearly.

5. Conclusion

In this paper, we proposed a method that uses a thesaurus to extend query keyword gather, and uses dictionary to get translation words. We filter search results by using common words extracted from the correct and erroneous results. The search reference information is generated by feedback processing through registering words information of correct search results and erroneous search results. The problem of words to words translation is translation ambiguity. The learning processing and feedback processing is a stratagem of deal with translation ambiguity.

At learning processing, the system uses ChaSen of morphological tool to segment correct and erroneous search results, and calculates the frequency of those words, extracts information of adapting user. For the search result, user judges the correct or erroneous, and then the system returns them to register in the reference information at feedback processing. For the correct search results, the queries that used in correct search will be increased authority. Otherwise, the queries that used in erroneous search will be decreased authority.

At last, the evaluation experiments are accomplished by inputting Japanese text and outputting Chinese text. The experiments show the validity of the proposed method.

We deal with the ambiguity of word translation by using the feedback process; and extend the search queries by thesaurus.

As the result shows the method of adapting user by feedback and learning process is effective. The feedback process updates the information of registered in the reference

information such as the relationship of translation words with other queries, so that the ability of adapting user is improved uninterruptedly.

For the future works, we plan to use this proposed method for open text on Internet and improve the method let's more effective. In addition, more detailed evaluation experiment is necessary.

Acknowledgements

This work was financially supported by Open Fund of Smart Education and Information Engineering (Harbin Normal University), the Scientific Research Foundation of Education Bureau of Heilongjiang Province: (12511127) and Natural Science Foundation of Heilongjiang province (F201243).

References

[1] H. Kataoka, N. Watanabe, K. Mizutani and H. Yoshiura, "DCNL: Disclosure Control of Natural Language Information to Enable Secure and Enjoyable E-Communications", International Journal of u- and e- Service, Science and Technology, vol. 3, no. 1, **(2010)**.

[2] Z. Yu, B. Deng, B. Hou, L. Han and J. Guo, "Word Sense Disambiguation Based on Bayes Model and Information Gain", International Journal of Advanced Science and Technology, vol. 3, **(2009)**.

[3] P. Srinivasanm, "Thesaurus Construction, Information Retrieval: Data Structures & algorithms", Frakes, W.B. and Baeza-Yates, R. (Eds.), Prentice Hall, NJ, **(1992)**, pp. 161-218.

[4] W. Song, J. Yang, C. Li and S. Park, "Intelligent information retrieval system using automatic thesaurus construction", International Journal of General Systems, vol. 40, no. 4, **(2011)**.

[5] A. Yamada and H. Esashi, "Proposal of a New Algorithm for Extracting Information Needs on Small Search System [in Japanese]", Bull, Tokyo Gakugei Univ. Sect., vol. 6, no. 55, **(2003)**.

[6] N. Yoshiki, "Dynamic Co-occurrence Analysis for Interactive Document Retrieval [in Japanese]", Information Processing Society of Japan (IPSJ), 96-NL-115, **(1996)**, pp. 99-106.

[7] K. Kishida and E. Ishita, "Translation disambiguation for cross-language information retrieval using context-based translation probability", Journal of Information Science, vol. 35, **(2009)**, pp. 481-495.

[8] M. Christof and J. D. Bonnie, "Iterative Translation Disambiguation for Cross-Language Information Retrieval", Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, **(2005)**, Salvador, Brazil, New York, NY, USA：ACM.

[9] A. Mirzal, "Design and Implementation of a Simple Web Search Engine", International Journal of Multimedia and Ubiquitous Engineering, vol. 7, no. 1, **(2012)**.

[10] A Japanese morphological analysis system: http://cl.naist.jp/en/index.

[11] ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System): http://www.ictclas.org/index.html.

[12] A. Pirkola, T. Hedlund, H. Keskustalo and Kalervo, "Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings", The fourth, fifth and sixth text retrieval conferences, **(1998-1999)**. URL: http://trec.nist.gov/.

[13] L. Ballesteros and B. Croft, "Dictionary-based methods for cross-lingual information retrieval", Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, **(1996)** September 9-13, Zurich, Switzerland.

[14] L. Ballesteros and B. Croft, "Resolving ambiguity for cross-language retrieval", Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, **(1998)** August 24-28; Melbourne, Australia.

[15] C. Buckley, A. Singhal, M. Mitra and G. Salton, **(1996)**, New retrieval approaches using SMART: TREC-4. The Fourth Text REtrieval Conference (TREC-4), Gaithesburg, MD. Available at: http://trec.nist.gov/pubs/trec4/t4_proceedings.html.

Author

Zhongjian Wang, Ph.D., professor, His main research interests include natural language process, Chinese sentence paraphrase, Chinese word segmentation and Information retrieval *etc*.