ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 130
2nd International Forum on Management, Education and Information Technology Application (IFMEITA 2017)

Cross Language Information Retrieval System

Zhidan Yang[1, a], Zhiting Yang[2,b]

[1]TSL School of Business and Information Technology, Quanzhou Normal University, Quanzhou 362000, Fujian, China

[2]Yuanyuan Zhu

Haoyouxing Information Technology LTD, Quanzhou 362000, Fujian, China

[a]yangmantis@aliyun.com, [b]doubleround@gmail.com

Keywords: CLIR; corpus; LDA; LCS; GIZA++; bilingual; in-memory DB; Crawler4j

Abstract. In the paper, we describe a Cross Language Information Retrieval System (CLIR), which allows user input English queries and search Chinese documents. We also explore the solution for online search engine, which can meet commercial requirements.

Introduction

CLIR is critical to communication among people from different countries, especially people speak in English and want to search Chinese information or vice versa. Trading, foreign education and traveling are more and more popular in China. In 2016, 5.4 million Chinese students studied in other countries [1], 122 million Chinese travelled aboard and 138 million foreigners travelled in China [2], and the import and export trading reached 24.33 trillion [3]. However, the current search engine or websites are not able to accommodate huge cross language information requirements. In this paper, we describe a new Cross Language Information Retrieval System to meet such requirements. There are two phases for development of CLIR system. Firstly, we implement a prototype with basic functions of bilingual information retrieval. In phase two, we make trials to transfer the prototype to an online search engine.

Prototype

In 2012, we develop CLIR system for academic experiments, which contains about 20,000 bilingual documents, 100 bilingual dictionaries, 11 English Queries, 237 Chinese relevant documents related to these queries. The bilingual documents are downloaded from Wikipedia by Crawler4j [4] and divided groups based on 13 majors of University of Albany (SUNY Albany). Each pair of bilingual documents is mapped to an English subject word and its Chinese meaning from subject dictionaries. These documents are used to train search algorithm and extract parallel sentences, which have same meaning between English and Chinese. Beside subject dictionaries, there are general dictionaries created by discrete web-based dictionaries under supervision. The translation of queries and documents are based on these dictionaries. The queries are English definitions of 11 majors, while the relevant documents are the related Chinese definitions downloaded from 8 Chinese wiki websites.

For searching algorithm, we use Longest Common Substring (LCS) algorithm [5] and TF-IDF [6] in prototype to calculate similarity and relevance judgments between queries and documents. To analyze the performance of different algorithm, the system will read queries and search the documents. Then the system evaluates the algorithm through the rank of related documents in search result. In addition to the above 4 kinds of downloaded files, there are created documents by system, which are the Chinese documents with inserted English meaning. During construction of corpus and dictionaries, LCS algorithm is additionally used to trim redundant head and tail of files downloaded by Web Crawler. The templates of head and tail are manually created to compare the content in bilingual corpus for deletion. The benefit of LCS is that it can skip minor difference between documents and templates. For example, the date of different documents is changing. Word by word

comparing cannot skip the minor different words as LCS. The following diagram demonstrates the system architecture of prototype.

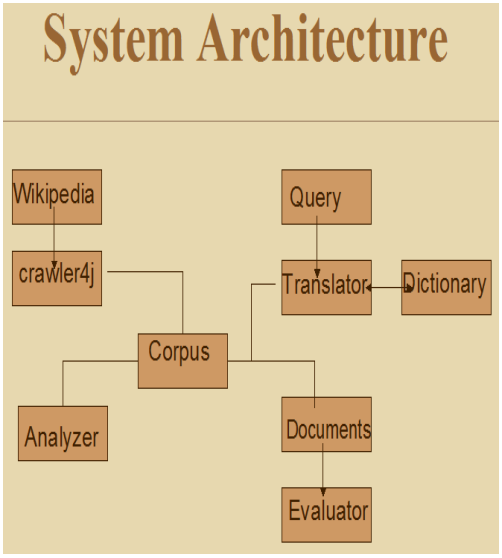

Figure 1

Online website

Since 2012, we start to develop online search engine available to Chinese users and others want to know about China. To convert the prototype to commercial website, we should tackle several issues. Firstly, documents downloaded from different website contains unpredictable information, which we cannot pre-classify them. What is more, the system may process billions or trillions level of documents every day from thousands of websites. Finally, users normally expect runtime response from search engine.

In phase two, webpages are downloaded from Wikipedia and Commercial, Education, and Government Website. The webpages from Wikipedia is mainly for training by our algorithm, while others are prepared for online search engine. Different from previous index method, we use Latent Dirichlet Allocation (LDA) algorithm [7] to extract topics from documents from single website and the documents are classified by topics, which are represented by words. Given a query as input, similarity and relevancy is calculated between topics words and terms of query. To process large scale documents, we use Hadoop Cluster Server [8] and the documents are distributed to different servers for parallel operation. To achieve runtime response of search engine, we pre-create queries from topic words and save them as decision tree in Derby In-Memory DB [9]. In this way, search engine can refresh search result when user type words one by one, and it also popup frequent appear query group.

Latent Dirichlet Allocation

TREC-9, a CLIR system between English and Chinese, uses hidden Markov models (HMM) as translation models. In their system, documents are ranked by the probability that a Chinese document D is relevant given an English query Q [10]. It also adopts query processing and query expansion to improve the performance. However, a small perturbation in the ranked document can lead to a big difference in performance. Microsoft uses NP translation Model, which is similar to template-based translation model, to improve it [15]. In our system, we use LDA to classify different topics in documents. Then we can use the probability of word occurrence frequency to locate which topic the query falls in.

LDA algorithm can extract topics from files, which describe their content. When we search files, we care about the topics they describe instead of the words they use. The relation between files and topics are many to many. The below diagram explains the relationship between topics and files.

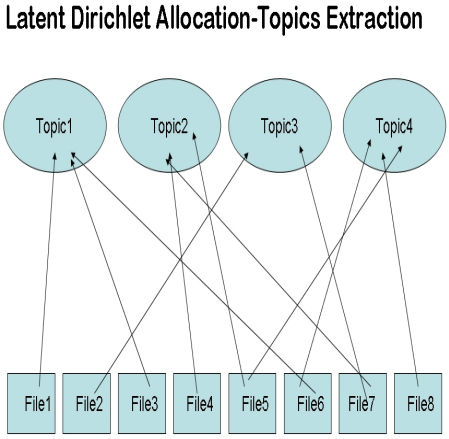

Figure 2

Blei et al used the LDA algorithm to extract topics for single language information retrieval. In cross language information retrieval, we apply LDA to extract topics from corpus of each single language itself and map the topics between different languages. Given a query as input, system find the topic it belongs to firstly, using TF-IDF to calculate the similarity between query and different topics. In the second step, the system locates the mapped topic in another language. In this way, the system can search the documents with the same topic as that of the query.

Machine Translation

Traditional machine translation translates both queries and documents using IBM Model 1 to improve the quality [11]. But IBM Model 1 only considers word for word probability. The fluency and adequacy is not good enough. In our system, a model similarly to IBM Model 3 is used as translation model. IBM Model 3 offers two additional features compared to IBM Model 1, fertility and distortion [12]. In our model, the probability of different meanings of one word is used. We design this probability according to the frequency of occurrence of each word in the query we input in corpus.

And the corpus will be classified to different topics by LDA. In this case, we can use this probability to locate which topic the query falls in. Then we locate the topic of the words of the query. There may be several meanings for one word, but once we detect which topic the word in the preprocessed corpus falls in, we narrow down the result set for the translation of this word. With these improvements, the accuracy is better.

Search Engine

Yahoo launch their CLIR in 2005, while google do a little later in 2007. In China, Baidu, 360 and Sogou also provide CLIR features. A research show the retrieval effectiveness of EC and CE cross lingual search in google and yahoo is much lower than that of EE and CC monolingual search [14]. The machine translation limitation is main problem in CLIR. As people are interesting in updating information, like news and government announcement, we use CLS to extract parallel sentences from unparalleled corpus of bilingual news. These sentences are used as corpus for GIZA++ [13]. The unparalleled corpus with parallel sentence are further used by LDA to extract topics, which map bilingually and improve precise in searching.

To enable the scalability of CLIR system, we use Hadoop as cluster server to preprocess downloaded webpages and document translation. We also use decision tree to save the link of keywords. So the webpages using Ajax technical can popup the next frequency words, when user input the first keyword for searching. The correlation analysis between keywords is based on the frequency of word-word sequence. The decision tree is saved in read-only in-memory DB, in this way the response time for the user request can be minimized.

Experiments

We tried two versions of LDA implemented by three languages. Initially we chose JGibbLDA as training programming for documents, which are preprocessed to delete stop words and restore the words to their stem formation. In experiments, the number of topics and number of words in each topic are adjusted manually, according to the knowledge of corpus. In the following round experiment, we extract topics documents through Blei LDA. According to our experiments, Blei LDA can process hundreds level of documents in seconds level time, while JGibblDA takes ten minutes level time to process same documents.

When we train the first batch of data, we changed the topics number from 2 to 200 topics. We find out when 3 topics is chosen as input argument, words in topics has best performance regarding words relations to present topics. We also find out the stemming brings a litter bit errors in the topics words. When the form of the word is changed by chopping suffix, the program may not find the exactly original words for the stem to restore.

We setup 3 Hadoop servers, with one for master and all for slaver servers. The status, including jobs and tasks can be checked through webpages. One problem is that the master server is not able to restart slaver server if it is down. The problem of In-memory DB is the loss of data, when database crashes.

Future works

In the further work, we will use the topics concept in data immigration and integration to decrease disk access times. However, the complexity of LDA prevent the usage in large number of documents. A fast version or improvement of LDA will be further focus.

Another important work is to finish Machine Translation module and integrate it into the current system. The IBM Models 1-6 for alignment are implemented in the open-source toolkit GIZA and its successor GIZA++ [Al-Onaizan et al. 1999; Och and Ney 2003]. These models require parallel corpus. How to construct a good parallel corpus is another problem to be solved. We will continue to use the longest common sequences (LCS) to generate a parallel corpus from the document we collected from the websites.

Acknowledgements

The authors wish to express their appreciation to Xintong Wang, Mei Yuan and Syed Shahzad Raza. They assisted on accounting on research on Baidu etc, experiments on LCS to extract parallel sentences from unparalleled corpus of news and theoretical analysis.

References

[1] Statistic information of students study aboard in 2016 , Ministry of Education of the People's Republic of China, March 1st, 2017, http://www.moe.edu.cn/jyb_xwfb/xw_fbh/moe_2069/xwfbh_2017n/xwfb_170301/170301_sjtj/201703/t20170301_297676.html

[2] Chinese travel industry statistics report in 2016, China Tourism Academy, Nov 8th, 2017, http://www.ctaweb.org/html/2017-11/2017-11-8-14-49-91372.html

[3] The international trading situation in 2016, General Administration of Customs of the People's Republic of China, Jan 13rd, 2017, http://www.customs.gov.cn/publish/portal0/tab65598/info836849.htm
[4] crawler4j, available at http://code.google.com/p/crawler4j/
[5] Allison, L., & Dix, T.I. (1986). A bit-string longest common subsequence algorithm. Information Processing Letters, 23(6), 305–310.
[6] Salton and M. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
[7] David M. Blei, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993-1022
[8] Apache Hadoop. Available at http://hadoop.apache.org
[9] Apache Derby. Available at http://db.apache.org/derby/
[10] Jinxi Xu, Ralph Weischedel. TREC-9 Cross-lingual Retrieval at BBN (2000). BBN Technologies.
[11] J. Scott McCarley. Should we translate the documents or the queries in Cross-language Information Retrieval (1999). IBM T.J. Watson Research Center.
[12] IBM Models and concepts, available at http://ppt.books5.net/s/statistical-machine-translation---university-of-maryland-institute-w1222-ppt.ppt
[13] GIZA++ statistical translation models toolkit, available at http://code.google.com/p/giza-pp/
[14] Retrieval Effectiveness of Cross Language Information Retrieval Search Engines, 2011, Schubert Foo, Nanyang Technological University, https://www.ntu.edu.sg/home/sfoo/publications/2011/2011ICADL-SF-fmt.pdf
[15] Statistical Query Translation Models for CrossLanguage Information Retrieval, Microsoft, 2005, https://www.microsoft.com/en-us/research/wp-content/uploads/2017/01/gao_nie_zhou.pdf