Multi-filtering Method Based Cross-lingual Link Discovery

Yingfan Gao, Hongjiao Xu, Junsheng Zhang, Huilin Wang
Institute of scientific and Technical Information of China
No 15, Fuxing Road, Haidian District,Beijing, China, 100038
{gaoyingf,xuhongjiao,zhangjs,wanghl}@istic.ac.cn

ABSTRACT
This paper describes cross-lingual link discovery method of ISTIC used in the system evaluation task at NTCIR-9. In this year's evaluation, we participated in cross-lingual link discovery task from English to Chinese. In this paper, we mainly describe our understanding for CLLD, the key techniques of our system, and the evaluation results.

Keywords
Multi-filtering; Stop words; Information Extraction

Team Name: ISTIC

Subtasks/Languages: English to Chinese CLLD

External Resources Used: Google Translate

1. INTRODUCTION
What is Cross-lingual link discovery (CLLD)[1]? NTCIR-9 website said that CLLD is a way of automatically finding potential linking between isolated documents in different languages. As it were, CLLD was born from CLIR (cross-lingual information retrieval). CLIR strives to find virtual link between the provided cross-lingual query and the retrieval documents. To CLLD, there are no the "provided cross-lingual query". So we should try our best to find the possible "query" in the documents of source language first and these "queries" are called "anchors" in CLLD. We consider, the most important task of CLLD is to find "key words or phrases" in source document.

On the basis of finding the precise anchors, there would be the good "queries" for CLLD. Next, we have to pay more attention to extract the significant words from documents. There are some translation tools available such as Google Translate [2], so query translation for CLIR is not the main tasks for us. Establishing the searching engine is also a subordinate task. We do all works on our existing IR platform.

This paper is organized as follows. In section 2, we will present the overview of our CLLD system. In section 3, we will introduce the multi-filtering methods used to find anchors in documents. In section 4, query translation method and information retrieval method will be narrated and listed. In section 5, experimental results will be shown and the analysis will be given. In section 6, we will see the conclusion for this paper.

2. SYSTEM OVERVIEW
Figure 1 presents our system architecture.

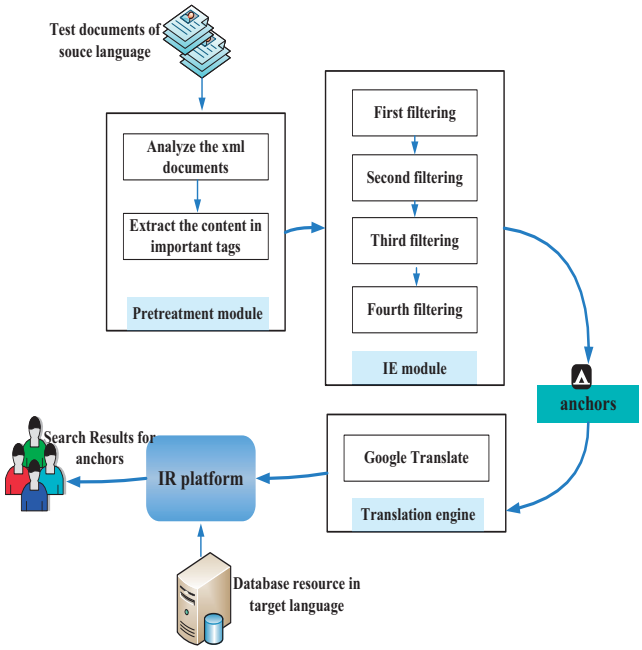

Figure 1. Our system architecture.

There are main four parts in our system for CLLD. They are pretreatment module, IE (Information Extraction) module, translation engine and IR (Information Retrieval) platform. Test documents of source language (English in our tasks) are files in XML format. Pretreatment module analyzed and processed these XML files. By doing this, the content in important tags (such as "title" tag, "id" tag, "bdy" tag et al.) of XML files would be extracted and XML files would be turned to TXT files.

In second part, four filtering methods would be used to extract the important words or phrases from files after pretreatment above. These extracted words or phrased would be call "anchors" in this paper.

In third and fourth part, anchors got from the second part would be translated into target language (Chinese in our tasks) and would be as the queries for CLIR. We use Google Translate as our translation engine. By IR platform,

Chinese documents would be searched and passed to the users in the end.

3. MULTI-FILTERING METHODs

3.1 Related Works

The process of finding anchors from documents is actually the process of extracting information from text. Extracting information from text is a challenging task for natural language processing researchers as well as a key problem for many real-world applications. In the last few years, the NLP community has made substantial progress in developing systems that can achieve good performance on information extraction tasks for limited domains [3]. The purpose of information extraction (IE) systems is to extract domain-specific information from natural language text. IE systems typically rely on two domain-specific resources: a dictionary of extraction patterns and a semantic lexicon. The extraction patterns may be constructed by hand or may be generated automatically using one of several techniques [4].

In the past, IE has been used on small, homogeneous corpora such as newswire stories or seminar announcements. As a result, traditional IE systems are able to rely on "heavy" linguistic technologies tuned to the domain of interest, such as dependency parsers and Named-Entity Recognizers (NERs). These systems were not designed to scale relative to the size of the corpus or the number of relations extracted, as both parameters were fixed and small. The problem of extracting information from the Web violates all of these assumptions. Corpora are massive and heterogeneous, the relations of interest are unanticipated, and their number can be large. Michele Banko [5] et al. proposed an Open Information Extraction (OIE) method. OIE is a new extraction paradigm that facilitates domain independent discovery of relations extracted from text and readily scales to the diversity and size of the Web corpus. In this paper, we present a novel and simple multi-filtering based IE method. No dictionaries used and no limited domains. By analyzing the document only, we can get the relatively accurate anchors soon.

3.2 Multi-filtering based IE method

(1)First filtering

There are different stop words list for different language. Stop words are terms that appear so frequently in text that they lose their usefulness as search terms, such as adverbs, articles, auxiliary, et al. In common preprocessing for documents, the stop words would be filtered first. But in our experiments, the stop words and the punctuation would be the separating characters to separate the long paragraph into short text segments and the process could be called the first filtering for documents. After this, one document would be separated into many little text segments. These text segments need more filtering for finding anchors.

(2)Second filtering

In this stage, the word frequency statistics could be done to get the low frequency words (some threshold should be set here). That is to say, the low frequency words which is lower than threshold value could act as the new stop words only for this document. We consider different words will act as different roles in different documents. These selected low frequency words would be the new separating characters to separate the short text segments above, and we could get some new words and multi-word phrases.

(3)Third filtering

To filter more useless words and phrases, we should employ the third filtering method to reduce the noise. In this phase, pos tagging would be done first. Then, some kinds of words (such as adjective, adverb and verb, et.al) and some POS collocation modes (such as noun+verb, adj.+adj., et.al) would be used to filter the words and phrased getting from the steps above. After the process here, we could get the candidate words and phrases with lesser noise.

(4) Fourth filtering

In this stage, some weighting rules would be used to select the final words and phrases as anchors. Specially, the candidate phrases could be sorted by their weights. We use an ingenious method to do this job. First, set a threshold value for word frequency. The candidate words got from part (3) above would be filtered and only the words which are higher than threshold value would be reserved. Then in each sentence, the cooccurrence times among these words would be computed and the string collection between words would be got and analyzed. If the string collections between words are just like the candidate phrases above, these phrases would be set to higher weight.

After these four filtering phases, the words and phrases we got would be as the anchors for our CLLD system.

4. TRANSLATION AND SEARCHING PLATFORM

To translate the anchors we found in section 3, we integrate the Google Translate API into our CLLD system. Moreover, our searching platform is established by Lucene open source software libraries. We use the basic TF-IDF retrieval model here. We are improving it by Okapi BM25 function. We hope the improved searching platform could get the higher precision.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1 Experiments for IE

Using the test topics and document collections provided by NTCIR-9 CLLD tasks, we do the English-Chinese Information Extraction experiments.

On the basis of the method described in section 3, we extracted the anchor words from documents. To any document, the extracted words would be highlighted and shown in figure2.

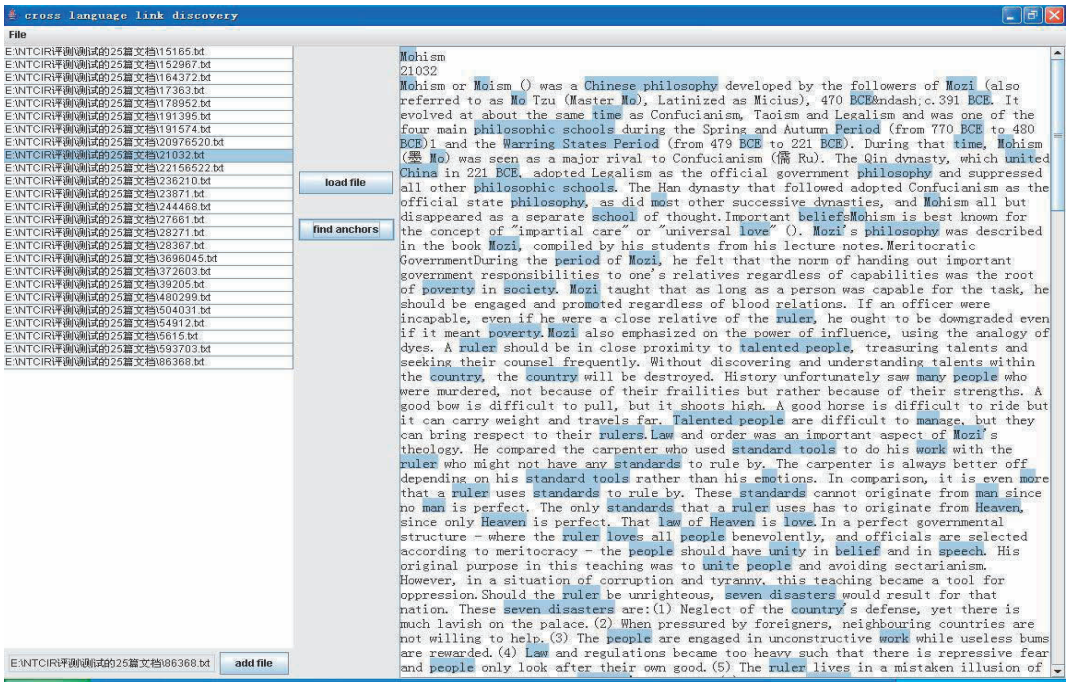

Figure 2. Our IE tools for CLLD of NTCIR-9.

5.2 Experiments for CLLD

Depending on the anchors extracted in section 5.1, we use the Google Translate tools to translate these anchors first. Then, these translated words would be as the searching words for IR platform. We wrote our experimental results into submitted file according to the format in NTCIR-9 CLLD website.

Using the Crosslink evaluation tools (Crosslink Evaluation-20110907.zip) provided by NTCIR-9 CLLD tasks, we evaluate our system and the result could be shown in figure 3.

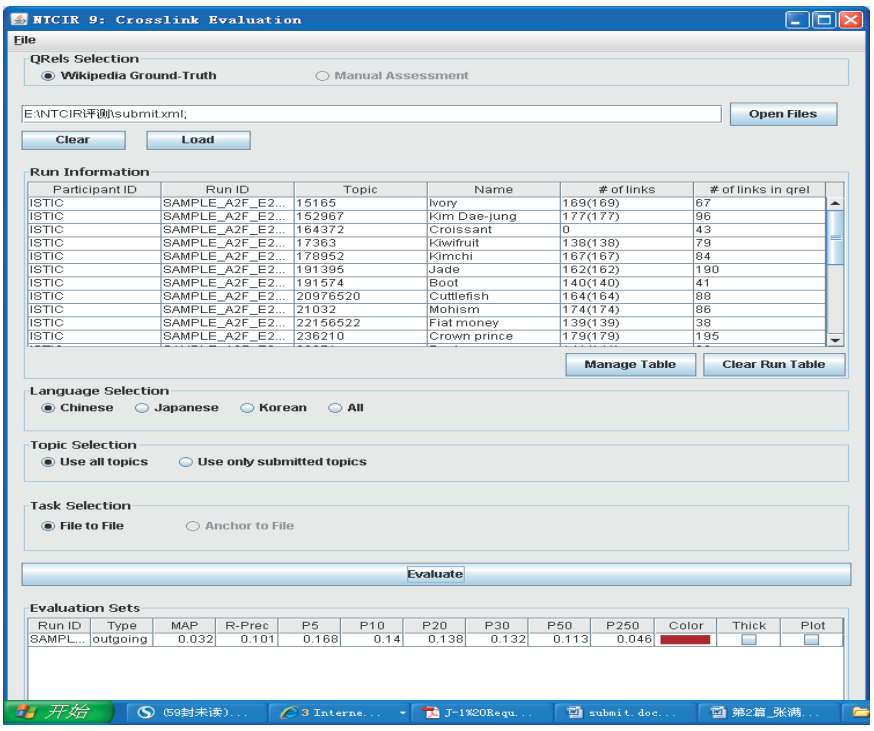

Figure 3. System evaluation results.

5.3 Analysis for experimental results

With the system evaluation results in figure 3, we can see the very low MAP and R-Prec in "File to File" task. Moreover, "Anchor to File" task is unelectable. Why does this happen? We consider the possible reasons are as follows:

(1) No efficient IR ranking algorithm

After getting anchors, we use these anchor words as the entry words of IR system. But our IR platform is based on the TF-IDF model. So when taking the preceding searching results, we could not ensure that these selected pages are good enough. In "File to File" task, the lower MAP and R-Prec proved this. We hope the improved ranking algorithm could get the higher precision.

(2) Failure to comply with the rules

Tasks of NTCIR-9 CLLD said that topic files including their CJK counterparts must be removed from document collections. We didn't pay more attention to this and didn't remove these files. Evaluation system has given us punishment in figure 3. Moreover, during preprocessing, we reserved the tags such as "title", "id", "bdy", et al. But we didn't obey the rules of NTCIR-9 CLLD that special case links (numbers, years, dates and century links) should be excluded in the runs. So some unimportant anchors have been obtained. These mistakes have important influence on our CLLD system performance.

(3) Wrong submitted file

In figure 3, "Anchor to File" task is unelectable. By analyzing the submitted file, we found the wrong calculation has been done for anchors' position. Incorrect offsets have fatal influence on performance for CLLD system. Indices like MAP,R-prec, et.al for A2F evaluation are all zero.

6. CONCLUSION

The multi-filtering method was used to solve the CLLD problems in our experiments. First, the punctuation and the common stop words could be the separating characters to separate the long paragraph into short text segments. Then, the word frequency statistics could be done to get the low frequency words (some threshold should be set here). These selected low frequency words would be the new separating characters to separate the short text segments above, and we could get some new words and multi-word phrases. In the third stage, pos tagging would be done first. Then, some kinds of words (such as adjective, adverb and verb, et.al) and some POS collocation modes (such as noun+verb, adj.+adj., et.al) would be used to filter the words and phrases getting from the steps above. In the fourth stage, according to the words which are above the thresholds, compute the cooccurrence times among these words in each sentence of the documents and acquire the new multi-words by the sequence of words above in sentence. After these four filtering phases, the words and phrases we got would be as the anchors. The translation tool we used is Google Translate and the searching platform is established by Lucene software package. We have established CLLD platform of our own and this is exciting

No dictionaries and no limited domains. The advantages of our system lie in high speed and flexible. The multi-filtering methods for finding anchors that we used in this system are novel and challenging. But there are many disadvantages and errors in our tasks. Just like the analysis for experimental results above, no efficient IR ranking algorithm, failure to comply with the rules, and wrong submitted file.

We are so regret about all these mistakes. This is the first time for us to take part in the evaluation of NTCIR. We didn't read the task instructions of CLLD carefully and we should draw a lesson from failure.

7. ACKNOWLEDGE

This research has been partially supported by ISTIC research foundation projects XK2011-6,ZD2011-3-3 ,YY-201121.

8. REFERENCE

[1] http://ntcir.nii.ac.jp/CrossLink

[2]http://translate.google.com/translate_t

[3] Ellen Riloff. Automatically Constructing a Dictionary for Information Extraction Tasks. Proceedings of the Eleventh National Conference on Artificial Intelligence, AAAI Press / MIT Press, 1993: 811–816.

[4] Ellen Riloff, Rosie Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99).

[5] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni. Open Information Extraction from the Web. In Procs. of IJCAI,2007:2670-2676.