WUST EN-CS Crosslink System at NTCIR-9 CLLD Task

Maofu Liu, Le Kang, Shuang Yang，Hong Zhang
College of Computer Science and Technology, Wuhan University of Science and Technology
Wuhan 430065, Hubei, P.R.China
{e_mfliu, kangle2010}@163.com

ABSTRACT
This paper describes our work in NTCIR-9 on the task of Cross-Lingual Link Discovery (Crosslink/CLLD). The work mainly focuses on two aspects to accomplish this task: (1) How to collect useful data for Crosslink and (2) How to use the data correctly and effectively. The system firstly uses online data collecting and text mining in Chinese Wikipedia articles to build the basic Crosslink database. And then these data and two-way expansion algorithm will be applied to identify the anchors and find out the relevant corresponding matchers.

Categories and Subject Descriptors
I.2.7 [Artificial Intelligence]: Natural Language Processing –text analysis.

I.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – linguistic processing.

General Terms
Experimentation.

Keywords
Cross-Lingual Link Discovery, Text Mining, Two-Way Expansion Algorithm

1. INTRODUCTION
Cross-Lingual Link Discovery (CLLD) is a way of automatically finding potential links between documents in different languages. It is not directly related to traditional cross-lingual information retrieval (CLIR). While CLIR can be viewed as a process of creating a virtual link between the provided cross-lingual query and the retrieved documents, CLLD actively recommends a set of meaningful anchors in the source document and uses them as queries (with the contextual information) from the article to establish links with documents in other languages.

CLLD usually includes two main important phrases: (1) identify all potential anchors from the given text, and then (2) match the potential anchors to the target documents in the given documents set.

The potential anchors extracted from the given text affect the performance of Cross-lingual link discovery. In fact, the potential anchors in CLLD can be different with distinct identification strategies, because even if the system can identify the longest named entity which the user exactly needs from the text, there might be some verbs or shorter named entities which are not selected from the text. The other major problem in CLLD lies in the second phrase when the system tries to link these potential anchors to the most appropriate target documents. The user can not be sure that a given anchor represents one or more documents. On the other hand, the title of one document which is the same as an anchor may have the different meaning to the anchor at the semantic level.

The easiest way to filter anchors out is to use English Wikipedia all-titles file. And our work also includes other resources such as the Crosslink database and named entity recognition tool. A well-formed database for CLLD is essential because it not only helps to recognize anchors, but also plays a key role in the linking processing. The system uses Chinese Wikipedia collection to extract useful information with certain pattern, trying to make the best use of the source to get potential linkable pairs.

The remainder of this paper is organized as follows. Section 2 delineates the architecture of our Crosslink system. Section 3 discusses our evaluation results. Finally, we conclude our paper in section 4.

2. System Description
Our Crosslink system includes three main modules, i.e. Crosslink database construction, anchors filtering, and anchors matching. Figure 1 illustrates the architecture of our Cross-Lingual Link Discovery System in detail.

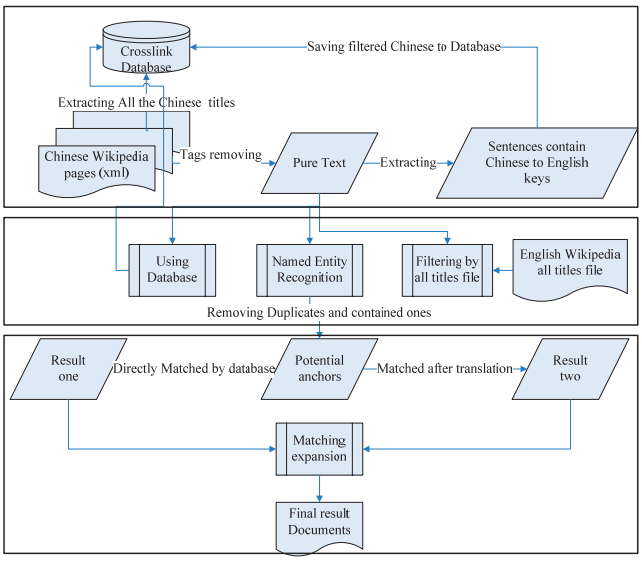


Figure 1. The Architecture of our Crosslink System

2.1 Crosslink Database Construction
Firstly, the system extracts all the Chinese titles and inserts them into the initial Crosslink database. The database is used to record the matched Chinese and English title terms and the relevant degree between the terms. Our work makes an assumption that the Chinese title could represent the corresponding documents perfectly well, while one title may be relevant to more than one English anchor.

The system tries to analyze the Chinese Wikipedia pages by going through its contents and find the relevant cross-lingual terms. Once a sentence meets a certain pattern, the system will use the Chinese titles in the database to locate the English term in this sentence to its Chinese title and update the database at the same time.

After carefully looking into the contents of Chinese Wikipedia pages (the documents set in xml format), our group has found out one useful pattern like "Chinese term(corresponding English term)" in the context such as the marked parts in the following Figure 2. We save these valuable "Chinese(English)" pairs into our Crosslink database. Examples of this kind of pairs are also illustrated in Figure 2.

Chinese_title	English_title
汤玛斯・史伯格	Thomas L. Sprague
科伦	Colon
巴尔博亚	Balboa

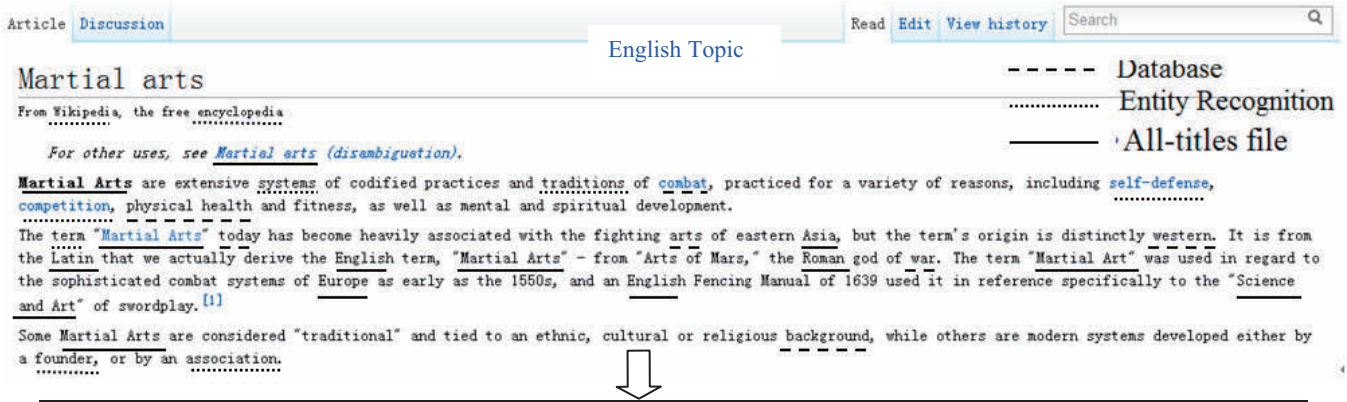
Figure 2. Examples of our Crosslink Database Construction

No matter we update or insert records, the "relevant" field will be set to "1", which indicates that these records are extracted in this way. After all the Chinese Wikipedia pages are done, we have got about 80,000 matched records. So far, our database has been initially established.

2.2 Anchors Filtering

When an English topic is given, we should identify all the potential anchors in the text. The potential anchors provide all the links that might be outgoing.

In our system, the potential anchors are generated from the merge of three potential anchor sets by removing the ones which are either duplicate or contained by another anchor. The first set is constructed by the Crosslink Database records which "English_title" is not "unknown", named DB-Set, and the second set is generated by Stanford Named Entity Recognizer (NER) [1], named NER-Set, and the third set is formed by using the English Wikipedia all-titles file[2], named EWiki-Set. Examples of anchor filtering are illustrated in Figure 3.



DB-Set	combat, physical health, today, arts, western, background
NER-Set	Wikipedia, encyclopedia, systems, traditions, self-defense, competition, term, founder, association
EWiki-Set	Martial arts, Asia, Latin, English, Roman, Europe, Science and Art

Figure 3. Examples of Anchors Filtering

2.3 Anchors Matching

We use the Crosslink database to match the links with English terms, then give the rest unmatched ones to Google translator and match the translated results (Chinese terms) with the Chinese titles in database.

The system directly matches the anchors which are originally in the first potential anchors set to the corresponding record in the database. Due to the possible N:M (N>1,M>1) relationships between the fields "Chinese_title" and "English_title", one English anchor might be matched more than one Chinese title.

As to the anchors remained, which are the unmatched ones in the first potential anchors set and those in the second and third potential anchors sets, we use Google online translator to help us get the Chinese terms of these English anchors. Then we use this translated result to match with the field " Chinese_title" in Crosslink database. If it matches with a certain Chinese title, we insert a new matching record into the database, and set the field "relevant" to "-1" to indicate how the record is formed.

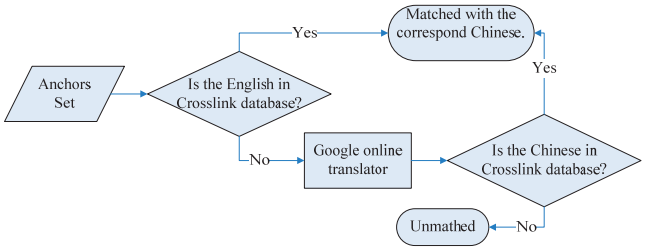


Figure 4. The Process of Anchors Matching

2.4 Matching Expansion

We use the temporary results got from the Anchor Matching phrase to expand a multi-relevance links set using two-way matching and synonyms.

Now we have got some initial matched pairs from the Anchor Matching phrase, and then we try to make every English anchor be relevant to several Chinese titles by finding synonyms for the Chinese titles [3]. Of course, the synonyms must be in the field "Chinese_title" in the Crosslink database to be effective outgoing documents.

The system does two-way expansion algorithm as four steps.

Step 1: Assume we successfully get a matched pair A2F (Anchor to File)

Step 2: Find out all the matched pairs whose left is A from Crosslink database.

Step 3: Find out all the matched pairs whose right is F from Crosslink database

Step 4: Link every A to all the F.

When we have done it, an expanded result is got, which is also the final result that needed to write into a document.

3. Experiments

We submitted four formal runs to NTCIR-9 and the description of our each run is in Table 1. The official evaluation results of performance are listed in the following tables. There are two types of assessments: automatic assessment using the Wikipedia ground truth (existing cross-lingual links) and manual assessment done by human assessors. From Table 2 and Table 3, we can find that WUST_A2F_E2C_03 achieve the best result.

Table 1. Description of our each run

Run ID	Method
WUST_A2F_E2C_01	Database, NER, Translation
WUST_A2F_E2C_02	Database
WUST_A2F_E2C_03	Database, Synonyms
WUST_A2F_E2C_04	NER, Translation

In Table 2 and 3, "Run ID" indicates the name of the run file we submitted. The "A2F" and "E2C" indicates that the subtask we participate in is mainly concerned with anchor to file and English to Chinese. R-Prec and Mean Average Precision (MAP) are the main metrics used to quantify the performance of the Crosslink system.

Table 2. F2F evaluation results with Wikipedia ground truth

Run ID	MAP	R-Prec
HITS_E2C_A2F_02	0.373	0.471
UKP_E2C_A2F_02	0.314	0.417
WUST_A2F_E2C_01	0.093	0.165
WUST_A2F_E2C_02	0.089	0.163
WUST_A2F_E2C_03	**0.108**	**0.207**
WUST_A2F_E2C_04	0.076	0.123

Table 3. F2F evaluation with manual assessment results

Run ID	MAP	R-Prec
HITS_E2C_A2F_02	0.241	0.315
UKP_E2C_A2F_02	0.308	0.429
WUST_A2F_E2C_01	0.070	0.102
WUST_A2F_E2C_02	0.065	0.103
WUST_A2F_E2C_03	**0.082**	**0.124**
WUST_A2F_E2C_04	0.038	0.056

Comparing with other groups, our system only achieves an intermediate official result. After looking into the evaluation result of Precision-at-N in the Table 4 and 5, we find the evaluation result of "P250" for our system is so low that it may be the main reason for our intermediate official result.

Table 4. F2F evaluation results with Wikipedia ground truth

Run ID	P5	P10	P20	P30	P50	P250
HITS_E2C_A2F_02	0.808	0.796	0.742	0.680	0.571	0.183
UKP_E2C_A2F_02	0.640	0.632	0.656	0.613	0.527	0.179
WUST_A2F_E2C_01	0.576	0.496	0.406	0.353	0.264	0.060
WUST_A2F_E2C_02	0.552	0.480	0.394	0.327	0.247	0.061
WUST_A2F_E2C_03	0.576	0.492	0.406	0.360	0.285	0.077
WUST_A2F_E2C_04	0.496	0.424	0.344	0.304	0.231	0.048

Table 5. F2F evaluation with manual assessment results

Run ID	P5	P10	P20	P30	P50	P250
HITS_E2C_A2F_02	0.752	0.772	0.748	0.735	0.701	0.288
UKP_E2C_A2F_02	0.688	0.684	0.716	0.712	0.678	0.417
WUST_A2F_E2C_01	0.744	0.692	0.572	0.516	0.407	0.096
WUST_A2F_E2C_02	0.712	0.668	0.552	0.476	0.381	0.098
WUST_A2F_E2C_03	0.744	0.684	0.576	0.528	0.440	0.120
WUST_A2F_E2C_04	0.568	0.536	0.536	0.521	0.519	0.244

For each topic we only filter out potential anchors starting with a capital letter. But if we ignore the word case, then the number of anchors would increase to be nearly close to 250 so that the evaluation result of "P250" for our system will not be particularly low. The additional experiment results after ignoring the word case are better and they are listed in Table 6 and Figure 5.

Table 6.The automatic evaluation comparative results of our system with word case ignorance

	Original result (Ground truth/Manual)	with case ignorance (Ground truth/Manual)
MAP	0.108/0.082	0.124/0.149
R-Prec	0.207/0.124	0.224/0.231
P5	0.576/0.744	0.528/0.768
P10	0.492/0.684	0.488/0.744
P20	0.406/0.576	0.392/0.67
P30	0.360/0.528	0.359/0.645
P50	0.285/0.440	0.293/0.559
P250	0.077/0.120	0.108/0.219

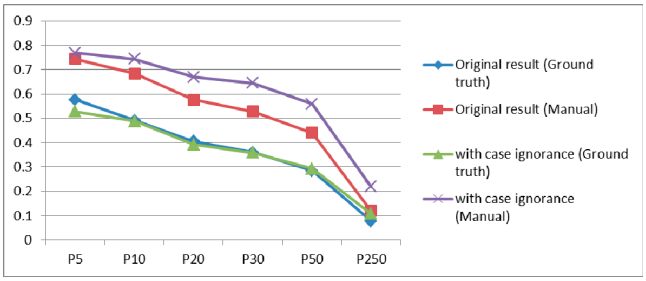


Figure 5. Precision-at-N with ignoring case ignorance

In Table 6, we could easily find out that after the enhancement of the number of anchors by ignoring word case, the value of "P250" increases significantly, especially the increment of MAP and R-Prec. By the comparative experiment, we can draw the conclusion that trying to find as many anchors as possible is a good way to achieve better result, and the value of "P250" plays a key role in the performance of the Crosslink system.

4. Conclusions

Our Crosslink system depends on the database we initially constructed to some extent, but the way in our system to judge whether a Chinese and English pair is relevant or not is so simple that can not be 100% correct. And also when considering even a slight difference in the form of a word may actually confuse our judgment, our system needs to tolerant to variation. We think the key lies in how to find more already relevant Cross-lingual pairs and save them to our database if we want to improve our system performance in a reliable way.

ACKNOWLEDGMENTS

The work in this paper was supported partially by the National Natural Science Foundation of China (No. 61100133 and No. 61003127), the Natural Science Grant of Hubei Province (No. 2009CDB311) and Social Science Grant of Department of Education of Hubei Province (No. 2011jyte126).

REFERENCES

[1] Stanford Named Entity Recognizer. World Wide Web. http://nlp.stanford.edu/software/CRF-NER.shtml

[2] enwiki-latest-all-titles-in-ns0.gz. World Wide Web. http://dumps.wikimedia.org/enwiki/latest/

[3] Technology, Peking University, Beijing 100871;Using Tongyici Cilin to Compute Word Semantic Polarity[A];[C];2007

[4] R. Schenkel, et al., "YAWN: A Semantically Annotated Wikipedia XML Corpus."

[5] Jagrati Agrawal, Yanlei Diao, Daniel Gyllstrom, and Neil Immerman. Efficient pattern matching over event streams. In SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 147{160, New York, NY, USA, 2008. ACM.

[6] Denoyer, L., & Gallinari, P. (2006). The Wikipedia XML corpus. In Proceedings of the Fifth Workshop of the Initiative for the Evaluation of XML Retrieval. Dagstuhl, Germany.

[7] Miad Faezipour and Mehrdad Nourani. Constraint Repetition Inspection for Regular Expression on FPGA. In HOTI '08: Proceedings of the 2008 16th IEEE Symposium on High Performance Interconnects, pages 111{118, Washington, DC, USA, 2008. IEEE Computer Society.

[8] F. Zemke, A. Witkowski, M. Cherniack, and L. Colby. Pattern Matching in Sequences of Rows. In Technical Report ANSI Standard Proposal, July 2007.

[9] Yi-Hua E. Yang, Weirong Jiang, and Viktor K. Prasanna. Compact architecture for high-throughput regular expression matching on FPGA. In Mark A. Franklin, Dhabaleswar K. Panda, and Dimitrios Stiliadis, editors, ANCS, pages 30-39. ACM, 2008.

[10] Ho, D., Imai, K., King, G., and Stuart, E. (2007), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Political Analysis, 15,199–236, http://gking.harvard.edu/files/abs/matchp-abs.shtml.