IISR Crosslink Approach at NTCIR 9 CLLD Task

Chun-Yuan Cheng
Department of Computer Science and Engineering
Yuan Ze University
Chungli, Taiwan
s1006005@mail.yzu.edu.tw

Yu-Chun Wang
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
d97023@csie.ntu.edu.tw

Richard Tzong-Han Tsai
Department of Computer Science and Engineering
Yuan Ze University
Chungli, Taiwan
thtsai@saturn.yzu.edu.tw

ABSTRACT

In this paper, we describe our approach to the English-Korean Cross-Lingual Link Discovery (CLLD) task in NTCIR 9. We propose a simple and effective approach to discover the links. Our method comprises preprocessing steps, anchor-target link mapping, and the ranking steps. For discovering the links, we use the English anchor names, the inter-language links, and the translation by the Google Translate as features and extract the possible links with the exactly matching among them. Our method also ranks the anchor candidates by the Wikipedia category sets and the PageRank method, and we select the Korean target pages with the mutual information between English anchors and Korean titles of Wikipedia articles. The official file-to-file evaluation with the manual assessment of our system is achieved from 0.6 to 0.7 in P10 precision, which shows that our approach can achieve satisfactory results.

Keywords

NTCIR, Wikipedia, Crosslink Discovery

1. INTRODUCTION

The Cross-Lingual Link Discovery (CLLD) task is a research topic in which the potential link between documents among different languages is discovered automatically. It recommends a set of anchors in the source document as queries to establish the links with documents in other languages. In NTCIR-9, CLLD is first held to discover links between two different language versions of Wikipedia [1].

There are three main aspects of this task. The first one is mining the crosslink information automatically and mapping the links with the related semantics between two different languages. The second aspect is to integrate and fulfill the links in Wikipedia. The third one is to break the language barrier for the purpose of knowledge sharing.

The CLLD task is comprised of three subtasks, including E2C (English to Chinese), E2J (English to Japanese) and E2K (English to Korean). In NTCIR-9, we participate in E2K task.

Mining the potential links in Wikipedia has become a famous research field in recent years. Milne and Witten [2] adopted an information retrieval technique to sketch a vast network of concepts and semantics in Wikipedia. Huang et al. [3] also proposed a system to identify the missing links automatically.

In addition, Sorg and Cimiano [4] presented a classification-based approach to extract new cross-language links between German and English Wikipedia recently.

In 2008, the crosslink mining study sprang up. Huang et al. [5] proposed a new virtual evaluation track that was known as Cross Language Link Discovery (CLLD).

2. Methods

For the NTCIR-9 CLLD task, we construct an English-Korean CLLD system with several components. Figure 1 shows the overview of our system that comprises three main steps. The detailed implementation of these three steps is described in the following subsections.

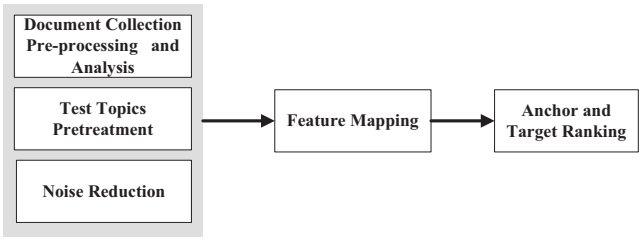

Figure 1. System overview.

2.1 Document Collection Preprocessing and Analysis

The NTCIR 9 E-K CLLD document collection set includes the mapping files of 201,596 anchors with a total of 1.2GB in size. We establish a relational database to record the anchor mapping information with extracting titles, categories, and cross-language links to other language versions from the Document Collection Set. Furthermore, in order to perform the following link discovery step, our system translates the titles of Korean anchors into English through the help of Google Translate [6] first, and we save them into the databases.

In our analysis, the document collection includes 39,798 independent categories and 3,044,968 anchor links. There are 129,696 anchors (39.2%) which do not link to others, and 204,305 anchors (61.7%) do not belong to any category.

2.2 Test Topics Preprocessing

For the 25 test topics without any anchor tags, our system adopts n-gram algorithm to extract the possible anchor candidates. We collect all the titles of the articles in English Wikipedia by using

Table 1. Feature correspond table

Feature	Candidate Anchor	Document Collection
A	English title	English title via Google translate API (Korean title)
B	Singular English title	English title via Google translate API (Korean title)
C	Singular English tilte	English title (If existed)
D	Korean title (If existed)	Korean title
E	Chinese title (If existed)	Chinese title (If existed)

Wikipedia API [7]. The n-gram anchor candidates are checked whether the candidates existed or not, where n is from 1 to 5.

If an anchor candidate exists in English Wikipedia and has the inter-language link to Chinese or Korean Wikipedia, our system stores the corresponding Chinese or Korean title into the database for further processing. In addition, for each English anchor candidate, our system transforms the anchor into a singular form if the English anchor is a noun phrase.

2.3 Noise Reduction

We collect all the possible English anchor candidates by the n-gram method in the previous step. However, there are still many useless or incorrect candidates. Thus, we use the following procedures to filter noises in candidate anchors.

1. Remove all the English stop-words with the NLTK library. [8]
2. Remove anchors that match the time or date formats in regular expressions.
3. Remove anchors that have lower-case initials.
4. Apply the maximum matching method to locate and select the longest candidate from the n-gram results.

2.4 Feature Mapping

After the document collection set and the test topics are processed, in this step, our system selects to map the English anchor candidates in the test topics with the Korean articles in the document collection. Table 1 shows the feature sets that our system uses to match the anchor candidates.

The first feature "A" matches the original English title of the anchor candidate in Wikipedia with the English translation of the Korean Wikipedia article via Google Translate. The feature "B" matches the singular form of the English anchor candidate with the English translation of the Korean Wikipedia article via Google Translate. The feature "C" maps the singular form of the English anchor candidate with the English title of the inter-language links in the Korean Wikipedia articles. The feature "D" matches the Korean title of the inter-language link existing in the English Wikipedia article of the anchor candidate with the title of the Korean Wikipedia article. The feature "E" matches the Chinese title of the inter-language link existing in the English Wikipedia article of the anchor candidate with the title of the Chinese inter-language link of the Korean Wikipedia article. The mapping method of our system is exact matching.

2.5 Anchor and Target Ranking

Since the CLLD task allows the results with a maximum of 250 anchor candidates under each test topic, we have to rank and select the most suitable anchor candidates. We select and sort all the anchor candidates via the Wikipedia categories to see which test topic they belong to. The number of the anchors in each category is counted and the anchors which belong to the category having the most number of anchors are ranked at the top. If the candidates occur in the same number of times, we then adopt PageRank Algorithm to rank these anchors.

For each anchor, the CLLD task permits up to 5 target links. Therefore, we have to rank the target links as well. We measure the mutual information score of each Korean target and the English anchor candiate in the web corpus of AltaVista search engine. The best top-5 Korean targets with the highest mutual information score are remained as the final results.

3. EVALUATION

3.1 Evaluation Metrics

In the CLLD task, there are two types of assessments: automatic assessment using the Wikipedia ground truth and manual assessment performed by human assessors. The performance of the cross-lingual link discovery system then is evaluated using Precision, Recall and Mean Average Precision metrics.

Table 2. Experiment evaluation by feature mapping

	MAP	R-Prec	P5	P10	P20	P30	P50	P250
Feature C	0.1975	0.3081	0.6667	0.6667	0.7000	0.7111	0.6267	0.1973
Feature D	0.1902	0.4034	0.4667	0.4667	0.3833	0.4000	0.5000	0.2733
Feature B	0.0800	0.2502	0.2667	0.2333	0.2833	0.3111	0.3733	0.1667
Feature A	0.0800	0.2502	0.2667	0.2333	0.2833	0.3111	0.3733	0.1667
Feature E	0.0729	0.1436	0.4000	0.5000	0.5833	0.5000	0.3800	0.1013

Table 3. Combined feature evaluation

	MAP	R-Prec	P5	P10	P20	P30	P50	P250
D + E	0.2106	0.4130	0.4000	0.4667	0.4833	0.5333	0.5667	0.2800
C + D	0.1834	0.4152	0.4667	0.4667	0.3167	0.3444	0.4267	0.2813
C + E	0.1518	0.3348	0.2000	0.2000	0.3000	0.4222	0.4400	0.2200
C + D + E	0.2057	0.4152	0.4000	0.3000	0.3333	0.4889	0.5467	0.2813

3.2 Feature Set Selection

In order to select the most suitable combinations of the different feature sets mentioned in Section 2.4, we perform several experiments to observe the performance of each feature. Table 2 and Figure 2 show the performance of each standalone feature.

From Table 2, it shows that feature C and D achieve better in MAP and R-Prec measurement. For other features, though features A and B are better than others in the R-Prec metric, but they achieve worse performance in Precision-at-N. After several experiments, we try to combine features C, D and E as the hybrid feature sets to achieve better results. Our experiment results are shown in Table 3 and Figure 3.

In Figure 3, the red line is the result of the combination with D and E; it achieves better results in the InteP-R measurement. Second, the blue line signals the combination of C and D. The green line represents the combination of C and E. The last one is the yellow line, which shows the combination of C, D and E. Combining more than 3 features cannot generate better results. The reason might be that the more candidate anchors we get, the more ambiguous the anchors are in affecting the result. Finally, we decide to submit the combination that is shown in Table 4.

3.3 Official Results

According to [9], the official results of File-to-File and Anchor-to-File are shown in Table 5 and Table 6.

The results show that tour system gets good performance in File-to-File Evaluation at P5 and P10. With the increases in the P range, the precision declines. The reason of the decreased precision could be divided into three cases. First, the maximum matching method we use in the test topic preprocessing may ignore some shorter but correct answers. It may decrease the overall accuracy. Second, the English anchors in the testing set cannot fully map onto the Korean anchors. It is the reason to explain why when the N value gets bigger, the precision declines.

Last, the ranking method used to select the anchor candidates is based on the anchor's category set; in some cases, it does not generate good results. For example, when the target anchor has a specific meaning, ranking the category sets in the article might achieve good precision. However, in the situation that the target has a general meaning, it might generate worse performance.

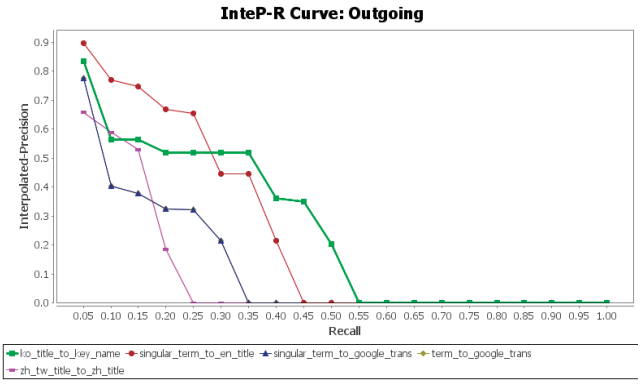

Figure 2. Interpolated Precision-Recall of English-2-Korean links by standalone feature

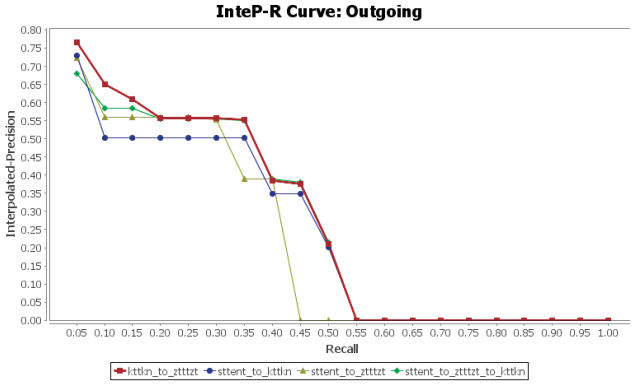

Figure 3. Interpolated Precision-Recall of English-2-Korean links by combined feature

4. CONCLUSION

For the NTCIR-9 CLLD task, we propose a simple and effective approach to discover the links between English and Korean Wikipedia collection. Our approach comprises preprocessing steps, anchor-target link mapping, and the ranking of anchors and targets. Our system has a relational database to store the preprocessing results and the detailed information of the Wikipedia articles in the document collection. For discovering the links, we use the English anchor names, the inter-language links, and the translation provided by Google Translate as the features

Table 5. File-to-File Evaluation with manual assessment results: Precision-at-N.

Run ID	P5	P10	P20	P30	P50	P250
IISR_singular_term_to_en_title	0.604	0.720	0.416	0.373	0.232	0.046
IISR_ko_title_to_key_name	0.692	0.712	0.606	0.532	0.402	0.086
IISR_kttkn_to_ztttzt	0.672	0.656	0.618	0.533	0.405	0.088
IISR_sttent_to_ztttzt_to_kttkn	0.660	0.648	0.610	0.527	0.401	0.088
IISR_sttent_to_kttkn	0.620	0.648	0.606	0.529	0.406	0.088

Table 6. Anchor-to-File Evaluation with manual assessment results: Precision-at-N.

Run ID	P5	P10	P20	P30	P50	P250
IISR_sttent_to_kttkn	0.168	0.200	0.172	0.161	0.134	0.032
IISR_singular_term_to_en_title	0.192	0.200	0.146	0.116	0.071	0.014
IISR_ko_title_to_key_name	0.204	0.184	0.176	0.163	0.135	0.031
IISR_sttent_to_ztttzt_to_kttkn	0.188	0.168	0.174	0.157	0.136	0.032
IISR_kttkn_to_ztttzt	0.192	0.152	0.178	0.161	0.133	0.032

and extract the possible links with the exact matching between them. To generate the final results, we rank the anchor candidates with the Wikipedia category sets and the PageRank method. To select the Korean target pages, we adopt the mutual information between English anchors and Korean titles of the Wikipedia articles. The official file-to-file evaluation with the manual assessment of our system achieves the level of precision in P10 from 0.6 to 0.7.

In future studies, we would like to adopt machine-learning methods by using these datasets as the training data to learn the link relationships between the two Wikipedia articles. Besides, we will try to integrate much more linguistic information of the English and Korean languages to improve the performance of CLLD.

5. REFERENCES

[1] NTCIR (NII Test Collection for IR Systems) Project. World Wide Web. http://research.nii.ac.jp/ntcir/ntcir-9/index.html

[2] David Milne, Ian H. Witten, Learning to Link with Wikipedia, Proceeding of the 17th ACM conference on Information and knowledge management, 2008.

[3] W.C. Huang, A. Trotman, S. Geva, Experiments and Evaluation of Link Discovery in the Wikipedia, Proceedings of SIGIR 2008 Workshop on Focused Retrieval, 2008.

[4] Philipp Sorg and Philipp Cimiano, Enriching the Cross-lingual Link Structure of Wikipedia – A Classification-Based Approach, Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence, 2008.

[5] W.C. Huang, A. Trotman, S. Geva, A Virtual Evaluation Track for Cross Language Link Discovery, SIGIR 2009.

[6] JSON/Atom Custom Search API - JSON/Atom Custom Search API - Google Code. World Wide Web.http://code.google.com/intl/en/apis/customsearch/v1/overview.html

[7] API:Main page - MediaWiki. World Wide Web. http://www.mediawiki.org/wiki/API:Main_page

[8] Natural Language Toolkit, http://www.nltk.org/

[9] Ling-Xiang Tang, Shlomo Geva, Andrew Trotman, Yue Xu and Kelly Itakura, Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery, The 9th NTCIR workshop Meeting, 2011.