NTCIR-10 CrossLink-2 Task: A Link Mining Strategy

Ling-Xiang Tang[1], Andrew Trotman[2], Shlomo Geva[1], Yue Xu[1]

[1]Faculty of Science and Technology,
Queensland University of Technology,
Brisbane, Australia
{l4.tang, s.geva, yue.xu}@qut.edu.au

[2]Department of Computer Science,
University of Otago,
Dunedin, New Zealand
andrew@cs.otago.ac.nz

ABSTRACT

At NTCIR-10 we participated in the cross-lingual link discovery (CrossLink-2) task. In this paper we describe our systems for discovering cross-lingual links between the Chinese, Japanese, and Korean (CJK) Wikipedia and the English Wikipedia. The evaluation results show that our implementation of the cross-lingual linking method achieved promising results.

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

Algorithms, Experimentation.

Keywords

NTCIR, CrossLink-2, Wikipedia, Link Probability, Page Name Matching.

Team Name

QUT

Subtasks

English to Chinese, English to Japanese, English to Korean, Chinese to English, Japanese to English, Korean to English

1. INTRODUCTION

At NTCIR-10, the main goal of cross-lingual link discovery (CrossLink-2) task is achieving CJK (Chinese, Japanese, and Korean) to English document linking. Plus, the subtasks for English to CJK language document linking as run at NTCIR-9 are also supported. We participated in both the English to CJK and the CJK to English subtasks.

Many good approaches to cross linking document from English to CJK languages were seen in the first cross-lingual link discovery task at NTCIR-9, and some of them were very effective in finding meaningful anchors and relevant links [1]. Among these approaches, our link mining method achieved encouraging results and the cross-lingual information retrieval method discovery the largest set of unique relevant links [2]. For the CrossLink-2 task, we continue employing the link mining method to link CJK documents to English ones without pre-processing the CJK text.

As the implementation of our CLLD system to link English documents to the CJK language ones has been detailed in the experiments for the CrossLink-1 task at NTCIR-9 [2], this paper focuses on the difference of our realisation in cross-lingual document linking from CJK language to English.

The remainder of this paper is organized as follows: First, we discuss the overall cross-lingual document linking strategy in Section 2. The experimental runs and results are discussed in Section 3. We then conclude in Section 4.

2. LINKING STRATEGY

We recorded our preliminary study on the Chinese-to-English document linking in Wikipedia with an automatic evaluation using the Wikipedia ground-truth [3]. Our previous experiments indicate that natural language processing such as Chinese segmentation for anchor identification is not absolutely required as segmentation is implicit in the anchor mining and the anchor identification processes. Also, as our systems were evaluated using only the Wikipedia ground-truth, the actual performance of the employed link mining method was unjustified. By participating in the new CrossLink-2 tasks, the performance of our CLLD system can then be benchmarked along with other systems by using the query relevance (*qrel*) obtained from the manually assessed links which are pooled from submissions of all participants of the task.

To implement a CLLD system for both E2CJK and CJK2E tasks, we simply applied a unified linking strategy for all the different language subtasks. The flowchart of how anchors are discovered and link are recommended by the unified cross-lingual linking method can be illustrated in Figure 1. The details of the linking process are given as follows.

First, prospective anchors are identified using the link mining method that relies on the pre-mined link graph of the Wikipedia corpus in source language. The target links of an identified anchor are the cross-lingual counterparts (if there are) of the existing links of that anchor. This kind of link translation is so-called triangulation [2, 3]. The biggest limitation of triangulation is that only a small portion of Wikipedia articles are cross-linked.

Secondly, if there are not enough anchors identified by the link mining method, page name matching method can be further used to search for additional anchors. Again, the link translation is still achieved through triangulation.

There are two important considerations involved in the entire link discovery process:

1) An anchor candidate will be dropped if there are no corresponding cross-lingual targets found for its associated links via triangulation.
2) As multiple targets for each anchor are allowed, many more links (if needed) can be further discovered by using the cross-lingual information retrieval method [2]. The returned items from the document retrieval system will be used as extra links of an anchor by searching it or its translation in the target corpus. The anchor translation can be obtained using an online machine translation service (specifically, Google Translate[1])

It needs to be noted that all anchor candidates identified by the link mining method or the page name matching method come with already known target link(s) as the result of link graph mining or page name mining in the Wikipedia collections. So once an anchor is identified, the target document of same language is also determined.

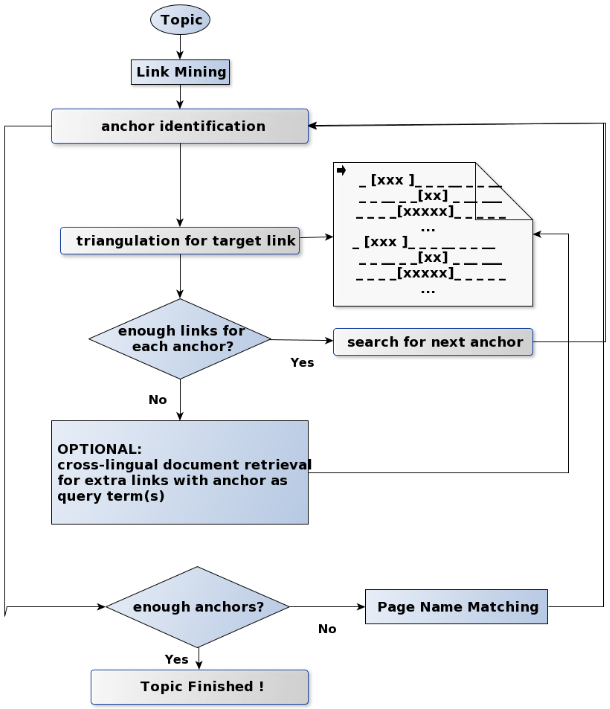


Figure 1: The CLLD System Design

3. EXPERIMENT

3.1 Collections

While our CLLD experiments relied on the standard Wikipedia collections provided in the CrossLink-2 task, we further employed additional corpora to test our Chinese / English document linking methods. The English-translated Chinese Wikipedia corpus [4] was used in the experiments of the English-to-Chinese task; and a Chinese translated (by Google Translate) English Wikipedia corpus was used for the document linking in the Chinese-to-English task. The details of the two additional pre-translated Wikipedia collections are given in Table 1.

Table 1: Statistics of the additional translated Wikipedia collections

Corpus	Documents	Size
English	400,654	2.6 GB
Chinese	3,108,756	23.0GB

3.2 Link Mining Statistics

As we mainly utilise both the link mining method and page name matching method for anchor identification, and also triangulation for translation [2, 3], the size of existing link information could decide how good our systems can be. Moreover, note that the inter Wikipedia language links created for articles of same topic do not have to be symmetrically existed [3], which means some could be missing or pointed to "incorrect" articles with similar topic. Therefore, the triangulation tables (even having same language pair but in different link direction) may contain different entries. Entries of triangulation table were mined according to the targeted language subtask in a specific link direction. Table 2 lists the statistics of triangulation tables, link mining tables and page name tables that were used in the CLLD experiments.

Table 2: Statistics of information tables used in the experiments

Triangulation Table	
Language Pair	**Entries #**
Chinese / English	233,433
Japanese / English	337,694
Korean / English	121,871
English / Chinese	211,108
English / Japanese	339, 204
English / Korean	106,704
Link Mining Table	
Corpus	**Entries #**
English	8,625,416
Chinese	860,337
Japanese	1,636,463
Korean	377,396
Page Name Table	
Corpus	**Entries #**
English	3,581,771
Chinese	370,632
Japanese	768,921
Korean	250,621

[1] http://research.google.com/university/translate/index.html

Table 3: Information of QUT Runs

Run ID	Description
CJK-to-English	
QUT_C2E_A2F_01_LinkProbPN	Primary run with link mining method plus page name matching method for supplemental anchors for topic in which there not enough 250 anchors recommended
QUT_C2E_A2F_02_LinkProbPN2	Secondary run, same as QUT_C2E_A2F_01_LinkProbPN, except for appending additional links to each anchor to make it 5 targets by using a information retrieval system—Atire. In this run, anchor is not translated, the ranked documents returned by searching anchor in the Chinese-translated English Wikipedia collection
QUT_J2E_A2F_01_LinkProbPN	Primary run with link mining method plus page name matching method for supplemental anchors for topic in which there are not enough 250 anchors recommended
QUT_J2E_A2F_02_LinkProbPN2	Secondary run, same as QUT_J2E_A2F_01_LinkProbPN, except for appending additional links to make it 5 targets for each anchor using Atire
QUT_K2E_A2F_01_LinkProbPN	Primary run with link mining method plus page name matching method for supplemental anchors for topic in which there are not enough 250 anchors recommended
QUT_K2E_A2F_02_LinkProbPN2	Secondary run, same as QUT_K2E_A2F_01_LinkProbPN, except for appending additional links to make it 5 targets for each anchor using Atire
English-to-CJK	
QUT_E2C_A2F_01_LinkProbPnCaseSensitive	Additional links are added to make it 5 targets for each anchor using Atire. In this run, anchor is not translated; the ranked documents returned by searching anchor in the Chinese-translated English Wikipedia collection
QUT_E2J_A2F_01_LinkProbPnCaseSensitive	Additional links are added to make it 5 targets for each anchor using Atire. In this run, anchor is translated with Google Translate; the ranked links are returned by searching the translation in the Japanese Wikipedia collection
QUT_E2K_A2F_01_LinkProbPnCaseSensitive	Additional links are added to make it 5 targets for each anchor using Atire. In this run, anchor is translated with Google Translate; the ranked links are returned by searching the translation in the Korean Wikipedia collection

3.3 Experimental Runs

In our experiments, a search engine named Atire[2] was employed as the document retrieval system when extra links are required. In total nine experimental runs were generated with the linking strategy discussed in section 2 for both the E2CJK and CJK2E tasks. The names and descriptions of runs are listed in Table 3.

For anchor identification in English articles, all anchor candidates were treated case sensitive; for anchor identification in CJK articles, text was not pre-segmented.

As discussed in section 2, anchors (either translated or not) could be used as query terms to obtain extra cross-lingual links from a search engine. For the Chinese / English language pair runs, with the availability of pre-translated target corpus in source language anchors need not to be translated but used directly as queries to obtain extra relevant links through the search engine; but for Japanese / English and Korean / English runs, anchors were translated by an online translation service, and then used as query terms for returning relevant links from the document retrieval system when additional links are needed.

3.4 Results and Discussion

The CrossLink-2 task uses LMAP, R-Prec, and P@N metrics for system evaluation. Performance score of a run can be easily computed against the query relevance (*qrel*) with an evaluation tool provided by the organisers [5]. The LMAP, R-Prec and P@N scores of these nine different runs are given in Table 4 and Table 5 for three evaluation scenarios (file-to-file evaluation with Wikipedia ground-truth, file-to-file and anchor-to-file evaluation with manual assessment results). In Table 4 runs are sorted on LMAP values; and in Table 5 runs are sorted on P@5 values. Interpolated precision and recall curves of runs for these three evaluation scenarios are given in Figure 2.

[2] www.atire.org

Table 4: Performance of experimental runs measured with LMAP and R-Prec

	Run ID	LMAP	R-Prec
f2f gt	QUT_E2K_A2F_01_LinkProbPnCaseSensitive	0.062	0.116
	QUT_E2C_A2F_01_LinkProbPnCaseSensitive	0.048	0.108
	QUT_E2J_A2F_01_LinkProbPnCaseSensitive	0.043	0.098
	QUT_J2E_A2F_01_LinkProbPN	0.171	0.281
	QUT_J2E_A2F_02_LinkProbPN2	0.171	0.281
	QUT_C2E_A2F_01_LinkProbPN	0.158	0.282
	QUT_K2E_A2F_02_LinkProbPN2	0.120	0.188
	QUT_K2E_A2F_01_LinkProbPN	0.120	0.188
	QUT_C2E_A2F_02_LinkProbPN2	0.059	0.111
f2f ma	QUT_E2K_A2F_01_LinkProbPnCaseSensitive	0.102	0.144
	QUT_E2C_A2F_01_LinkProbPnCaseSensitive	0.099	0.102
	QUT_E2J_A2F_01_LinkProbPnCaseSensitive	0.086	0.114
	QUT_K2E_A2F_02_LinkProbPN2	0.196	0.204
	QUT_K2E_A2F_01_LinkProbPN	0.196	0.204
	QUT_J2E_A2F_01_LinkProbPN	0.145	0.161
	QUT_J2E_A2F_02_LinkProbPN2	0.145	0.161
	QUT_C2E_A2F_01_LinkProbPN	0.069	0.132
	QUT_C2E_A2F_02_LinkProbPN2	0.037	0.049
a2f ma	QUT_E2C_A2F_01_LinkProbPnCaseSensitive	0.229	0.245
	QUT_E2K_A2F_01_LinkProbPnCaseSensitive	0.220	0.127
	QUT_E2J_A2F_01_LinkProbPnCaseSensitive	0.187	0.125
	QUT_J2E_A2F_02_LinkProbPN2	0.270	0.068
	QUT_J2E_A2F_01_LinkProbPN	0.270	0.068
	QUT_K2E_A2F_01_LinkProbPN	0.137	0.040
	QUT_K2E_A2F_02_LinkProbPN2	0.137	0.040
	QUT_C2E_A2F_01_LinkProbPN	0.089	0.087
	QUT_C2E_A2F_02_LinkProbPN2	0.089	0.087

3.4.1 Evaluation of English to CJK runs

From the *a* - plots of Figure 2, the performance of three different language runs for the English to CJK task can be easily compared: 1) all three runs has similar performance when measured against Wikipedia ground truth in F2F evaluation as showed in plot *a-(1)*; 2) plot *a-(2)* shows that English to Chinese run differentiates itself with other two language runs by having a steady precision curve across all recall points,; and 3) plot *a-(3)* indicates that in the manual assessment among all English to Chinese runs run QUT_E2C_A2F_01_LinkProbPnCaseSensitive should have the most anchors and their associated links identified correctly.

3.4.2 Evaluation of CJK to English Runs

From the *b* - plots of interpolated precision-recall curves of CJK to English runs as showed in Figure 2, the performance of these three runs is quite different from task to task.

Unlike the English to CJK CLLD tasks where all runs share a same English link table for anchor recommendation, in the experiments of CJK to English CLLD tasks we have three different link tables mined from different Wikipedia collections (Chinese, Japanese and Korean separately).

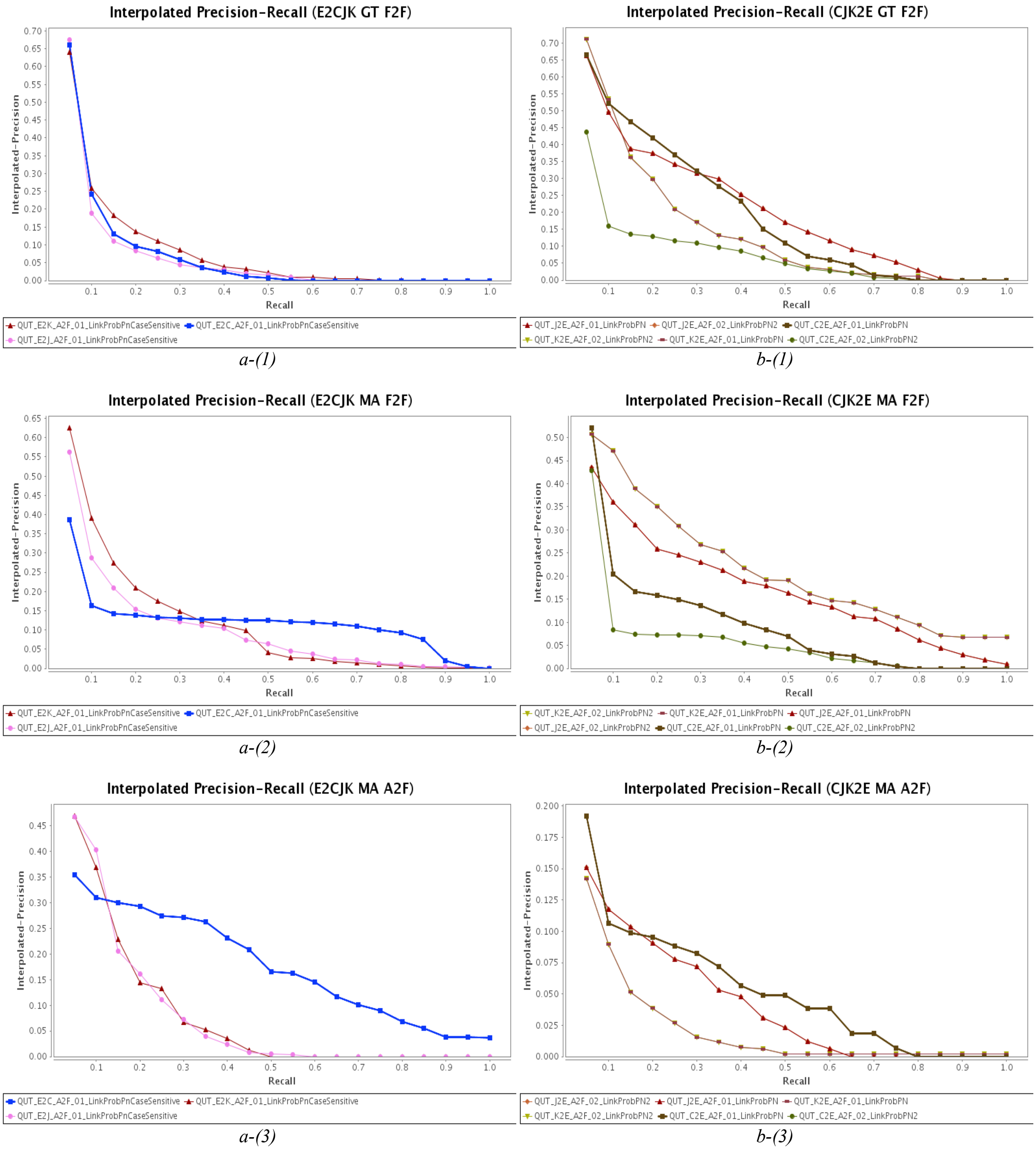
From the *b*-plots of Figure 2, it can be seen that Japanese to English runs have a fairly stable performance in all there evaluation scenarios. Interestingly, for the Chinese to English task, run QUT_C2E_A2F_02_LinkProbPN2 doesn't score well in all the F2F evaluations, but it is considered a good run, as showed in plot *b-(3)*, when evaluated in anchor-to-file level with manual assessment results.

The difference of two Chinese to English runs lies in that for run QUT_C2E_A2F_02_LinkProbPN2 extra links were retrieved by searching the anchor candidate (where not enough links recommended, as up to 5 allowed) in the pre-translated target corpus. With the boost of the LMAP scores in the anchor-to-file evaluation when measured against manual assessment results, it is further proven that using a pre-translated corpus is a very effective way for discovering relevant cross-lingual links.

Except for the Chinese to English one all other secondary runs, which have the names with the *LinkProbPN2 suffix, have the same scores as of their primary ones (*_LinkProbPN), and their interpolated precision-recall curves are overlapped too as showed in the *b*-plots of Figure 2. This could be caused by the failed translation of anchor candidates, given the fact that 1) the number of links contributed by link mining method and page name method is limited due to the natures of these methods; and 2) even a few relevant links found by searching the translated anchor (if they are correctly translated) in the target corpus could result in different evaluation scores, let alone the overlapped interpolated precision-recall curves.

Table 5: Performance of experimental runs measured with P@N

Run ID	P@5	P@10	P@20	P@30	P@50	P@250
f2f evaluation with metric scores computed against *qrel* from Wikipedia ground-truth						
QUT_E2K_A2F_01_LinkProbPnCaseSensitive	0.192	0.180	0.136	0.119	0.093	0.036
QUT_E2C_A2F_01_LinkProbPnCaseSensitive	0.216	0.188	0.150	0.124	0.105	0.042
QUT_E2J_A2F_01_LinkProbPnCaseSensitive	0.160	0.172	0.130	0.125	0.114	0.050
QUT_J2E_A2F_01_LinkProbPN	0.269	0.300	0.315	0.308	0.285	0.136
QUT_J2E_A2F_02_LinkProbPN2	0.269	0.300	0.315	0.308	0.285	0.136
QUT_C2E_A2F_01_LinkProbPN	0.352	0.376	0.368	0.332	0.285	0.104
QUT_K2E_A2F_02_LinkProbPN2	0.408	0.336	0.268	0.229	0.175	0.062
QUT_K2E_A2F_01_LinkProbPN	0.408	0.336	0.268	0.229	0.175	0.062
QUT_C2E_A2F_02_LinkProbPN2	0.080	0.084	0.104	0.104	0.108	0.080
f2f evaluation with metric scores computed against *qrel* from manual assessment						
QUT_E2K_A2F_01_LinkProbPnCaseSensitive	0.192	0.164	0.120	0.103	0.082	0.032
QUT_E2C_A2F_01_LinkProbPnCaseSensitive	0.152	0.112	0.092	0.093	0.103	0.112
QUT_E2J_A2F_01_LinkProbPnCaseSensitive	0.136	0.124	0.104	0.107	0.097	0.042
QUT_K2E_A2F_02_LinkProbPN2	0.264	0.184	0.146	0.129	0.102	0.036
QUT_K2E_A2F_01_LinkProbPN	0.264	0.184	0.146	0.129	0.102	0.036
QUT_J2E_A2F_01_LinkProbPN	0.168	0.176	0.188	0.175	0.153	0.066
QUT_J2E_A2F_02_LinkProbPN2	0.168	0.176	0.188	0.175	0.153	0.066
QUT_C2E_A2F_01_LinkProbPN	0.176	0.152	0.144	0.135	0.147	0.108
QUT_C2E_A2F_02_LinkProbPN2	0.080	0.068	0.060	0.059	0.055	0.060
a2f evaluation with metric scores computed against *qrel* from manual assessment						
QUT_E2C_A2F_01_LinkProbPnCaseSensitive	0.104	0.112	0.176	0.201	0.201	0.097
QUT_E2K_A2F_01_LinkProbPnCaseSensitive	0.192	0.156	0.124	0.104	0.078	0.022
QUT_E2J_A2F_01_LinkProbPnCaseSensitive	0.192	0.156	0.120	0.105	0.089	0.028
QUT_J2E_A2F_02_LinkProbPN2	0.048	0.04	0.072	0.065	0.072	0.037
QUT_J2E_A2F_01_LinkProbPN	0.048	0.04	0.072	0.065	0.072	0.037
QUT_K2E_A2F_01_LinkProbPN	0.048	0.048	0.042	0.039	0.034	0.014
QUT_K2E_A2F_02_LinkProbPN2	0.048	0.048	0.042	0.039	0.034	0.014
QUT_C2E_A2F_01_LinkProbPN	0.032	0.048	0.034	0.041	0.048	0.052
QUT_C2E_A2F_02_LinkProbPN2	0.032	0.048	0.034	0.041	0.048	0.052



a-(1)

b-(1)

a-(2)

b-(2)

a-(3)

b-(3)

Figure 2: The interpolated precision-recall curves of all experimental runs. Plot *a)* is the evaluation of English to CJK runs; plot *b)* is the evaluation of CJK to English runs; *(1)* is the f2f evaluation with Wikipedia ground-truth; *(2)* is the f2f evaluation with manual assessment result; *(3)* is the a2f evaluation with manual assessment result.

Unlike in the English to CJK CLLD tasks where different language runs has similar performance when measured with Wikipedia ground-truth, it is a different situation for the CJK to English runs. There may be a few reasons that can explain the overall inconsistent performance of these runs of different language pairs in the CJK to English CLLD tasks. First, the effectiveness of the link mining method depends on the availability of existing links in the corpus. And various anchors inserted by volunteer editors or bots[3] for articles even on same topic could be different and asymmetric across Wikipedia of different languages.

Furthermore, Wikipedia ground truth extraction relies on the triangulation for finding the cross-lingual counterparts of existing links in topics. So with these considerations it is not surprising that even with the same unified cross-lingual document linking strategy, the outcomes of different language runs could vary.

3.4.3 Comparison with Other Teams

The comparison of our runs with the ones from other participant teams is obtained from the evaluation data provided in the task overview paper [5].

English to CJK Task

File-to-File Evaluation with Wikipedia Ground-Truth

Our runs didn't score well in the file-to-file evaluation with Wikipedia ground-truth.

File-to-File Evaluation with Manual Assessment Results

In the file-to-file evaluation with the manual assessment results, run QUT_E2C_A2F_01_LinkProbPnCaseSensitive climbs up to the third place in the English to Chinese task, the results of our English to Japanese and English to Korean runs still unsatisfactory though. This indicates that using a translated target corpus is a very effective way for looking up relevant cross-lingual links, but a shortcoming of this method is that the target corpus has to be pre-translated.

Anchor-to-File Evaluation with Manual Assessment Results

In the English to Chinese task, our run jumps to the first place showing that our run has more anchors and their associated links being considered relevant than others.

CJK to English Task

File-to-File Evaluation with Wikipedia Ground-Truth

As the same in the English to CJK task, our runs didn't score well in the file-to-file evaluation with Wikipedia ground-truth.

File-to-File Evaluation with Manual Assessment Results

In the file-to-file evaluation with manual assessment results, our run has the best performance in Chinese to English task.

Anchor-to-File Evaluation with Manual Assessment Results

When the relevance of anchors is taken into consideration, our runs are ranked second in the evaluation of the Chinese to English task, and have the top rankings in the Japanese to English task.

Overall, our runs perform reasonable well in many tasks when measured against the manual assessment results. In some cases, our runs are ranked number one when measured with LMAP or R-Prec metric.

4. CONCLUSION

In this paper we present our experiments in realising cross-lingual linking between English and CJK Wikipedia. We used the same CLLD system developed for the first cross-lingual link discovery task at NTCIR-9 with some improvements according to the new settings.

Our unified cross-lingual document linking strategy with the link mining method was proven effective, but the performance varied when it was applied on the different language tasks with different link directions. It seemed this strategy worked better on the CJK to English tasks than on the English to CJK tasks. Our experiments also indicated that by using a pre-translated corpus for extra relevant links the performance of our CLLD system can be greatly improved especially when measured against query relevance taken from the manual assessment results.

5. REFERENCES

[1] L.-X. Tang, S. Geva, A. Trotman, Y. Xu, and K. Y. Itakura, "Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery," in *Proceedings of NTCIR-9*, Tokyo, Japan, 2011, pp. 437-463.

[2] L.-X. Tang, D. Cavanagh, A. Trotman, S. Geva, Y. Xu, and L. Sitbon, "Automated Cross-lingual Link Discovery in Wikipedia," in *Proceedings of NTCIR-9*, Tokyo, Japan, 2011, pp. 512-519.

[3] L.-X. Tang, A. Trotman, S. Geva, and Y. Xu, "Cross-Lingual Knowledge Discovery: Chinese-to-English Article Linking in Wikipedia," in *The Eighth Asia Information Retrieval Societies Conference (AIRS 2012)*, Tianjin, China, 2012.

[4] L.-X. Tang, S. Geva, and A. Trotman, "An English-translated parallel corpus for the CJK Wikipedia collections," presented at the Proceedings of the Seventeenth Australasian Document Computing Symposium, Dunedin, New Zealand, 2012.

[5] L.-X. Tang, I.-S. Kang, F. Kimura, Y.-H. Lee, A. Trotman, S. Geva, *et al.*, "Overview of the NTCIR-10 Cross-Lingual Link Discovery Task," in *Proceedings of NTCIR-10*, Tokyo, 2013.

[3] http://en.wikipedia.org/wiki/Wikipedia:Bots