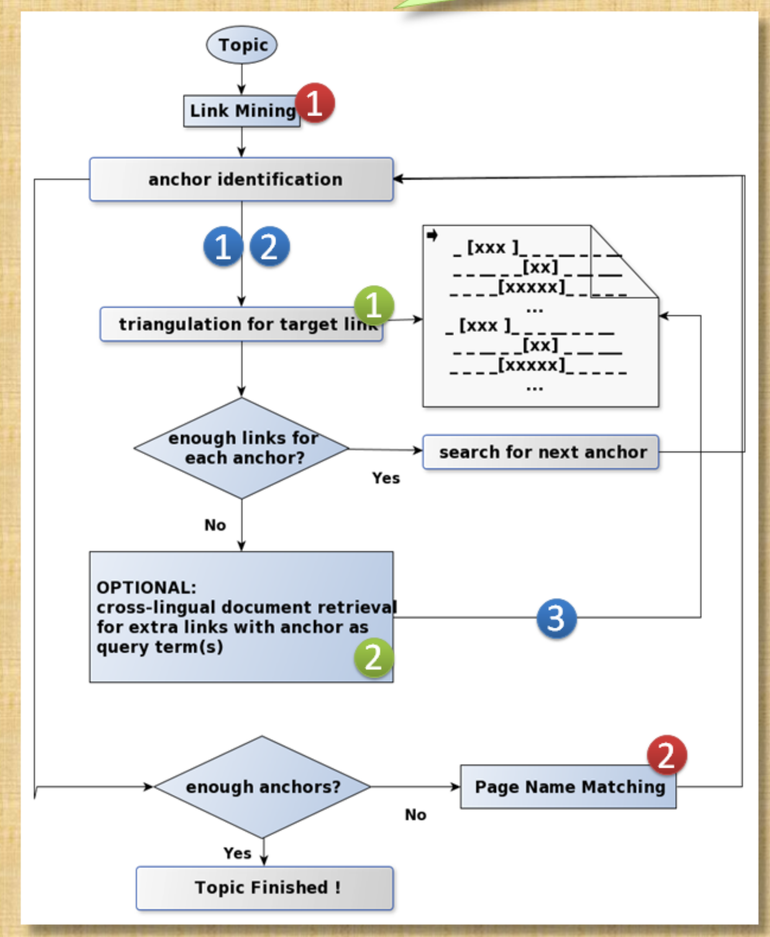
CrossLink-2 Task: A Link Mining Strategy

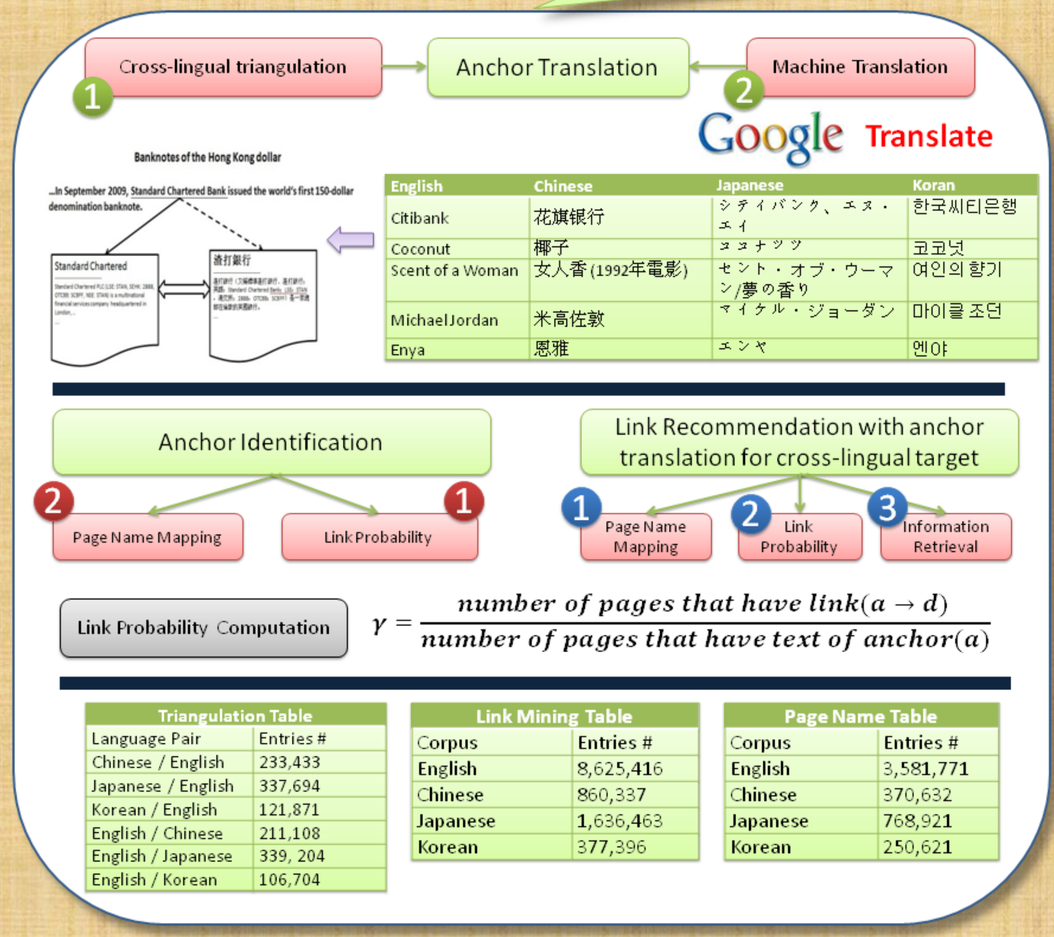
Ling-Xiang Tang[1], Andrew Trotman[2] , Shlomo Geva[1] , and Yue Xu[1]
[1]Science and Engineering Faculty, Queensland University of Technology [2]Department of Computer Science , University of Otago

System Design

A Unified Cross-Lingual Document Linking Strategy For CJK languages / English CLLD



The Implementations of Anchor Translation and Link Recommendation



The Detailed Steps

- An anchor candidate will be dropped if there are no corresponding cross-lingual targets found for its associated links via triangulation.
- 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 (if needed) can be obtained using Google Translate.

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. In these two tasks, anchors need not to be translated but used directly as queries when information retrieval method is used for finding cross-lingual targets.

The Statistics of the additional translated Wikipedia collections

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

Experiments and Results

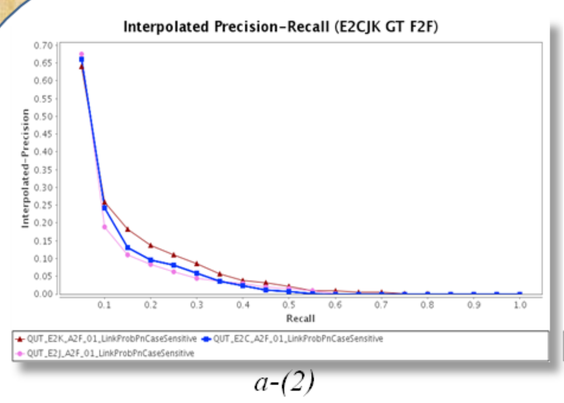
Submitted experimental runs

CJK-to-English
QUT_C2E_A2F_01_LinkProbPN: Primary run with link mining method plus page name matching method for supplemental anchors if there not enough 250 anchors recommended for the topic.
QUT_C2E_A2F_02_LinkProbPN2: Secondary run, same as QUT_C2E_A2F_01_LinkProbPN, except for appending additional links to each anchor to make all have full 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 if there are not enough 250 anchors recommended for the topic.
QUT_J2E_A2F_02_LinkProbPN2: Secondary run, same as QUT_J2E_A2F_01_LinkProbPN, except for appending additional links to each anchor to make all have full 5 targets using Atire.
QUT_K2E_A2F_01_LinkProbPN: Primary run with link mining method plus page name matching method for supplemental anchors if there are not enough 250 anchors recommended for the topic.
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: Using the unified cross-lingual document linking strategy with that additional links return from Atire are appended to each anchor to make all have 5 targets. The English link mining table is treated case sensitive.
QUT_E2J_A2F_01_LinkProbPnCaseSensitive: as run QUT_E2C_A2F_01_LinkProbPnCaseSensitive except for on the English to Japanese document linking.
QUT_E2K_A2F_01_LinkProbPnCaseSensitive: as run QUT_E2C_A2F_01_LinkProbPnCaseSensitive except for on the English to Korean document linking.

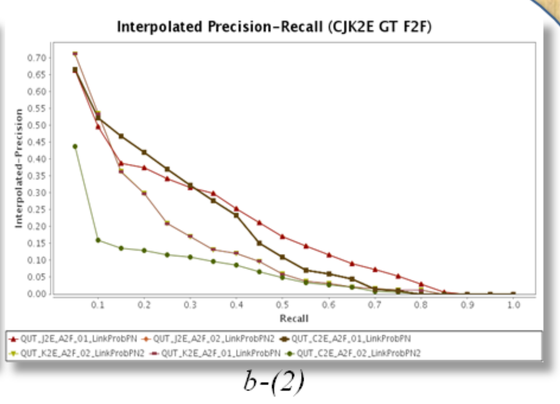
Atire (www.atire.org) - the search engine was employed as the document retrieval system when extra links are required.

Interpolated precision-recall curves of experimental runs

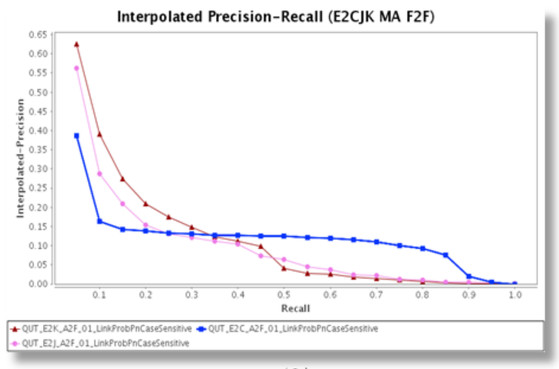
a-(1)



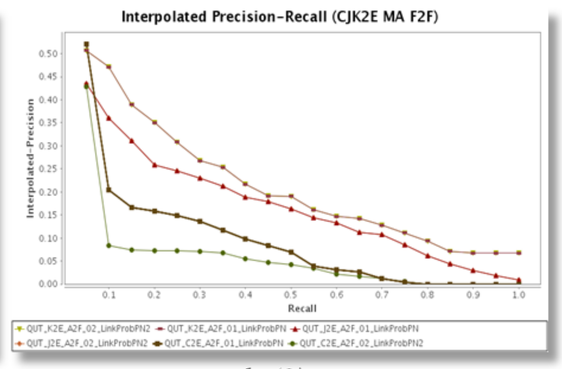
b-(1)



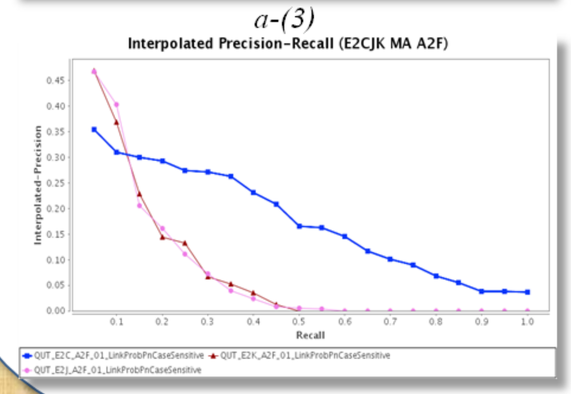
a-(2)



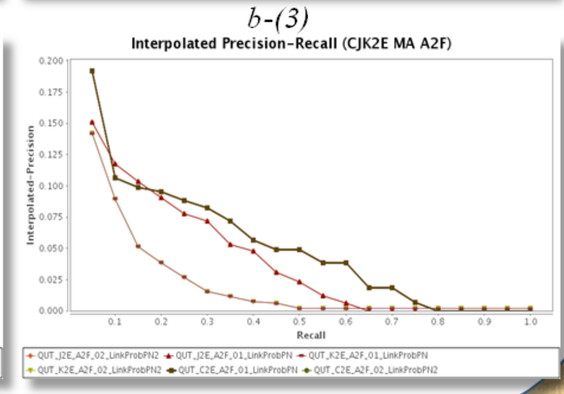
b-(2)



a-(3)



b-(3)



English to CJK Task

- In Anchor-to-File evaluation of the English to Chinese task with manual assessment results, run **QUT_E2C_A2F_01_LinkProbPnCaseSensitive** 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

- In Anchor-to-File Evaluation with Manual Assessment Results where the relevance of anchors is taken into consideration, run **QUT_C2E_A2F_01_LinkProbPN** is ranked second in the evaluation of the Chinese to English task, and run **QUT_J2E_A2F_01_LinkProbPN** have the top rankings in the Japanese to English task.

Conclusions

❖ 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.

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.