CrossLink-2 Task: A Link Mining Strategy

Ling-Xiang Tang¹, Andrew Trotman², Shlomo Geva³, and Yue Xu¹
¹Science and Engineering Faculty, Queensland University of Technology ²Department 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 disambiguated 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.

Experiments and Results

Submitted experimental runs

Interpolated precision-recall curves of 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 are 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 S targets by using a information retrieval system—Aire. In this run, anchor is not translated; the ranked documents returned by searching anchor in the Chinese-translated English Wikipedia collection.

QUT_IE2E_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_IE2E_A2F_02_LinkProbPN2: Secondary run, same as QUT_IE2E_A2F_01_LinkProbPN except for appending additional links to each anchor to make all have full S targets using Aire.

English-to-CJK

QUT_E2C_A2F_01_LinkProbPNCaseSensitive: Using the unified cross-lingual document linking strategy with that additional links return from Aire are appended to each anchor to make all have S targets. The English link mining table is treated case sensitive.

QUT_C2E_A2F_01_LinkProbPNCaseSensitive: as run QUT_E2C_A2F_01_LinkProbPNCaseSensitive except for on the English to Japanese document linking.

QUT_E2C_A2F_02_LinkProbPNCaseSensitive: as run QUT_E2C_A2F_01_LinkProbPNCaseSensitive except for on the English to Korean document linking.

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

English to CJK 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_IE2E_A2F_02_LinkProbPN have the top ranking 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


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 target.