Introduction

KMI submitted 15 runs in the NTCIR-10 CrossLink-2 evaluation achieving the best overall results in the English to Chinese, Japanese and Korean (E2CJK) task and being the top performer in the Chinese, Japanese, Korean to English task (CJK2E). All KMI methods are language agnostic and can be easily applied to any other language combination with sufficient corpora and available pre-processing tools.

Link discovery methods

1. Anchor detection
   - Link up all occurrences of dictionary terms in the orphan document.
   - Dictionaries of candidate anchors are pre-computed for each source language.
   - Each anchor corresponds to at least one concept. For example, the English dictionary contains about 34 million terms corresponding to about 4.2 million concepts.

2. Anchor filtering
   - Discard anchors with low probability
   \[ p(a) = \frac{N_a}{N} \]
   where \( N_a \) is the number of terms appearing as an anchor and \( N \) is the number of terms in the collection.

3. Disambiguation
   - Out of possible concepts, select the one with the highest score
   \[ h_c(a) = \alpha p(a) + \beta \text{sim}(ctx_a, ctx_c) \]
   where \( p(a) \) is the conditional probability of concept \( c \) given anchor \( a \) and \( \text{sim}(\cdot, \cdot) \) is the similarity of anchor's context \( ctx_a \) with the text describing concept \( ctx_c \), calculated using either Explicit Semantic Analysis (ESA) or link similarity (LIS) method.

4. Cross-language step
   - Find an equivalent concept in the target Wikipedia version to the concept selected in the disambiguation step.
   - If a cross-language link is missing for the desired language combination, we make use of the fact that the cross-language relation is transitive. Therefore, the cross-language link can be sometimes acquired using other Wikipedia language versions.

5. Ranking
   - All anchor-concept pairs are ranked, sorted and re-ranked in the specified output format.

We have experimented with 3 ranking methods:
1. Anchor probability ranking
2. Machine learned ranking
3. Oracle ranking

Evaluation methodology

The use of ESA for disambiguation in CJK2E: ESA was applied in E2CJK tasks, where it performed consistently better than link similarity.

Anchor detection: Our system did not detect anchors that were only part of a term, contrary to other systems.

Tuning parameters in the disambiguation step: It might be possible to determine more optimal disambiguation parameters by further tuning or machine learning.

Considering more than one disambiguation per anchor in the first step: Our methods currently select the best disambiguation for each anchor in the first round and the second best, third best, etc. disambiguation only in the following rounds. It might be possible to achieve better performance in manual assessment if more than one disambiguation is assigned in the first round.

What have we learned?

ESA vs link similarity disambiguation: Our experiments show that ESA outperforms link similarity.

Ranking strategy: While the optimal ranking technique (ORC runs) with the Wiki ground-truth (WT), for which they were optimised, achieve substantially higher performance than our anchor probability ranking runs, the ESA runs perform equally well when applied to a different GT.

Conclusion

We understood the importance of the ranking phase, experimentally confirmed the impact of high variance in the ground-truth on the CLLD results, measured the maximum (theoretical boundary) performance of an ideal CLLD system and analysed some of the evaluation pitfalls.

We believe this knowledge will help us to better understand how to more representatively measure the performance in the future, which will, in turn, enable further evidence-based improvements of link discovery systems.