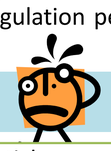
Automated Cross-lingual Link Discovery in Wikipedia

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

ABSTRACT

At NTCIR-9, we participated in the cross-lingual link discovery (Crosslink) task. In this paper we describe our approaches to discovering Chinese, Japanese, and Korean (CJK) cross-lingual links for English documents in Wikipedia. Our experimental results show that a link mining approach that mines the existing link structure for anchor probabilities and relies on the "translation" using cross-lingual document name triangulation performs very well. The evaluation shows encouraging results for our system.

1. CROSS-LINGUAL LINKING IN WIKIPEDIA



Among all language sub-sets of Wikipedia, English Wikipedia contains the largest number of articles. However, the links in the current English Wikipedia are mainly pointed at articles of the same language. Without direct links to articles in other languages, it may cause difficulties when viewing cross-lingual materials for people who are bi-lingual readers or knowledge contributors, or second language acquisition students (e.g. English learners of Chinese).

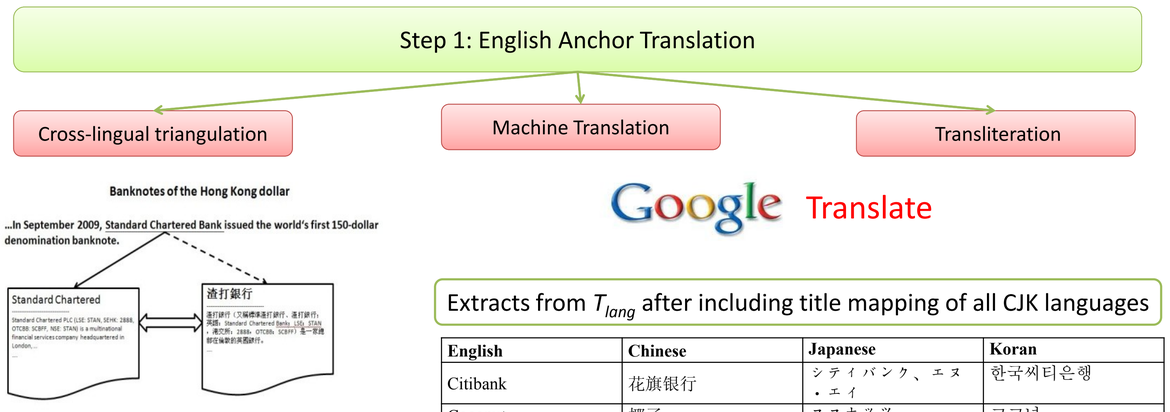
2. CLLD METHODS

To locate CJK cross-lingual links for English Wikipedia articles, we separate the link discovery into two phases:

1) **detecting prospective anchors in the source document;**
2) **and for each anchor, identifying relevant documents in the target language corpus. Once the anchor is identified, a link, *a*→*d*, is created (where *a* is the anchor, *d* is the target document).**

What to linK?

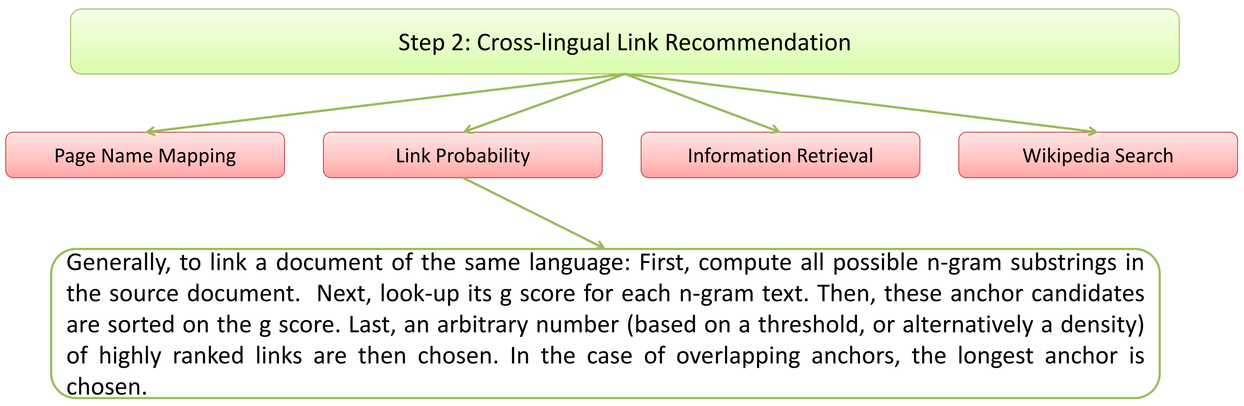
- Cross-lingual Link Probability (English-to-Chinese)
- Cross-lingual Page Name Matching (English-to-Chinese, English-to-Japanese, English-to-Korean)
- Cross-lingual Information Retrieval (English-to-Chinese)
- Named Entity Recognition with Transliteration (English-to-Japanese)



An example of cross-lingual triangulation. It can be used in page name matching and link probability methods for anchor translation.

Extracts from T_{lang} after including title mapping of all CJK languages

English	Chinese	Japanese	Koran
Citibank	花旗银行	シティバンク、エヌ・エイ	한국씨티은행
Coconut	椰子	ココナッツ	코코넛
Scent of a Woman	女人香 (1992年電影)	セント・オブ・ウーマン/夢の香り	여인의 향기
Michael Jordan	米高佐敦	マイケル・ジョーダン	마이클 조던
Enya	恩雅	エンヤ	엔야

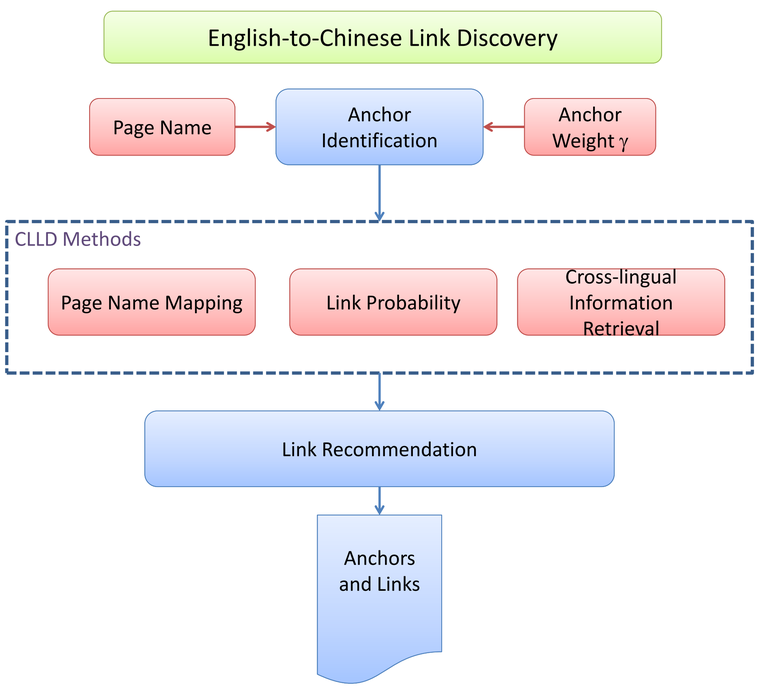


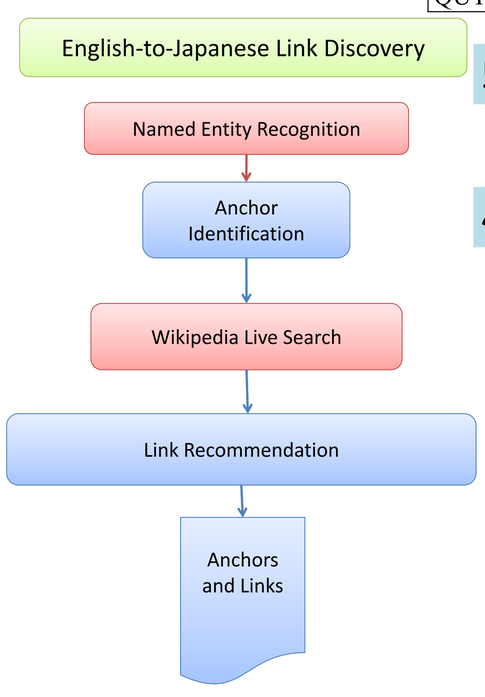
Generally, to link a document of the same language: First, compute all possible n-gram substrings in the source document. Next, look-up its g score for each n-gram text. Then, these anchor candidates are sorted on the g score. Last, an arbitrary number (based on a threshold, or alternatively a density) of highly ranked links are then chosen. In the case of overlapping anchors, the longest anchor is chosen.

$$\gamma = \frac{number\ of\ pages\ that\ have\ link(a \rightarrow d)}{number\ of\ pages\ that\ have\ text\ of\ anchor(a)}$$

Method	Pros	Cons
ML	More accurate, less noisy	Only finds links already in the corpus
PNM	Simple, effective	Only finds links matched with the page title
CLIR	Finds links not seen elsewhere in the corpus	May be noisy
TRANSLITERATION	Simple	May not be very accurate

3. IMPLEMENTATIONS OF CROSS-LINGUAL LINK DISCOVERY





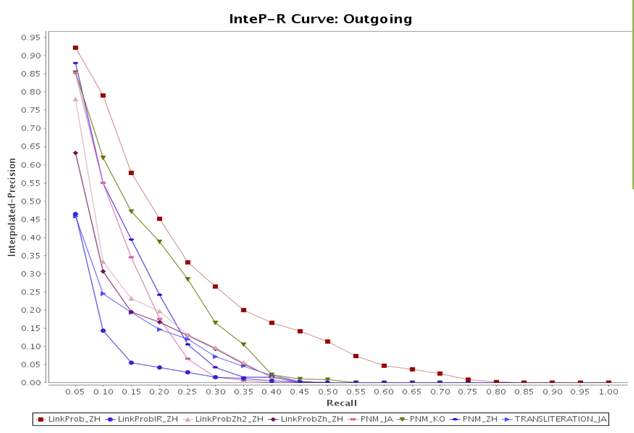
CONCLUSION

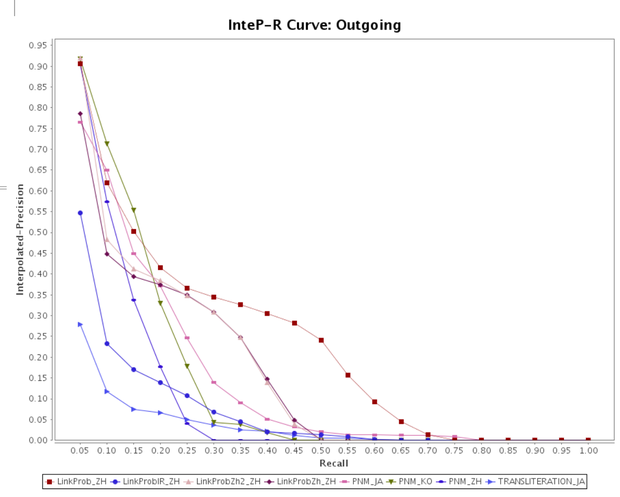
Several automatic linking methods were tested. The methods employed include: link mining, page name matching, cross-lingual information retrieval and transliteration with online Wikipedia search service.

Link mining method with Wikipedia cross-lingual document name triangulation (run: *QUT_LinkProb_ZH*) performed the best among all implementations, and also achieved encouraging results in the overall evaluations of Crosslink task. This method requires pre-mining on the existing link structure of Wikipedia. In order to compute a list of English anchor / target probabilities, additional English Wikipedia corpus from INEX[5] was employed for this English link mining.

7. RESULTS AND DISCUSSIONS

	Run ID	MAP	R-Prec	P@5	P@10	P@20	P@30	P@50	P@250
	metric scores computed with *qrel* from Wikipedia ground-truth								
f2f	LinkProb_ZH	0.179	0.244	0.776	0.588	0.480	0.404	0.319	0.132
	PNM_KO	0.122	0.208	0.552	0.460	0.384	0.321	0.244	0.062
	PNM_ZH	0.088	0.166	0.592	0.472	0.362	0.307	0.242	0.064
	PNM_JA	0.076	0.143	0.624	0.504	0.394	0.333	0.262	0.079
	LinkProbZh2_ZH	0.069	0.154	0.360	0.284	0.248	0.221	0.187	0.082
	LinkProbZh_ZH	0.059	0.148	0.304	0.208	0.168	0.161	0.156	0.082
	TRANSLITERATION_JA	0.047	0.145	0.160	0.136	0.126	0.139	0.152	0.099
	LinkProbIR_ZH	0.023	0.067	0.184	0.160	0.118	0.109	0.084	0.044
	metric scores computed with *qrel* from manual assessment								
a2f	LinkProb_ZH	0.115	0.133	0.336	0.308	0.294	0.288	0.277	0.172
	LinkProbZh_ZH	0.094	0.119	0.320	0.244	0.260	0.273	0.269	0.158
	LinkProbZh2_ZH	0.090	0.117	0.312	0.312	0.304	0.299	0.271	0.155
	PNM_JA	0.087	0.016	0.128	0.124	0.108	0.096	0.077	0.020
	PNM_KO	0.043	0.043	0.136	0.200	0.220	0.217	0.193	0.047
	PNM_ZH	0.030	0.033	0.208	0.204	0.214	0.220	0.187	0.045
	LinkProbIR_ZH	0.008	0.026	0.104	0.104	0.072	0.073	0.070	0.033
	TRANSLITERATION_JA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000





The interpolated precision/recall curves of runs.

Left:
The f2f evaluation using Wikipedia ground-truth

Right:
The a2f evaluation using manual assessment result

Overall, our runs, especially those submitted for the English-to-Japanese and English-to-Korean tasks, have medium performance when compared to the other good runs submitted to the task. But we contribute the largest number of unique relevant links that users might think they deserve further reading.

6. EXPERIMENTAL RUNS

Run ID	Description
QUT_PNM_ZH	Use the PNM algorithm, Chinese Wikipedia Corpus for title-to-target table
QUT_LinkProbIR_ZH	Use the anchors recommended by link probability, and retrieve relevant links using a search engine with anchors as query terms
QUT_LinkProbZh2_ZH	Same as QUT_LinkProbZh_ZH , except for that anchors are sorted based on Chinese link probability table.
QUT_LinkProbZh_ZH	Use two set of link probability tables (one Chinese; one English mining from English Wikipedia corpus from INEX), and tables are connected by translation. Anchors are sorted based on English link probability table.
QUT_LinkProb_ZH	Use link probability for anchor sorting and link recommendation
QUT_PNM_JA	Use the PNM algorithm, Japanese Wikipedia Corpus for title-to-target table
QUT_TRANSLITERATION_JA	The Stanford Named Entity Recogniser is used With Google Translate, and connect to Wikipedia's Japanese search engine to identify suitable pages to link to.
QUT_PNM_KO	Use the PNM algorithm, Korean Wikipedia Corpus for title-to-target table

5. INFORMATION RETRIEVAL: CHINESE DOCUMENTS INDEXING

Unigrams, bigrams and words are all common tokens used when indexing Chinese text.

4. INFORMATION RETRIEVAL: WEIGHTING MODEL – BM25

A slightly modified BM25 ranking function was used for document ordering.

$$IDF(q_i) = log\frac{N}{n}$$

Where N is the number of documents in the corpus, and n is the document frequency of query term . The retrieval status value of a document d with respect to query is calculated as:

$$rsv(q,d) = \sum_{i=0}^{m} \frac{tf(q_i,d) * (k_1 + 1)}{tf(q_i,d) + k_1 * \left(1 - b + b * \frac{len(d)}{avgdl}\right)} * IDF(q_i)$$

Parameters k_1 and b were 0.7 and 0.3 respectively (values previously shown to be effective).