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).

- Cross-link Probability (English-to-Chinese)
- Cross-lingual Information Retrieval (English-to-Chinese)
- Named Entity Recognition with Transliteration (English-to-Japanese)

What to link?

Step 1: English Anchor Translation

Cross-lingual translation

Google Translate

Extracts from T1map after including title mapping of all CJK languages

Step 2: Cross-lingual Link Recommendation

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

Generally, to link a document of the same language: First, compute all possible n-gram substrings in the source text. 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.

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

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

4. INFORMATION RETRIEVAL: WEIGHTING MODEL – BM25

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

\[ IDF(q_i) = \log \frac{N}{\sum_{j=0}^{m} t_f(q_i,d_j)} \]

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 q is calculated as:

\[ rBM25(q,d) = \sum_{d} t_f(q,d) \times (k_1 + 1) \times \frac{t_f(q,d)}{k_1 + t_f(q,d)} \times \text{len}(d) \text{avg} \times IDF(q) \]

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