Wikipedia and Web document based Query Translation and Expansion for Cross-language IR

Ling-Xiang Tang\textsuperscript{a}, Andrew Trotman\textsuperscript{b}, Shlomo Geva\textsuperscript{c} and Yue Xu\textsuperscript{d} \textsuperscript{a}Faculty of Science and Technology, Queensland University of Technology \textsuperscript{b}Department of Computer Science, University of Otago

Abstract

In this paper we describe our approaches to retrieving cross-lingual documents for question answering in the NTCIR-IRA QA task. A few Chinese indexing techniques were used in our experiments. We mainly focused on using external resources: web documents and Wikipedia for the key phrase identification, translation and query expansion. The evaluation shows encouraging results of our system.

1. Questions in IRAQA for ACLIA

1.1. How to deal with the ODV issues in translation?
1.2. How to deal with language variants?
1.3. How to deal with the ambiguity in translation?
1.4. How to deal with the 0.1040 indexing issues?
1.5. How to deal with the 0.4905 indexing issues?
1.6. How to achieve bettered 0.2656 performance?
1.7. How to get the 0.5882 performance?
1.8. How to get the 0.5555 performance?
1.9. How to get the 0.3086 performance?
1.10. How to get the 0.1673 performance?
1.11. How to get the 0.4139 performance?

2. Observations

Wikipedia, including 100,000 entries in simplified, traditional Chinese, or different Chinese variants describing various up-to-date events, people, organizations, locations, and facts.

Bilingual Web documents. When people post information on the Internet, they often provide a translation (where necessary) in the same web documents. These pages contain bilingual phrase pairs.

e.g. a youtube entry: YouTube - Sean Chen [视频] dumps on Yao Ming

A web search engine such as Google can identify Wikipedia entries, and return popular bi-lingual web documents that are closely related to a named entity.

Statistical machine translation relying on parallel corpora such as Google Translate can achieve very high translation accuracy.

3. An automated online live translation solution

Web Search Engine (Google, Bing, etc.)

Wikipedia

Bilingual Web documents

General Machine Translation (Google Translate)

An automated online live translation solution

4. IRAQA System Design

1. Remove stop words from the English question to generate term chunks.
2. Search Google using the term chunks with options set to return only the Chinese and English page. If there is a Wikipedia page at the top 10 then go to step 4, else go to step 3.
3. Calculate the frequency of words appearing in the title of the results from step 2. Choose the top ranking ones with a frequency above a threshold. The threshold is set to 3 to begin with then linearly decreased by one until at least one term is found. Search Google using these words, but restrict the search to the English Wikipedia. The highest ranking result containing any term from the question is considered the correct result.
5. Translate the corresponding Chinese Wikipedia page by following the languages link from the English page. If there is no Chinese link found, then jump to step 7.
6. Extract the title of the Chinese Wikipedia page, and use it as the query.
7. Process the clue text if there is any for query expansion. Either the whole or some frequent words of the clue text could be used as expansion terms. Append these additional terms to the Chinese query. The frequent words are chosen from the NGMI segmented terms in the clue text. The frequency of those words has to be above a threshold. The threshold is set to 5 to begin with then linearly decreased by one until at least one term is found. Search Google using these words, but restrict the search to the English Wikipedia. The highest ranking result containing any term from the question is considered the correct result.
8. Search the corpus using either: single character segmentation; or single character segmentation plus NGMI segmented terms.

5. Weighting Model – BM25

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

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:

6. Chinese Documents Indexing

Unigrams, bigrams and words are all common tokens used when indexing Chinese text. The performance of various IR systems combining different segmentation algorithms can be very different.

In our experiments, we used n-gram mutual information (NGMI) to segment Chinese text. NGMI is an unsupervised n-gram word segmentation approach. It is derived from character-based mutual information, but can additionally recognize words longer than two characters.

In order to test how NGMI can make a difference in Chinese IR and find a suitable segmentation strategy for our CLIR system, unigram indexing and dual indexing (unigrams and NGMI segmentation) were used in different experimental runs.