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

Ling-Xiang Tang[1], Andrew Trotman[2] , Shlomo Geva[1] and Yue Xu[1] [1]Faculty of Science and Technology, Queensland University of Technology [2]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 ACLIA-IR4QA task. A few Chinese indexing techniques were used in our experiments. We mainly focused on using external recourses: web documents and Wikipedia for the key phrase identification, translation and query expansion. The evaluation shows encouraging results of our system.

## 1. Questions in IR4QA for ACLIA

Is traditional translation by a MT software/toolkit to achieve CLIR absolutely necessary?

How to deal with the OOV issues in translation?

How to resolve the ambiguity in translation?

How to deal with language variants?

Liu Xiang - 刘向? Or 刘翔?

What is "laser printer" called in Chinese?

| | |
|---|---|
| laser printer (American English) | 激光打印机 (Mainland China) |
| laser printer (Australian English) | 鐳射打印機 (Hongkong) |
| laser printer (British English) | 雷射印表機 (Taiwan) |

## 2. Observations

Wikipedia, including +100,000 entries in simplified , traditional Chinese , or differnt 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* ( 陳信安) dunks 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 corpus 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

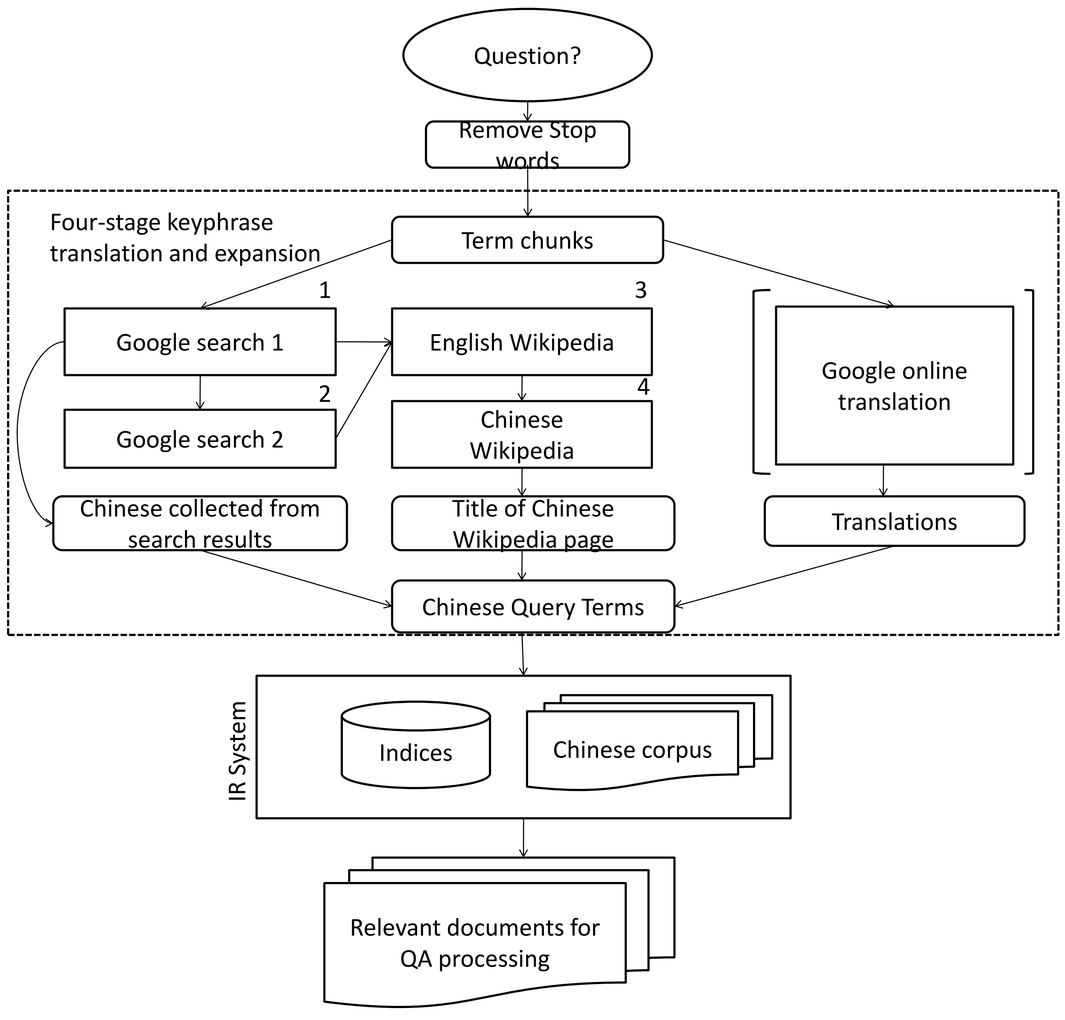
+

General Machine Translation (Google Translate)

=

An automated online live translation solution

## 4. IR4QA System Design




## 5. 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:

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

## Conclusion

The strategy relies on an external resource such as web documents and the Wikipedia can effectively tackle the out-of-vocabulary problem.

It is possible to totally rely on Web search engine and Wikipedia to obtain key phrase translation for question in QA.

Also, this hybrid translation approach can greatly extend the size of query terms, and make related documents be seen as many as possible.

## 9. Official IR4QA Results(BEFORE BUG FIXES) and Discussions

| EN-CS Runs | # EN Wiki | # ZH Wiki | # Clue text | # FAIL |
|---|---|---|---|---|
| 01, 02, 03, 04 | 89 | 65 | 24 | 27 |
| 05 | 57 | 43 | 93 | 6 |

| EN-CT Runs | # EN Wiki | # ZH Wiki | # Clue Text | # FAIL |
|---|---|---|---|---|
| 01, 02, 03, 04 | 86 | 71 | 36 | 25 |
| 05 | 70 | 52 | 96 | 2 |

The statistics of the EN-CT and EN-CS runs. #FAIL is the number of total topics for which the web search could not obtain either a Chinese Wikipedia page or Chinese clue text.

| RUN ID | MAP | MQ | MnDCG |
|---|---|---|---|
| EN-CS best | 0.4139 | 0.4499 | 0.6509 |
| QUTIS-EN-CS-01-T | 0.1420 | 0.1689 | 0.3527 |
| QUTIS-EN-CS-02-T | 0.1673 | 0.1967 | 0.4028 |
| QUTIS-EN-CS-03-T | 0.2504 | 0.2886 | 0.5127 |
| QUTIS-EN-CS-04-T | **0.3198** | **0.3607** | **0.5882** |
| QUTIS-EN-CS-05-T | 0.2752 | 0.3086 | 0.5245 |

| RUN ID | MAP | MQ | MnDCG |
|---|---|---|---|
| EN-CT best | 0.4900 | 0.5263 | 0.7175 |
| QUTIS-EN-CT-01-T | 0.1943 | 0.2218 | 0.3997 |
| QUTIS-EN-CT-02-T | 0.2161 | 0.2501 | 0.4374 |
| QUTIS-EN-CT-03-T | 0.2656 | 0.2957 | 0.4905 |
| QUTIS-EN-CT-04-T | **0.3231** | **0.3569** | **0.5555** |
| QUTIS-EN-CT-05-T | 0.1040 | 0.1167 | 0.2492 |

Runs 01 and 02 that use just web documents and the Wikipedia (where possible) achieve a relatively low precision. This, however, does indicate that web documents and the Wikipedia may be good resources for query translation.

In both cases the runs that restricted the web search to a particular domain (run 05) performed worse than the equivalent run that did not site restrict.

The runs using NGMI word dual indexing (runs 02, 04, and 05) bettered the equivalent character based indexing in all cases.

The QUTIS-EN-CS runs, QUTIS-EN-CS-01-T particularly, contribute the highest number of unique relevant documents in all CS submissions according to the official NTCIR assessment results.

## 8. Experimental Runs

| RUNID | Index Units | Translation | Query Units | IR Model |
|---|---|---|---|---|
| QUTIS-EN-CS-01-T | unigram | 1.Web search + 2. [Google Translate, if 1 fails] | unigram | BM25 |
| QUTIS-EN-CS-02-T | unigram + word | 1.Web search + 2. [google translate, if 1 fails] | unigram + word | BM25 |
| QUTIS-EN-CS-03-T | unigram | Web search + Google Translate | unigram | BM25 |
| QUTIS-EN-CS-04-T | unigram + word | Web search + Google Translate | unigram + word | BM25 |
| QUTIS-EN-CS-05-T* | unigram + word | Web search (site:cn) + Google Translate | unigram + word | BM25 |

| RUNID | Index Units | Translation | Query Units | IR Model |
|---|---|---|---|---|
| QUTIS-EN-CT-01-T | unigram | 1.Web search + 2. [Google Translate, if 1 fails] | unigram | BM25 |
| QUTIS-EN-CT-02-T | unigram + word | 1.Web search + 2. [Google Translate, if 1 fails] | unigram + word | BM25 |
| QUTIS-EN-CT-03-T | unigram | Web search + Google Translate | unigram | BM25 |
| QUTIS-EN-CT-04-T | unigram + word | Web search + Google Translate | unigram + word | BM25 |
| QUTIS-EN-CT-05-T* | unigram + word | Web search (site: tw) + Google Translate | unigram + word | BM25 |

## 7. How this hybrid translation approach works

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. In both cases Chinese text present in the English results list is collected – this clue text is used for query expansion.
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.
4. Retrieve the English Wikipedia page.
5. Retrieve 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.
8. If no Chinese query terms have been found then use the translation from Google Translate on the term chunks.. Otherwise, the results of this machine translation are optionally appended to the query.
9. Search the corpus using either: single character segmentation; or single character segmentation plus NGMI segmented terms.

The resulting query can be thought-of as:
*Chinese query*
*=*
*The title of the Chinese Wikipedia document found by searching Google for the English Wikipedia document and following the languages link*
*+*
*Chinese clue text collected from Google search (all or high frequency words)*
*+*
*Result from Google Translate (optional)*

## 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.