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

Ling-Xiang Tang[1], Andrew Trotman[2], Shlomo Geva[1], Yue Xu[1]

[1]Queensland University of Technology, Australia
[2]University of Otago, New Zealand

Introduction

- Cross-lingual QA is getting more attentions because of the needs for cross-lingual information/answer seeking in the fast-changing, multilingual world of information.

- For information/answer seekers, nothing is better than to be able to use their natural languages to compose their questions precisely without worrying the burden of translation.

Introduction (cont'ed)

- A simple approach to achieving CLIR(IR4QA) for CLQA system to process the information and return the answer automatically is to translate the query into the language of the documents and to use a mono-lingual IR system.

- However, it is difficult for general Machine Translation to catch up the changes of the world. For the question containing term that is new, OOV, general MT could produce noise words rather than to provide the correct translation.

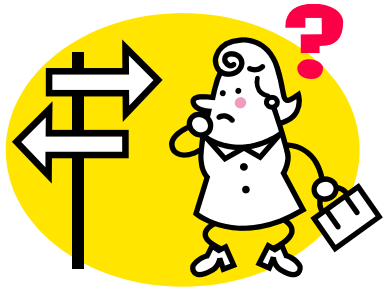
Questions

- 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)



Observations

- The Wikipedia has over 100,000 Chinese entries describing various up-to-date events, people, organizations, locations, and facts. Most importantly, there are links between English articles and their Chinese counterparts.

- 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. For example, if an English term/phrase is used in a Chinese article, it is often followed by the Chinese translation enclosed in parentheses.

Observations (cont'ed)

- 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.

Answers

- Web Search Engine (Google, Bing, etc.)

↓

- Wikipedia
- Bilingual Web documents

+

- General Machine Translation (Google Translate)

=

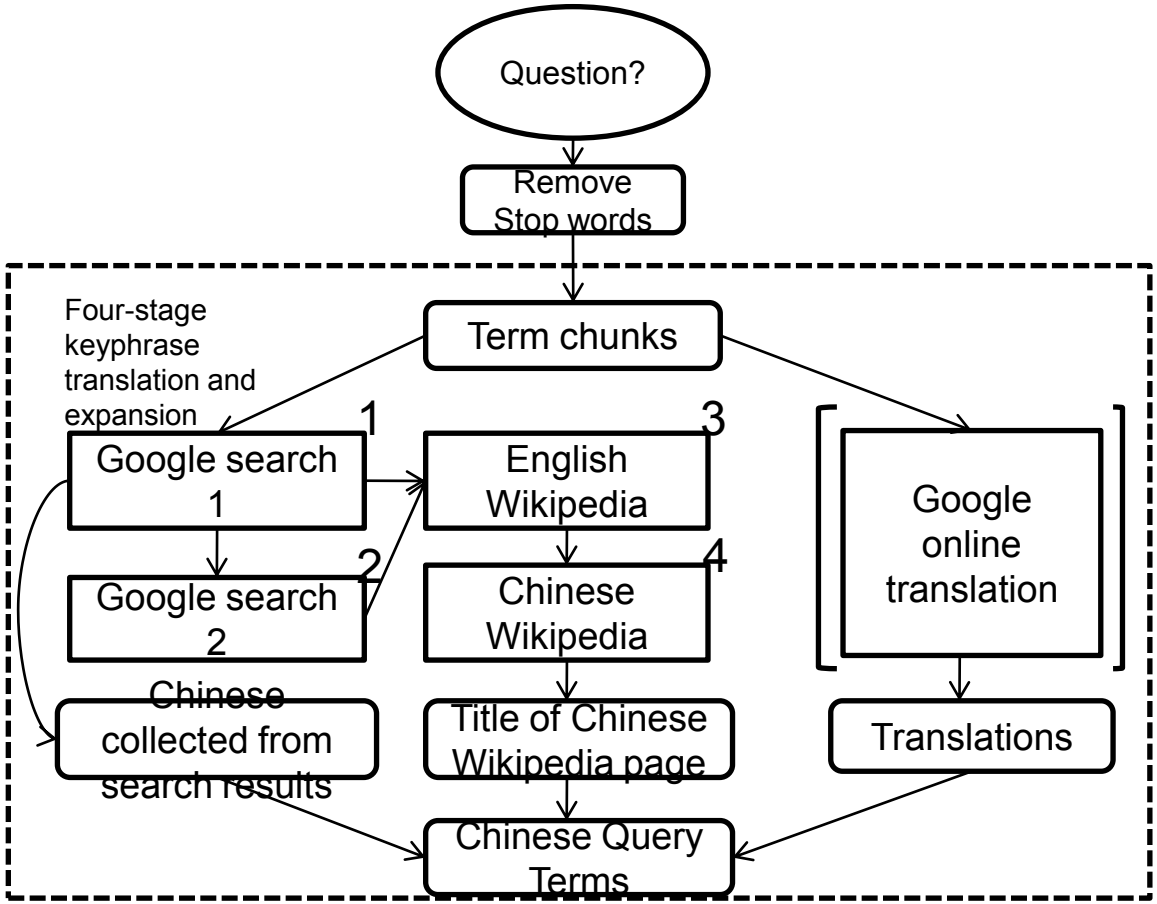
An automated online live translation solution

Previous Works

- Ferrández, et al(2007) used Wikipedia to solve low coverage issue on named entities in EuroWordNet.

- Researchers in previous NTCIR workshops employed Wikipedia for person name or other NEs particularly OOV terms translation.

Query Generation



Query Generation (cont'ed)

- *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)*

Weighting Model – BM25

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

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

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:

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

Chinese Document 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.

Experiment Runs (CS)

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

* Run EN-CS 05 searched only in the cn domain

Experiment Runs (CT)

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

* Run EN-CT 05 search only in the tw domain.

Runs Statistics

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.

Official IR4QA Results BEFORE bug fixes

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

Discussions

- 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.

Discussions (cont'ed)

- 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.

Conclusions

- 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.

QA