Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment

K.L. Kwok
Computer Science Dept., Queens College, City University of NY,
Flushing, NY 11367, USA.
kklqc@cunyvm.cuny.edu

Abstract

We conducted retrieval experiments using a collection of 170MB Chinese text from TREC-5. The supplied queries (topics) are in both English and Chinese, the latter to a large extent may be considered as very good translation of the former. Another set of Chinese queries were created by translating the English via a simple dictionary look-up procedure. Retrieval effectiveness for both types of queries were evaluated, essentially giving upper and lower bounds of how machine translation may contribute to this cross-lingual retrieval experiment. Results show that naive translation returns effectiveness about 30% to 50% worse than good translation.

1 Introduction

Cross-Lingual Text Retrieval (CLTR) aims at combining machine translation (MT) tools with an information retrieval (IR) system to facilitate users to retrieve documents in one language while posing queries in another. Circumstances under which such capability may be useful has been surveyed in [Oard96]. Both MT and IR have long histories in computing, and they are difficult tasks involving conceptual accuracy. The output from MT is a piece of text in another language that is usually aimed for human consumption, and has stringent requirements such as being grammatic, having good linguistic style, readability, etc. in addition to content correctness. Large scale IR for unrestricted domain on the other hand still functions at a content word level augmented with phrases, and relies on statistical evidence for retrieval. Knowledge-intense methodologies in such an environment are difficult and not too successful. Because of this mismatch, it is possible that some rudimentary MT tool is sufficient for CLTR, such as translating only content terms and noun phrases in a piece of text.

This paper intends to explore the range of retrieval effectiveness that MT tools may affect CLTR. We assume that a user of an CLTR system has control over query composition but not the documents. Knowing the difficulties in translation, a user would know to use simple direct wordings in his/her mother language to convey what s/he wants and avoid difficult texts such as idioms, acronyms, etc. We employ queries that are 'well-translated' to a target language used in the documents, as well as the simplest cross-lingual procedure of looking up content terms in a dictionary. Both types of queries are then used for retrieval in a monolingual fashion, and a quantitative comparison is made between the two. This essentially provides an upper and lower bound to CLTR because more sophisticated MT procedures having some sense disambiguation capability would probably give performance lying between the two. The specific languages involved is English queries on Chinese documents. Other researchers have also done CLTR work. Examples for other language pairs include: [Davis & Dunning 1996, Davis 199x] English-Spanish, [Fluhr], [Hull & Grefenstette, 1996] English-French and [Sheridan & Ballerini 1996] German-Italian. [Lin & Chen 1996] has worked on automatic classification of documents containing both Chinese and English texts.

2 Experimental Set-Up

2.1 The TREC-5 Chinese Collection

Over the past several years the NIST-DARPA sponsored Text REtrieval Conference (TREC) have provided the IR community with large, realistic document collections for experimentation as well as unbiased manual relevance judgments between queries and documents for evaluation purposes. In the 1996 TREC-5 cycle [Harman 9x], Chinese language IR experiments are included for the first time, with queries (topics) expressed in both English and Chinese. This provides an opportunity to study cross-lingual retrieval issues employing a large collection using a set of queries that has unbiased judgments for evaluation.

The collection of documents consists of 24,988 Xinhua and 139,801 People's Daily news articles totaling about 170 MB. To guard against very long documents which can lead to outliers in frequency estimates, these are segmented into subdocuments of about 550 characters in size ending on a paragraph boundary. The total number of subdocuments is 231,527. These subdocuments are indexed based on a simple 4-step segmentation algorithm that aims at discovering short-words of 2 or 3 ideographs:

1) facts - lookup on a manually created lexicon list of about 3000 items. Each item is tagged as useful, useless (stopword), numeric, punctuation and a few other codes. Longest match is performed when searching on this lexicon, and this results in breaking a sentence into smaller chunks of texts.

2) rules - common usage ad-hoc rules, also manually determined, are then employed to further split the chunks into short words of 2 or 3 characters. Examples of rules that serves to segment text are: a) any two adjacent similar characters XX; b) AX, where A is a small set of special characters; c) a remaining sequence of even number of characters are segmented two by two.

3) filtering - a first pass through the test corpus using steps 1 and 2, and frequency of occurrence is used to threshold the captured short words to extract the common words that are mainly domain-related.

4) iteration - expand the initial lexicon list in step 1) based on step 3) to about 15000 in our case and re-process corpus.

We are essentially segmenting texts using the most commonly used short-words, while other short-words are 'mined' from the collection. The result is 494,288 unique index terms.

2.2 TREC-5 Supplied Queries

Provided with the TREC-5 collection are 28 very long and rich queries (topics), mostly on current affairs, in English and Chinese versions. For this investigation we retain only the small 'title' section and examples are shown in Fig.1. This is done because in most realistic situations users would input only a few words as queries for retrieval. These Chinese queries reflect closely those of the English in content, and we consider them as approximations to 'expert translation' of the English. (In fact, they were composed by the same authors, not deliberate translations - - E. Voorhees, private communication). These are indexed like documents. Fig.1 also shows how our simple segmentation procedure segments it with all stopwords kept. Comparing with a manual segmentation based on short-word identification shows that we achieve 91.3% recall and 83% precision for the full queries, not great from a linguistic viewpoint but probably sufficient for IR. The average size of these queries is 6.0 terms, after stopword removal. Retrieval and evaluation using this version of the queries then provides us with results that may be considered as approximate 'upper-bound' for cross-lingual text retrieval.

2.3 Queries Based on Simple Dictionary Look-Up

We simulate the situation of an English-speaking user who has some rudimentary reading capability in Chinese attempting retrieval on Chinese documents. S/he composes queries in English and relies on an English-Chinese dictionary for translation. The English queries are those of TREC-5 discussed above, and we manually look up each content-bearing word on a small size (23000 words and phrases) pocket dictionary [Concise 1986]. A mechanical procedure is simulated: take at most three translations of each word, one from each of the first three senses; if there are less than 3 senses, synonyms are taken from the first, then the second until we use 3 translations or exhaust all definitions. If an adjacent pair of English words also appears in a definition entry under the first word, we will take the phrase translation. The procedure is naive and no attempt is made to resolve senses or ambiguities. We chose three translations simply as a compromise between effort and coverage. Using fewer than 3 will cause many missing translations, while using all translations of each word will not only require much more effort, but also often result in long queries with many inappropriate terms. If a machine readable English-Chinese dictionary is available, this procedure can be automated. Each of the translated Chinese word is separated by a blank to facilitate segmentation. Examples of this simple translation are also shown in Fig.1.

A major problem is with names such as Bosnia, Yugoslavia, etc. and acronyms such as WTO (World Trade Organization), APEC (Asian Pacific Economic Commission) or AIDS. For these, we assume the user would consult a different larger dictionary to obtain the translations, or spell out the abbreviations so that individual words can be mapped.

2.4 PIRCS Retrieval Engine

For retrieval, we use our PIRCS (acronym for Probabilistic Indexing and Retrieval - Components - System) engine that has been documented elsewhere [Kwok 1990,1995] and has participated in the past five TREC experiments with admirable results [e.g. Kwok & Grunfeld 1996]. It combines different probabilistic methods of retrieval that can account for local term frequencies within documents as well as global inverse collection term frequencies. Our strategy for ad-hoc retrieval involves two stages. The first is the initial retrieval where a raw query is used directly. The d best-ranked documents from this retrieval are then regarded as relevant without user judgment, and employed as feedback data to train the initial query term weights and to add new terms to the query - query expansion. This process has been called pseudo-feedback. This expanded

Retrieval:	Initial	Expanded Query
Total number of documents over all queries		
Retrieved:	28000	28000
Relevant:	2182	2182
Rel_ret:	1615	1709
Interpolated Recall - Precision Averages:		
at 0.00	0.5826	0.6599
at 0.10	0.4507	0.5774
at 0.20	0.4134	0.5151
at 0.30	0.3655	0.4697
at 0.40	0.3344	0.4249
at 0.50	0.3044	0.3969
at 0.60	0.2744	0.3558
at 0.70	0.2372	0.3212
at 0.80	0.2081	0.2805
at 0.90	0.1578	0.2250
at 1.00	0.0461	0.0842
Average precision (non-interpolated) over all rel docs		
	0.2930	0.3837
Precision:		
At 5 docs:	0.3786	0.5071
At 10 docs:	0.3893	0.5179
At 15 docs:	0.3762	0.4952
At 20 docs:	0.3607	0.4821
At 30 docs:	0.3512	0.4536
At 100 docs:	0.2504	0.3054

Table 1: Initial and Expanded Query Retrieval Results Based on 'Good Translation' of Queries

Retrieval:	Initial	Expanded Query
Total number of documents over all queries		
Retrieved:	28000	28000
Relevant:	2182	2182
Rel_ret:	991	1225
Interpolated Recall - Precision Averages:		
at 0.00	0.4364	0.4599
at 0.10	0.2246	0.2911
at 0.20	0.1674	0.2533
at 0.30	0.1379	0.2290
at 0.40	0.1199	0.2144
at 0.50	0.1050	0.2021
at 0.60	0.0833	0.1789
at 0.70	0.0666	0.1387
at 0.80	0.0395	0.1041
at 0.90	0.0318	0.0780
at 1.00	0.0197	0.0232
Average precision (non-interpolated) over all rel docs		
	0.1111	0.1819
Precision:		
At 5 docs:	0.2000	0.2643
At 10 docs:	0.1893	0.2643
At 15 docs:	0.1905	0.2595
At 20 docs:	0.1732	0.2482
At 30 docs:	0.1571	0.2274
At 100 docs:	0.1164	0.1654

Table 2: Initial and Expanded Query Retrieval Results Based on Dictionary Translation of Queries

query retrieval then provides the final result. This second retrieval in general can provide substantially better results than the initial if the initial retrieval is reasonable and has some relevants within the d best-ranked documents. The process is like having a dynamic thesaurus bringing in synonymous or related terms to enrich the raw query.

3 Results and Discussion

Table 1 shows the precision and recall table for the 'well-translated' queries, while Table 2 shows the same for the queries based on naive translation. These are the standard tables used for the TREC evaluation. Precision is defined as the proportion of retrieved documents which are relevant, and recall that of relevant documents which are retrieved. In general when more documents are retrieved, precision falls as recall increases.

The two columns in Table 1 using 'well-translated' queries corresponds to initial and expanded query retrieval. In our experiment, a retrieval returns 1000 ranked documents for each query, hence a total of 28,000 are 'Retrieved'. TREC assessors, based on pooled estimates, determines that 2,182 documents out of the collection are 'Relevant' to all 28 queries, while our engine returns 1615 (initial) and 1709 (ExpQry) of them within the 28,000, about 74% of the maximum for initial and 78.3% for ExpQry. The next rows contain precision values averaged over 28 queries and interpolated to eleven recall points from 0.0 to 1.0 in steps of 0.1. This gives users an idea of how the system behaves, especially at low recall or high recall regions. The non-interpolated precision calculates the average precision at every point where a relevant document is retrieved. Following that are the precision at various retrieved document cut-offs. For example, in Table 1, 'precision at 10 docs' of 0.3893 and 0.5179 means that out of the first 10 highest ranked documents, nearly 3.9 for initial retrieval and 5.2 for expanded query retrieval are relevant on average. 5.2 is quite reasonable for a mainly statistical system. For users, this is the most meaningful measure.

These values show that the second stage retrieval performs in general much better than initial first stage. We

use this second stage as our standard for comparison. These 'well-translated' queries return quite respectable results, considering the fact that they have on average only 6 characters. These values may serve as 'upper bound' retrieval effectiveness to this cross-lingual experiment.

Looking at Table 2 which employs naive translation, we see again that second stage retrieval is preferable to first stage. However, the effectiveness measures in all cases are much worse than those for Table 1, the 'well-translated' queries. For example, total relevants retrieved (1225 vs 1709) is 28.3% worse; precision at 10 retrieved documents (0.2643 vs 0.5179) is 49% worse, and average non-interpolated precision (0.1819 vs 0.3837) is 52.5% worse. Thus, the quality of translation has a major effect on retrieval and there is a lot of room for improvements.

The difficulty is well-known and arises mainly because in dictionary look-up, each English word can have several translations depending on how the word is used in context, and we did not use methods to help choose the correct one. Our simple procedure of picking the first three translations can lead to irrelevant as well as sometimes missing the correct translations. For example, 'world' was translated into Chinese words meaning 'earth', 'this life', and 'out of this world', and the usual meaning of 'all the land/countries' simply happens not to be in the right place. Employing all translations listed may lead to a long text that will dilute a query's intent. Using another dictionary may help in some cases but hinder in others. A dictionary that places all these senses in a strictly probability of usage order may provide better translation without context. Using context to help determine the sense of a word appears crucial. On the other hand, it is not easy to determine context when users compose queries of only a few words, which they often do.

This experiment may be seen as providing upper and lower bounds to English-Chinese CLTR using PIRCS' statistical retrieval engine. Using more sophisticated translation techniques, sizable decrements in effectiveness are also reported in English-Spanish CLTR [Davis & Dunning 95,96] and in English-French [Hull & Grefenstette 1996].

4 Acknowledgments

This work is partially supported by a Tipster grant from the U.S. Department of Defense. Xianlin Zhang and Jing Yan provided the dictionary look-up translation.

References

Concise English-Chinese Chinese-English Dictionary. Oxford University Press & The Commercial Press, 1986

Davis, M. (199x). New experiments in cross-language text retrieval at NMSU's Computer Research Lab. In: The Fifth Text REtrieval Conference (TREC-5). Harman, D.K. (ed.) to be published.

Davis, M. & Dunning, T. (1996). A TREC evaluation of query translation methods for multi-lingual text retrieval. In: The Fourth Text REtrieval Conference (TREC-4). Harman, D.K. (ed.) MIST Special Publication 500-236, 1996, pp.483-497.

The Fifth Text REtrieval Conference (TREC-5). Harman, D.K. (ed.). to be published.

Hull, D.A. & Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Proc. 19th Ann. Intl. ACM SIGIR Conf. on R&D in IR. Frei, H.P., Harman, D., Schauble, P. & Wilkinson, R (eds). pp.49-57.

Kwok, K.L. & Grunfeld, L. (1996). TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS. In: The Fourth Text REtrieval Conference (TREC-4). Harman, D.K. (ed.) MIST Special Publication 500-236, 1996, pp.145-152.

Kwok, K.L (1995). A network approach to probabilistic information retrieval. ACM Transactions on Office Information Systems, 13:325-353.

Kwok, K.L (1990). Experiments with a component theory of probabilistic information retrieval based on single terms as document components. ACM Transactions on Office Information Systems, 8:363-386.

Lin, C. & Chen, H. (1996). An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. IEEE Transactions on Systems, Man and Cybernetics, 26, pp.75-88.

Oard, D.W. & Dorr, B.J. (1996). A survey of multilingual text retrieval. Technical Report, University of Maryland. http://www.ee.umd.edu/medlab /mlir/mlir.html.

Sheridan, P. & Ballerini, J.P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In: Proc. 19th Ann. Intl. ACM SIGIR Conf. on R&D in IR. Frei, H.P., Harman, D., Schauble, P. & Wilkinson, R (eds). pp.58-65.

Query 001: U.S. to separate the most-favored-nation status from human rights issue in China.

Chinese Version: 美国决定将中国大陆的人权状况与其是否给予中共最惠国待遇分离

Segmentation result:
美国 | 决定 | 将 | 中国 | 大陆 | 的 | 人权 | 状况 | 与其 | 是否 | 给予 | 中共 | 最 | 惠国 | 待遇 | 分离 |

Translation by Dictionary Look-up:
美国 分离 分开 分手 最多的 最 大多数 好感 喜爱 赞成 民族 国家 地位 身份 人权 出版 问题 结果 中国

Query 002: Communist China's position on reunification.

Chinese Version: 中共对于中国统一的立场

Segmentation result:
中共 | 对于 | 中国 | 统一 | 的 | 立场 |

Translation by Dictionary Look-up:
共产主义者 共产主义的 中国 位置 阵势 姿势 统一 使一致

Query 003: The operational condition of nuclear power plants in China.

Chinese Version: 中共核电站之营运情况

Segmentation result:
中共 | 核电站 | 之 | 营运 | 情况 |

Translation by Dictionary Look-up:
操作的 业务的 可使用的 条件 状况 环境 原子核的 原子能的 能力 体力 力 植物 仪器 工厂 中国

Query 004: The newly discovered oil fields in China.

Chinese Version: 中国大陆新发现的油田

Segmentation result:
中国 | 大陆 | 新 | 发现 | 的 | 油田 |

Translation by Dictionary Look-up:
新近 最近 以新的方式 发现 油 田野 场地 矿田 中国

Fig.1: Examples of TREC-5 English Queries, Chinese Version, Segmentation Result and Chinese Version by Dictionary Look-up.