Effects of Term Segmentation on Chinese/English Cross-Language Information Retrieval

Douglas W. Oard and Jianqiang Wang
Digital Library Research Group
College of Library and Information Services
University of Maryland, College Park, MD 20742
{oard|wangjq}@glue.umd.edu

Abstract

The majority of recent Cross-Language Information Retrieval (CLIR) research has focused on European languages. CLIR problems that involve East Asian languages such as Chinese introduce additional challenges, because written Chinese texts lack boundaries between terms. This paper examines three Chinese segmentation techniques in combination with two variants of dictionary-based Chinese to English query translation. The results indicate that failure to segment terms, particularly technical terms and names, can have a cascading effect that reduces retrieval effectiveness. Task-tuned segmentation algorithms and alternative term weighting strategies are suggested as productive directions for future work.

1. Introduction

In Cross-language Information Retrieval (CLIR), the objective is to find documents written in one language (such as English) using queries that are expressed in another language (Chinese, for example). Fully automatic CLIR techniques for searching unrestricted texts typically extract term relationships from bilingual electronic dictionaries or bilingual text collections and use them to translate query and/or document representations into a compatible set of representations with a common feature set. Several types of terms can be used in information retrieval systems, including words, word roots, word stems, and character n-grams.

In this paper we focus on the retrieval of English documents using Chinese queries. The vast majority of recent CLIR research has focused on European languages [10]. Chinese/English CLIR introduces two additional challenges: a need to accommodate multiple character sets and a need to accommodate the lack of orthographic cues for term segmentation in Chinese. When lexical resources with good general coverage are available, the first problem is most acute for names and technical terms. Names and technical terms are usually highly selective when used as search terms, and so-called "natural language" systems that ranked retrieved documents in relevance order typically give high weight to such terms when they appear in query. CLIR systems in which both languages share a common character set typically retain unknown terms unchanged because names and technical terms (for which translations may not be known) are sometimes rendered the same way in both languages. When the character sets differ, transliteration would be needed to achieve the same effect. Knight and Graehl presented a technique for generating English "reverse transliterations" of Japanese terms that might be adapted to match English and Chinese terms [7], but we have not yet incorporated such a capability in our system. So in this paper we focus on the second problem: accommodating the lack of orthographic cues for term segmentation in Chinese.

2. Chinese Segmentation

Term segmentation is an important issue for CLIR with agglutinating languages such as Chinese. Written Chinese sentences lack explicit delimiters between terms, appearing as a linear sequence of equally spaced ideograph characters. Term segmentation, the process of taking a sequence of character strings and producing meaningful morphological units, has been widely studied because it is a prerequisite for many types of natural language processing (NLP) of Chinese texts [3]. Accurate segmentation is challenging because in many cases a Chinese character can be either a term by itself or part of a compound term. Furthermore, some Chinese terms can equally well be viewed as part of a still-larger compound term. This diversity creates the problem of *segmentation ambiguity*. Native Chinese speakers will, in fact, often disagree about the proper segmentation for a sentence.

Three general approaches to term segmentation for Chinese queries are possible: users could manually segment the query, the system could automatically segment it, or query segmentation could be avoided by indexing overlapping n-character sequences (character

n-grams) rather than words. For example, Bian and Chen used manual segmentation in their CLIR experiments [2]. Although manual segmentation of short user-entered queries may be practical in many applications, it does not scale well to cut-and-paste queries, query-by-example, and relevance feedback. N-grams offer the promise of broad coverage (perhaps with reduced retrieval performance), but we have not yet assembled translation resources that would be suitable for use with n-grams. So in this paper we have thus restricted our attention to automatic query segmentation.

2.1. One-best Segmentation (NMSU and LDC)

Many techniques have been proposed for automatic segmentation of Chinese text. Fundamentally, there are four possible sources of evidence about segmentation: lexical representations such as a list of known terms, algorithmic knowledge such as a heuristic preference for the longest substring, statistical evidence acquired from representative collections of text, and the user. Each source of evidence brings advantages and disadvantages, and practical segmentation schemes typically exploit multiple sources of evidence. For example, the simplest commonly implemented approach uses a greedy left-to-right search for the longest matching substring in a term list. The New Mexico State University (NMSU) segmentation software that we used in our experiments applied a variant on this approach, with a more thorough search of the space of alternative segmentations than the greedy algorithm and special processing to recognize Chinese names.[1] Segmentation algorithms that depend on manually coded lexical information generally fail when they encounter unknown terms, however. Statistical evidence can help overcome this problem, and can also help to improve the selection among alternative possible segmentations. The segmenter that we obtained from the Linguistic Data Consortium (LDC) incorporated this second idea, using dynamic programming to search for the most likely segmentation based on the product of the frequencies of the segmented terms.[2] This method is simple, but its effectiveness naturally depends on the degree to which the statistics on which segmentation decisions are based are representative of the texts that are presented for segmentation.

2.2. Exhaustive Segmentation (EXH)

One-best segmentation strategies such as those implemented in the NMSU and LDC segmenters might not be optimal for information retrieval applications, because longer Chinese query terms might contain meaningful substrings appear alone in relevant documents. For example, when a query containing 乙 肝 病 毒 (hepatitis B virus) is issued, several meaningful included terms might be useful in the query (e.g., 乙 肝 (hepatitis B), 乙 肝 病 (hepatitis B disease), and 病毒 (virus)). Of course, this could sometimes introduce inappropriate query terms as well. Full-text information retrieval systems are, however, known to be remarkably tolerant of ambiguity when relatively long queries are presented. Long queries, which are typical in cases where automatic segmentation would be needed, would be expected to provide sufficient context for co-occurrence relationships within the documents being searched to favor the correct terms over the incorrect ones. Sanderson saw a similar effect with conflated terms, for example [11].

Based on this insight, we chose to also explore a task-tuned segmentation strategy that we call exhaustive segmentation. In exhaustive segmentation, every substring for which a translation is known is extracted from the query. This represents the opposite extreme from the one-best segmentation implemented in the NMSU and LDC segmenters. The Chinese term list that we used as a basis for exhaustive segmentation was the Chinese half of the bilingual term list (described below) that we subsequently used to perform query translation. We performed exhaustive segmentation with a Perl script that implemented the following simple algorithm:

- Create a hash table of all Chinese dictionary entries
- Set *k* to the maximum length of any dictionary entry
- Given an unsegmented input text of *n* two-byte Chinese characters, for each text position from 1 to *n-k* do
    - For each string (starting at the current text position) with a length varying between 1 and *k*, search the hash table for that string
        - If the string is found in the hash table, add it to the output text

Assuming that a hash table search can be performed in constant time, for an input text of length *k* and a dictionary with maximum string length of *k*, the time complexity of this algorithm is *O*(n^k). This could clearly be improved by using a more appropriate data structure that would eliminate repeated rechecking of the same substring, but we found this simple algorithm to be adequate for our purpose because both n and k are relatively small in our experiments.

[1] Available at http://crl.nmsu.edu/software/

[2] Available at http://morph.ldc.upenn.edu/Projects/Chinese/

3. Query Translation

CLIR is more complex than traditional information retrieval because some method for query or document translation must generally be applied before term matching and document-ranking algorithms can be invoked. Query translation essentially transforms the CLIR problem into a monolingual information retrieval problem for which useful solutions already exist, so it has proven to be a popular approach. One commonly used query translation approach, known as Dictionary-based Query Translation (DQT) replaces each query term with one or more translations that are automatically extracted from a bilingual term list built from an online bilingual dictionary (cf., [1, 6]). We used a bilingual term list constructed from the "Optilex" bilingual dictionary that was developed by the Chinese-English Translation Assistance (CETA) group.[3] Our term list contains 177,063 bilingual pairs in which each pair consists of one term in Chinese and the corresponding word or phrase in English. It is quite common for single-character Chinese terms to have several translations, some with very different meanings. The number of unique Chinese terms in our bilingual term list is thus far smaller than 177,063 – perhaps around 60,000. When multiple translations are known for a single Chinese term, the bilingual term pairs in our term list sorted in a weak predominance order that seeks to put the most common translation first. The Optilex dictionary was constructed from many smaller dictionaries, and the resulting predominance order is sometimes noticeably incorrect. We have previously explored six DQT techniques that together explore the effects of winner-take-all, word-match and stem-match approaches [9], and we have chosen two of the techniques for this evaluation:

- **First Translation (FT).** Choose the first match in the bilingual term list. Terms that are not found in the list are ignored.
- **Every Translation (ET).** Choose every match in the bilingual term list. Terms that are not found in the list are ignored.

In either case, we replace each Chinese term in the query with the corresponding English term(s) from matching bilingual pair(s) to produce a version of the query that is expressed in English. In addition to simple mappings from Chinese terms to English words, term-to-phrase mappings are possible (and, in fact, common). So translated queries sometimes contain repeated words. Furthermore, translated queries could contain multiple words with the same stems. Such words would be treated by our English information retrieval system as if they too were identical.

In our initial experimental runs we discovered that each occurrence of a few Chinese terms generated many English words that had little relationship to the query. Closer inspection revealed that these terms were definitions of Chinese particles that we should have treated as stopwords. We minimized the effect of this problem by deleting all translations that contained more than three English words from the bilingual term list.

In language pairs for which Machine Translation (MT) systems exist, CLIR applications could leverage the investment in those systems by using them to translate either each query or all of the documents. As we were completing our experiments, we obtained the SYSTRAN Professional 2.0 Chinese to English machine translation system. We were thus able to explore MT-based Query Translation (MQT) as well. SYSTRAN includes a proprietary segmentation algorithm, so none of our other three segmenters were needed in this case.

4. Experiment

The document collection used in our experiments was the *Financial Times* collection from TREC disk 4. It contains 210,158 English articles from the *Financial Times* newspaper in the United Kingdom that were generated between 1991 and 1994. The topics used in the experiment were TREC topics 351-400, which are English language topics. The documents, topics, and relevance judgement are available from the National Institute of Standards and Technology (NIST). The title and description fields of each topic were translated into Chinese manually by a native Chinese speaker. Translation of 50 topics required approximately 3 hours, including data entry. Each test query was formed automatically from the entire translated title and description fields of the associated topic. No relevant documents are known in the *Financial Times* collection for two of the topics (358 and 379), so the retrieval effectiveness results reported below are based on 48 queries.

Version 3.1pl of the Inquery information retrieval system from the University of Massachusetts was run on a SPARC 20 to index and retrieve the English documents. The Inquery "kstem" stemmer and the standard English Inquery stopword list were used. We ran the eight experiments shown in Table 1. Monolingual retrieval offers some insight into the best performance that a CLIR system might be expected to achieve, so we included that as a baseline condition using the title and description fields from the original English query.

[3] Available from MRM Corporation, Kensington, MD USA

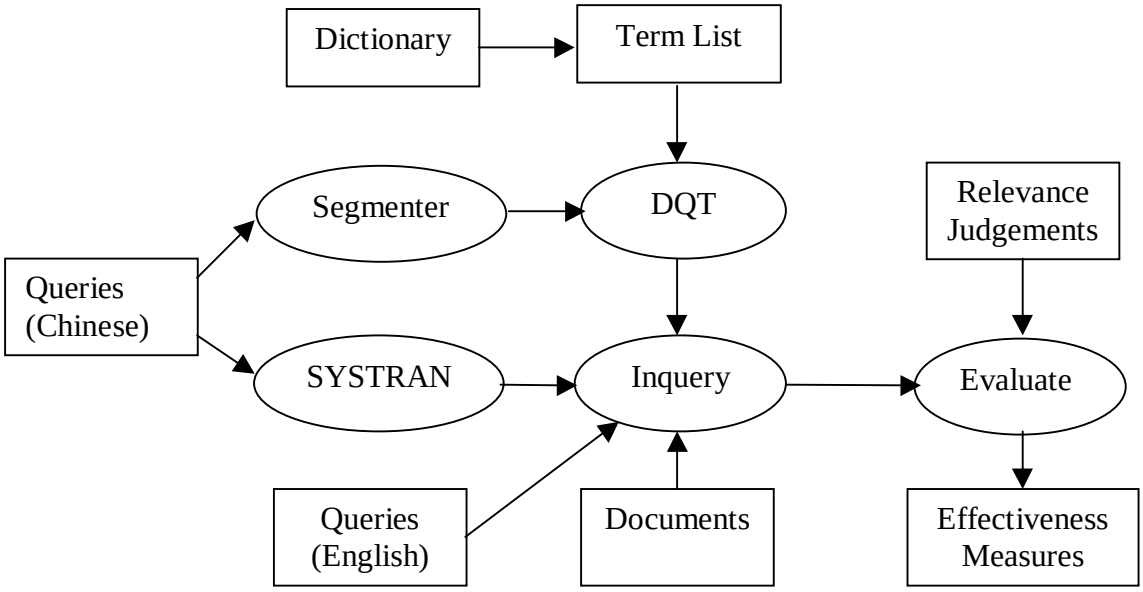

Figure 1. Experiment design.

Run name	**EXHET**	**EXHFT**	**LDCET**	**NMSUET**	**LDCFT**	**NMSUFT**	**SYSTR**	**MONO**
Segmention	Exhaustive	Exhaustive	LDC	NMSU	LDC	NMSU	SYSTRAN	N/A
Translation	DQT-ET	DQT-FT	DQT-ET	DQT-ET	DQT-FT	DQT-FT	SYSTRAN	Mono-lingual
Segment	7 sec	7 sec	3.6 sec	2 sec	3.6s	2sec	N/A	N/A
Translate	40 sec	30 sec	28 sec	26 sec	12 sec	12 sec	0.3 sec	N/A
Retrieve	63 sec	31 sec	26 sec	25 sec	10 sec	11.6 sec	5.4 sec	10 sec
Avg. Prec.	0.0212	0.0346	0.0422	0.0455	0.0470	0.0492	0.0891	0.1805
Std. Dev.	0.0441	0.0819	0.0780	0.0823	0.0845	0.0907	0.1729	0.2109
% MONO	12%	19%	23%	25%	26%	27%	49%	100%
Prec@10	0.0521	0.0625	0.0833	0.0896	0.1000	0.0979	0.1146	0.2417

Table 1. Retrieval effectiveness (avg. over 48 queries) and processing time (avg. over 50 queries).

	EXHET	**EXHFT**	**LDCET**	**NMSUET**	**LDCFT**	**NMSUFT**	**SYSTR**
EXHFT	.286						
LDCET	.060	.651					
NMSUET	**.037**	.489	.552				
LDCFT	.062	.314	.613	.859			
NMSUFT	.061	.489	.522	.722	.559		
SYSTR	**.009**	.054	.054	.074	.074	.107	
MONO	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.001**

Table 2. Paired *t*-test significance values for average precision (48 trials, **bold** significant at 0.05).

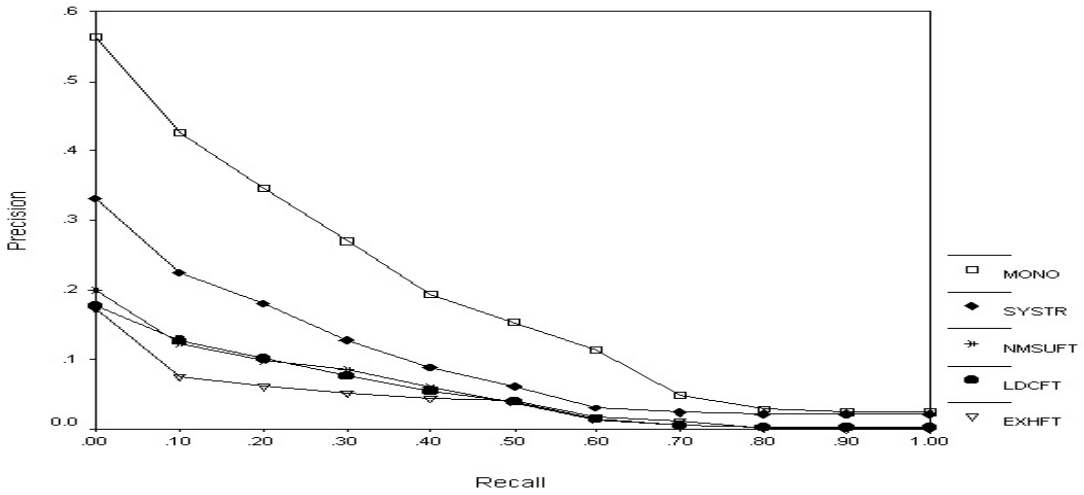

Figure 2. Comparison of DQT-FT, SYSTRAN, and monolingual retrieval.

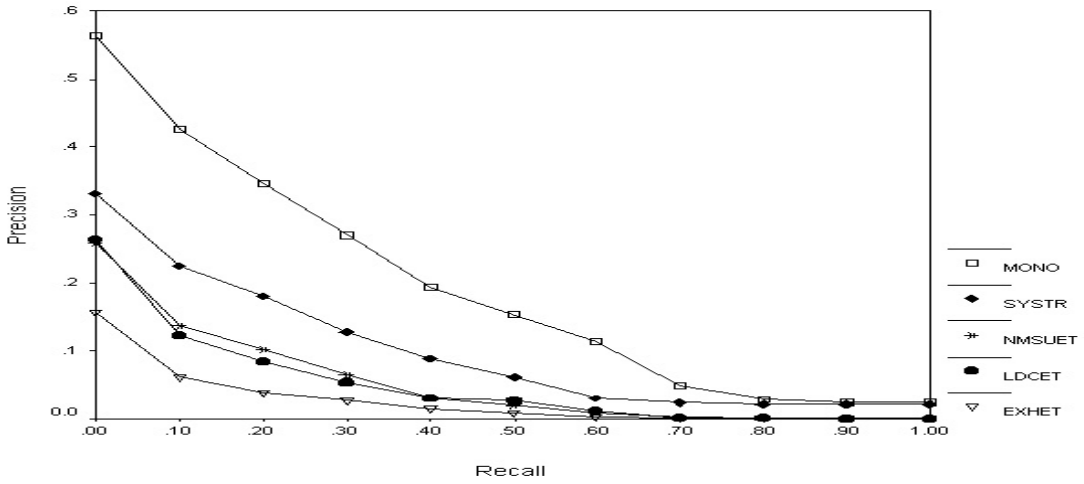

Figure 3. Comparison of DQT-ET, SYSTRAN, and monolingual retrieval.

5. Results

Average precision is widely used as a measure of effectiveness for ranked retrieval systems, and simple DQT techniques of the type we have implemented generally achieve between 40% and 60% of the average precision of a corresponding monolingual system [10]. We were thus surprised to obtain only 27% of monolingual average precision using our best system (NMSUFT, see Table 1). Since segmentation is the new factor in these experiments, we explored the results further in order to understand the nature of the interaction between our segmentation, translation, and retrieval techniques

5.1 Statistical Significance Testing

Table 2 shows the results of *t*-tests between each pair of techniques. The paired *t*-test treats the 48 queries as random samples from a query population. We chose the (11-point) average precision for each query as the dependent variable as recommended by Hull [5], and treated the CLIR technique as the independent variable. Under these conditions, the null hypothesis would be that two CLIR techniques produces values for average precision that are drawn from the same distribution. We set our significance level as 0.05, a commonly used value. Obtaining significance values below 0.05 would indicate that the

retrieval effectiveness values were unlikely to have been drawn from the same distribution. From this we could conclude that the observed difference in the average precision values would be sufficient to claim that a real difference in the retrieval effectiveness of the measures resulted when one CLIR method was used in place of another.

From the results in Table 2 we can draw the following conclusions:

- NMSU segmentation is significantly better than exhaustive segmentation when every translation is used;
- SYSTRAN is significantly better than exhaustive segmentation when every translation is used;
- Monolingual IR is significantly better than all of the CLIR techniques that we tried.

However, we do **not** have sufficient evidence to demonstrate that:

- Using the first translation would be significantly different from using every translation;
- NMSU segmentation would be significantly different from LDC segmentation;
- Using SYSTRAN would be significantly better than using either the NMSU or LDC segmenters with dictionary-based query translation.

The relatively small significance values for comparisons between SYSTRAN and the other methods suggest that with more samples it might be possible to demonstrate that SYSTRAN is actually outperforming every CLIR technique that we have implemented (see Figures 2 and 3).

5.2 Failure Analysis

Inspection of the segmented queries used in the experiment indicates that some query terms, particularly some technical terms, are not segmented correctly. The following is a list of some important query terms (terms that, if removed, would make accurate retrieval unlikely) that each segmenter failed to segment correctly.

- NMSU segmenter: 后更年期 (postmenopausal), 氰化物 (cyanide), 厌食症 (nervosa), 易饿症 (bulimia), 欧元 (Euro), 安乐死 (mercy killing), 军工厂 (arsenal)
- LDC segmenter: 福克兰 (Falkland), 人质 (hostage), 后更年期 (postmenopausal), 埃尔尼诺 (El Nino), 氰化物 (cyanide), 狂犬病 (rabies), 厌食症 (nervosa), 易饿症 (bulimia), 国际法庭 (international court), 欧元 (Euro), 安乐死 (mercy killing), 军工厂 (arsenal), 亚马逊 (Amazon)
- Exhaustive segmentation: 福克兰 (Falkland), 血液 (blood), 后更年期 (postmenopausal), 埃尔尼诺 (El Nino), 厌食症 (nervosa), 易饿症 (bulimia), 诺贝尔 (Nobel), 欧元 (Euro), 并发症 (syndrome), 军工厂 (arsenal), 安乐死(mercy killing)，亚马逊 (Amazon).

These are either technical terms or proper names. Failure to segment a Chinese term correctly is not merely a matter of missing important query terms; more seriously it produces the wrong query terms. When failing to correctly segment a term, one-best Chinese segmenters typically produce several single characters. For example, for the term 福克兰 (Falkland), the LDC segmenter produced three single characters: 福, 克, and 兰. Most single characters are common terms in Chinese (for example, 福, 克, and 兰 individually are all valid Chinese terms). Sometimes these common terms have relatively rare translations, though. For example, *orchid*, one of the translations of 兰, is a relatively rare English word. Inquery (and most other ranked retrieval systems) favors matches on rare terms over matches on common words because the rare terms are highly selective. As a result, CLIR effectiveness can be severely degraded. Segmentation failures thus have a cascading effect through translation step to produce adverse effects on retrieval effectiveness that greatly exceed that which would be seen in monolingual applications.

There are two general causes of segmentation errors. The first is dictionary coverage. Technical terms and proper names, such as 狂犬病 (rabies) and 亚马逊 (Amazon), may be missing from the segmentation dictionary. New terms, such as 欧元 (Euro) pose an additional challenge in this regard since electronic dictionaries typically lag behind the creation of new terms. The other cause of difficulties is segmentation ambiguity. For example, 国际法庭 (international court) includes the terms 国际法 (international laws), 国际 (international), and 法庭 (court). The impact of this problem might be minimized by incorporating more context information into the segmentation algorithm, but there are undoubtedly practical limits to how far we can productively proceed in that direction.

5.3 One-best vs. Exhaustive Segmentation

Although relaxing the requirement for one-best segmentation might be a good idea, the precision vs. recall graphs in Figures 2 and 3 make it clear that exhaustive segmentation goes too far in that direction. The paired *t*-tests in Table 2 show that the NMSU segmenter is significantly better than exhaustive segmentation when DQT-ET is used, and inspection of

the query-by-query results in Figure 4 show few cases in which exhaustive segmentation is of any help when DQT-FT is used. When compared with either one-best segmenter, exhaustive segmentation produced many more unwanted single-character terms. This simply exacerbated the cascading error problem described above. Some simple modification to our exhaustive segmentation algorithm (e.g., eliminating all single-character terms) might result in improved retrieval effectiveness, but we have not yet had time to explore that possibility.

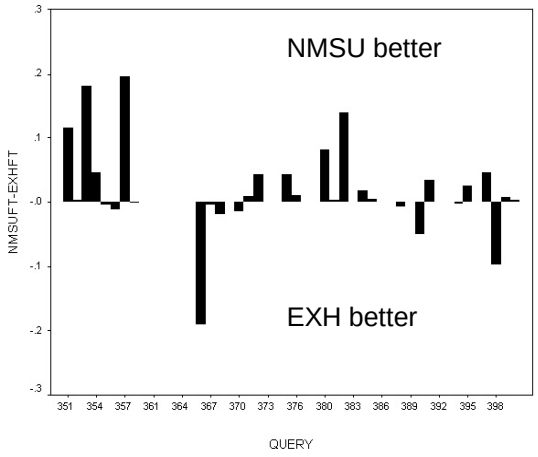

Figure 4. Average precision difference between NMSU and exhaustive segmenters with DQT-FT.

We did not make any effort to align the lexical resources used in these experiments. In particular the LDC and NMSU segmenters incorporated Chinese term lists that contained terms for which no translation was known. Thus, some terms were correctly segmented but DQT then failed to translate them. A richer bilingual term list would certainly be desirable, but it would not be advisable to remove terms from the segmenter's term list simply because no translation is known. As we described above, segmentation failures can cause cascading errors, and no benefit would accrue from introducing additional segmentation errors.

5.4 NMSU vs. LDC Segmenters

Overall, the two one-best segmenters that we tried achieved similar retrieval effectiveness. There was little separation between the average precision achieved by each under comparable conditions (e.g., 0.0492 for NMSU and 0.0470 for LDC with DQT-FT), and the query-by-query comparison in Figure 5 shows

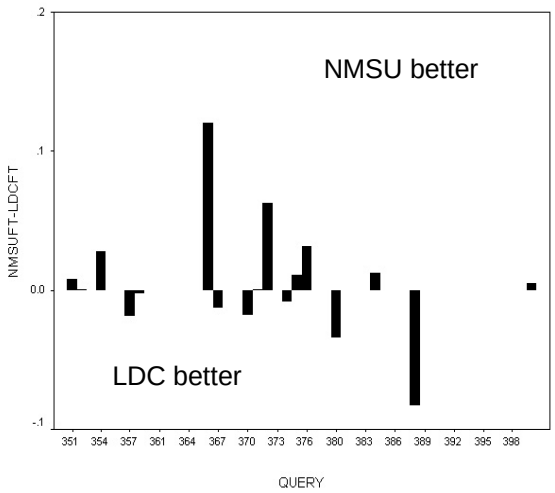

Figure 5. Average precision difference between NMSU and LDC segmenters with DQT-FT.

that each has an advantage over the other on some queries. This suggests that it might be worth exploring merging strategies that could leverage the strengths of each.

5.5 First Translation vs. Every Translation

Overall, DQT-FT achieves retrieval effectiveness that is comparable to that achieved by DQT-ET (see Figure 6). This is consistent with the results we have seen in previous experiments between English and German [9]. The Optilex dictionary that we used is sorted in an order that approximates the predominance in common usage, but we made no effort to tune this ordering to our application. Further attention to this might improve the performance of DQT-FT. DQT-FT and DQT-ET are, of course, extremes on a spectrum of options, and exploring top-n DQT techniques might also be productive.

Averaged effectiveness measures can serve as a useful source of insight about the utility of a retrieval system, but other factors are important as well. As Table 1 shows, DQT-ET is somewhat slower than DQT-FT. This occurs because the time required for query processing in information retrieval systems typically grows roughly linearly with the number of terms in the query. Consistent behavior is also an important issue, but we detected no significant difference in cross-query variability in our experiment. The standard deviations in Table 1 do in fact show a slight trend towards greater variance from DQT-FT, but the amount is more than would be expected given the slightly larger average precision values achieved by

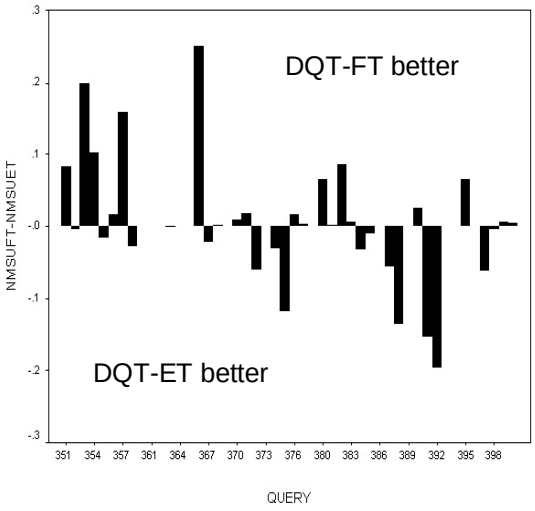

Figure 6. Average precision differences between DQT-FT and DQT-ET with the NMSU segmenter.

DQT-FT. On balance, DQT-FT thus seems to be the better of the two choices, with better efficiency and with effectiveness and consistency not appreciably different from DQT-ET.

5.6 MT-based Query Translation vs. DQT

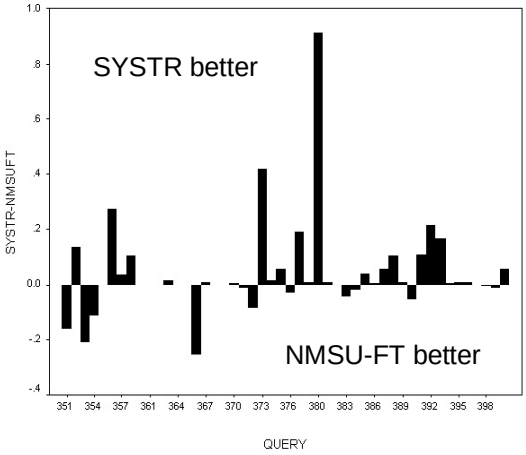

Figure 7. Average precision differences between SYSTRAN and NMSU segmenter with DQT-FT

Our results indicate that MT-based query translation is at least as effective as any other CLIR method that we studied, and sometimes it is significantly better (see Figure 7). Inspection of the translated queries reveals that SYSTRAN successfully segmented (and translated) some technical terms that all of the other three segmenters handled incorrectly. Furthermore, SYSTRAN applies more sophisticated linguistic processing than our simple DQT techniques. This suggests that incorporating additional linguistic knowledge, for example constraining the set of possible translations for a term using part-of-speech information, could produce improvements in Chinese/English CLIR effectiveness similar to the substantial gains reported for CLIR between European languages [4].

6. Discussion and future work

We have examined Chinese/English cross-language information retrieval using three segmentation techniques and three query translation techniques. Our results reveal that term segmentation errors can have an effect on retrieval effectiveness that is of the same magnitude as the effects typically seen from errors in term translation. Failure to correctly segment technical terms and proper names seems to be the direct cause of this effect. This, in turn, reflects the heavy reliance of the Chinese segmenters that we used on lexical knowledge that is encoded in (unavoidably incomplete) lists of Chinese terms. The cascading effect of incorrect segmentation, erroneous translation, and reliance on rare terms in ranked retrieval systems yields at best 27% of monolingual retrieval effectiveness with the architecture that we have used..

Among the three segmenters we studied, the two one-best segmenters (NMSU and LDC) perform at a comparable level, while exhaustive segmentation does not seem to be as good. The major reason for this appears to be that exhaustive segmentation produces too many single-character terms, so the simple expedient of eliminating all of the single-character terms might be useful in this case.

More fundamentally, we believe that it is important to explore approaches to Chinese segmentation that are tuned to the requirements of cross-language retrieval applications. Just as English phrases may not be helpful unless their constituent words are also indexed, indexing only the longest recognized Chinese terms may not be as good as indexing meaningful constituent parts of those terms as well. Exhaustive segmentation is a first crude step in this direction, but more sophisticated techniques have been proposed by Song and we believe that it would be productive to explore them [12].

We also found that MT-based query translation appeared to outperform all of our present dictionary-based techniques. The observed differences in query processing time may not be too important since our dictionary-based techniques are not optimized for speed, but the differences in retrieval effectiveness deserve our attention. The use of shallow linguistic

processing such as part-of-speech information might be helpful, and it would also be interesting to investigate the middle ground between taking all known translations and using only one. We expect that the lessons that we are learning will apply to some degree to any agglutinating language There are presently no useful machine translation systems for several important language pairs that include at least one such language, so reliance on an existing MT system will not always be possible.

Perhaps the most important focus for further work is neither segmentation nor translation, but rather retrieval. In our experiments we used term weighting strategies and retrieval algorithms that were optimized for queries expressed in the same language as the documents, but we presented those systems with queries obtained through segmentation and translation. Rethinking the term weighting strategy and designing retrieval algorithms that exploit the structure induced by the translation process may ultimately offer the best way to interrupt the cascading errors that we experienced in these experiments. One simple step in this direction would be to assign term weights before translation and then map those weights appropriately into English, perhaps using a vector translation strategy such as that we have used in earlier experiments between English and Spanish [8].

The techniques that we have explored in this paper offer the potential to expand the range of practical cross-language information retrieval applications by enabling query-by-example and relevance feedback with agglutinating languages. Such languages are used by a significant fraction of the Earth's population, so fully developing this capability will ultimately move us one step closer to the dear of a truly global information infrastructure.

Acknowledgments. The authors are grateful to Dekang Lin for suggesting exhaustive segmentation, to Ruth Sperer for implementing it, to Philip Resnik and Gina Levow for help with the bilingual term list and the term substitution software, to New Mexico State University and the Linguistic Data Consortium for the use of their segmenters, and to Laurie Gerber for the use of SYSTRAN. We also wish to thank Mun Kew Leong, Song Rou, Huang Changning, Guo Zhili, Bob Allen and the anonymous reviewers for their comments on earlier presentations of the ideas in this paper. This work was supported in part by DARPA contract N6600197C8540.

References

[1] Lisa Ballesteros and W Bruce Croft, "Phrase Translation and Query Expansion Techniques for Cross-language Information Retrieval", in *Proceeding of the 20th International ACM SIGIR Conference on Research ad Development in Information Retrieva*l, July 1997.

[2] Guo-wei Bian and Hsin-hsi Chen, "Integrating Query Translation and Document Translation in a Cross-language Information System", in David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Third Conference of the Association for Machine Translation in the Americas*, Springer, October 1998, pp. 250-265.

[3] Lee-Feng Chien and Hsiao-Tieh Pu, "Important Issues on Chinese Information Retrieval", *Computational Linguistics and Chinese Language Processing*. Vol. 1, No. 1, August 1996, pp. 205-221.

[4] Mark W. Davis and William C. Ogden, "QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System", in *Proceeding of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, pp. 92-97.

[5] David A. Hull, "Using Statistical Testing in the Evaluation of Retrieval Performance", in *Proc. of the 16th ACM/SIGIR Conference*, 1993, pp. 329-338.

[6] David A. Hull and Gregory Grefenstette, "Querying across Languages: A Dictionary-based Approach to Multilingual Information Retrieval", in *Proceeding of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[7] Kevin Knight and Johnathan Graehl, "Machine Transliteration", *in Seventeenth International Conference of the Association for Computational Linguistics*, 1997.

[8] Douglas W. Oard, "Adaptative Filtering of Multilingual Document Streams", *in Conference Proceeding of RIAO'99*, Montreal, 1997, pp. 233-254.

[9] Douglas W. Oard, "A Comparative Study of Query and Document Translation for Cross-language Information Retrieval", in *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, October 1998.

[10] Douglas W Oard and Anne Diekema, "Cross-language Information Retrieval", *In Annual Review of Information Science and Technology*, volume 33, American Society for Information Science, 1998.

[11] Mark Sanderson, "Word Sense Disambiguation and Information Retrieval", in W Bruce Croft and C. J. van Rijsbergen, editors, Proceedings of *the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Springer-Verlag, July 1994, pp. 142-151.

[12] R. Song, C.J. Qiu, L.G. Ou Yang, X. Wang, "Bi-Orderly-Neighborship and its Application to Chinese Word Segmentation and Proof, *ICCC'96*, National University of Singapore, Singapore, June 1996, pp. 428-433 (in Chinese).