Structured Translation for Cross-Language Information Retrieval

Ruth Sperer*
eMotion, Inc.
2600 Park Tower Drive, Vienna, VA 22180
ruth.sperer@emotion.com

Douglas W. Oard
College of Library and Information Services
and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
oard@glue.umd.edu

ABSTRACT

The paper introduces a query translation model that reflects the structure of the cross-language information retrieval task. The model is based on a structured bilingual dictionary in which the translations of each term are clustered into groups with distinct meanings. Query translation is modeled as a two-stage process, with the system first determining the intended meaning of a query term and then selecting translations appropriate to that meaning that might appear in the document collection. An implementation of structured translation based on automatic dictionary clustering is described and evaluated by using Chinese queries to retrieve English documents. Structured translation achieved an average precision that was statistically indistinguishable from Pirkola's technique for very short queries, but Pirkola's technique outperformed structured translation on long queries. The paper concludes with some observations on future work to improve retrieval effectiveness and on other potential uses of structured translation in interactive cross-language retrieval applications.

1. INTRODUCTION

Cross-language Information Retrieval (CLIR) is the task of finding documents that are written in one language (e.g., English) using queries that are expressed in another (e.g., Chinese). One common approach, known as "query translation," is to translate each query term and then perform monolingual retrieval in the language of the document [11]. Bilingual dictionaries have proven to be a useful source for potential translations, and methods for selecting the appropriate translation(s) for each term have been the subject of extensive research. Two fundamental approaches to this problem of translation selection have emerged, techniques guided by encoded linguistic knowledge and/or statistics that characterize term usage in large collections of text (cf., [8, 9]) and techniques that treat each known translation as an equally valid alternative [14]. In this paper we introduce structured translation, a framework for query translation in which we assume that the dictionary employed has the target language translations grouped into distinct concepts. We use the first approach to select among alternative concepts that a query-language term might represent and then apply the second approach to accommodate the range of document-language terms that might be chosen to express that concept. We present a technique that transforms an unstructured bilingual dictionary into a structured one, and experimental results obtained using that technique.

*Work by the first author was performed while at the University of Maryland Computer Science Department.

2. STRUCTURED TRANSLATION

Structured translation offers four possible advantages:

- It models the query translation process with greater fidelity than unstructured translation. By exploiting the greater fidelity we may be able to improve retrieval effectiveness.
- The use of translation clusters rather than individual translations can reduce the number of comparisons when using local context to select the correct translation, thus decreasing the time required for query translation.
- With structured translation it becomes possible to display alternate gloss translations that illustrate different meanings, possibly improving the user's ability to recognize the most desirable documents in the retrieved set.
- Structured translation could facilitate interactive query refinement by allowing alternative translations of each term to be grouped meaningfully.

Our principal goal in this paper is to explore the first point. The second point is addressed in section 3.3, and the remaining points are discussed briefly in Section 6.

2.1 Vocabulary Selection

Figure 1 compares the vocabulary selection process for monolingual and cross-language free-text retrieval systems that use the so-called "bag-of-words" approach. Reduced to its simplest form, in monolingual retrieval users attempt to choose query terms that will be present in documents that the user wishes to see. Terms are present in documents because authors put them there, and authors are free to choose any terms that express the concepts they wish to convey. For monolingual searchers, the vocabulary selection problem is essentially the searcher's effort to guess which terms an author might have used to express the concepts that are of interest [1]. Searchers can employ a range of strategies for this "vocabulary discovery" task, including reliance on prior knowledge, use of a thesaurus, and analysis of search results (cf., [17]). The terms chosen by the searcher and the author are then compared to compute a score (belief value) for each document with respect to the query, and the resulting belief values are sorted to produce a ranked list in which the most promising documents will (hopefully) appear near the top.

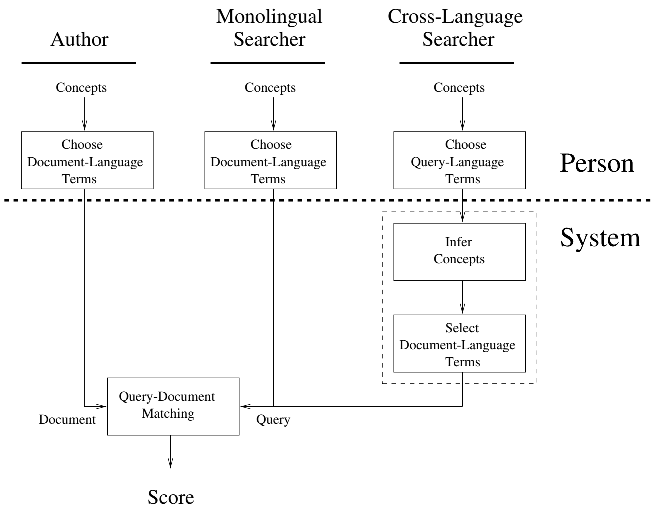

Figure 1: The vocabulary selection process for free-text retrieval. The dashed box encloses the structured translation model.

Such a direct approach is not possible in cross-language retrieval, however. In the cross-language case, we assume that the searcher wishes to choose query terms in a language different from that which the author has used. Under these conditions, the best that a searcher could hope to do is choose query-language terms that have the same *meaning* as the terms that the author of a desirable document might have chosen. The system must then select document-language terms that express those concepts. Critically, in the cross-language case, the responsibility for selecting the *same* term(s) that the author would use to represent each concept rests with the system rather than with the user.[1] The processes by which concepts are inferred and the process by which terms are selected to represent those concepts are the key points that distinguish cross-language retrieval from its monolingual counterpart.

2.2 Improving Translation Effectiveness

As shown by the dashed box in Figure 1, all present approaches to cross-language retrieval with which the authors are familiar group the two stages together, treating the problem as one of translating from terms in one language to terms in another. Several researchers have achieved good results by modeling query translation as a statistical process (cf. [2, 9]. The basic approach is to construct a probability mass function over the document-language terms that is conditioned on the term to be translated and proximate terms found within some local context of the term to be translated. Pirkola recently introduced an alternative technique in which every known translation is treated as if it were synonymous for the purposes of retrieval [14]. The basic idea in this case is that the occurrence of any known translation is treated as an instance of the query-language term that was specified by the user.

The query translation model depicted in Figure 1 offers an attractive way of combining these apparently contradictory approaches. When a user chooses a term that has multiple meanings, it is reasonable to assume that only one of the possible meanings is intended. Probabilistic techniques offer a principled way to express the system's degree of certainty about which meaning was intended. But even if the intended meaning were known with certainty, a different perspective may be needed to choose the document-language term(s) that an author might use to express that concept. Pirkola's approach seems appropriate after the concept has been determined, reflecting the intuition that in the absence of evidence about how authors choose terms it is reasonable to assume that they could choose *any* appropriate term for the concept that they wish to represent. By separating concept inference from translation selection. structured translation seeks to model the term selection problem with greater fidelity than previous methods.

If we are to treat concept selection and term choice separately, each concept must be defined in a way that facilitates generation of the terms that can be used to express that concept. In WordNet, sets of synonymous terms are used to represent concepts [3]. The same idea can be applied directly to bilingual dictionaries, grouping translations with similar meanings into clusters. Each cluster then represents one possible meaning of the query-language term, and the document-language terms that can be used to express that concept are the terms that make up the cluster. Dictionaries with such a structure may be available,[2] and Section 3.2 presents

[1] In monolingual retrieval, automatic query expansion techniques seek to achieve a similar effect.

[2] Sakhr's Arabic/English CLIR system is one example

an automated technique for converting an unstructured (term-to-term) translation dictionary into a structured dictionary. Once a structured translation dictionary is available, any desired technique can be used to perform the concept inference process (translation from term to cluster). The terms within each cluster can then be treated as synonyms, or an attempt can be made to build source-specific or genre-specific probabilistic models of the way in which terms are used in the document collection. In Section 4 we explore three alternative ways of performing concept inference in conjunction with the synonym-based approach to term selection.

3. STRUCTURED DICTIONARIES

In this section we approach the task of building structured dictionaries as one of agglomerative clustering of alternate translations in an existing bilingual term list. We begin by describing the dictionary clustering task, and then summarize related work on word sense disambiguation. Finally, we present an automatic technique for clustering bilingual dictionaries in which the target language is English.

Figure 2 shows the entry for a single term in the bilingual term list that we used for the experiments reported in Section 4. Six translations are known for *deng1lu4*.[3] Three have very similar meaning (login, log-in, and logon); the other three are clearly related in some way to the concept of beginning a session on a computer, but their meanings are both more general and somewhat different from each other. It would be easy to argue for any of the groupings in that figure.

record	entry	register	login	log-in	logon
(record)	(entry)	(register)	(login, log-in, logon)		
(record, register)		(entry)	(login, log-in, logon)		
(record, entry)		(register)	(login, log-in, logon)		
(record, entry, register)			(login, log-in, logon)		
(record, entry, register, login, log-in, logon)					

Figure 2: Dictionary clustering for deng1lu4.

This is clearly a clustering problem, for which the key issues are definition of a function that expresses the degree of similarity among a cluster of terms and choice of a strategy to search the space of possible clusters. For the moment we focus on similarity measures—our complete clustering technique is presented at the end of this section. For the purposes of CLIR, it seems clear that the appropriate basis for constructing a similarity function is the differential effect on retrieval if both terms were considered to represent the same concept. In other words, given the rank order produced through the use of one translation, what would be the effect of treating the other word as part of the same cluster? The key test is whether a searcher who knew the document language might have thought the second translation would be a good choice to add to a search that already included the first one. Such a test would be useful if we were to manually cluster a dictionary, but we must turn to other sources of evidence if the process is to be performed automatically.

3.1 Prior Art

Term similarity has been the subject of extensive research, principally in the context of word sense disambiguation [4]. Similarity measures have been developed that exploit both knowledge-based and corpus-based sources of evidence. Knowledge can be encoded algorithmically or in some form of knowledge structure. For example:

- Semantic relationships. Ontologies such as thesauri and semantic networks encode knowledge of relationships between terms that can be used to compute similarity. WordNet can be used for this purpose by computing a similarity measure based on the link structure of the semantic network.
- Morphological analysis. Morphological knowledge can be applied to determine whether two words originate from the same stem. This includes inflectional morphology (e.g., walked → walk) as well as derivational morphology (e.g., destruction → destroy).
- Orthographic and phonological evidence. Terms that are spelled similarly but not identically might have been intended as the same term, perhaps with an unintentional misspelling. This might also be true of terms with similar pronunciation—they could represent alternate transliterations of the same proper name, for example.

Statistical analysis of term usage in a document collection provides another source of evidence. Two types of statistical evidence have been used:

- Syntactic dependency. Lin used a broad-coverage parser to extract dependency triples containing (headword, modifier, dependency type) and then computed the word similarity based on occurrence in similar syntactic structures with similar heads or modifiers [7].
- Proximity. Parsing is a relatively expensive process, and some statements are not sufficiently well formed to permit reliable parsing. Term cooccurrence within a sliding window provides a weaker (but more easily obtained) source of evidence that can be used in a similar way by treating proximity as a dependency type.

Similarity functions based on a combination of evidence are often used because the alternative sources of evidence have complementary strengths and weaknesses.

of a system with such a dictionary.

[3]Throughout this paper, we use the pinyin transliteration for Chinese characters.

McRoy, for example, used a linear scheme for combining semantic, syntactic, and proximity evidence for word sense disambiguation [10] and Mandala, et al. used a similar approach to control query expansion in monolingual context [8]. Approaches based on term clustering have been used for query expansion in monolingual retrieval, and similar ideas have been applied to CLIR by Sheridan and Ballerini [15] and others. We are, however, not aware of any prior application of automatic clustering techniques to translations contained in bilingual dictionaries.

3.2 Implementing Dictionary Clustering

We used complete link clustering with early termination. Our similarity measure was based on the following sources of evidence:

- WordNet similarity. For words and phrases that appeared in WordNet we use Lin's information theoretic similarity measure, equation (**??**), which is based on the intuition that the similarity of two synonym sets depends on the informativeness of each and the informativeness of their most specific subsumer [7]. After examining a few cases we selected an *ad hoc* threshold of 0.5 and treated two terms as similar if their computed similarity exceeded that value.
- Inflectional morphology. For single words, we treat two English words as similar if morphological analysis revealed that they were inflectional variants of the same root. We used the freely available WordNet morphological analysis software for this purpose.
- Orthographic similarity. For single words, we treat two English words as similar if adding, deleting, or changing a single character would convert one into the other.

Our approach to the combination of evidence is extremely simple—we consider two terms to be similar if they meet at least one of the criteria identified above:

```
for each Chinese term
 for each English translation 'a'
  for each existing cluster 'C'
   for each translation 'b' included in cluster
    if (Wordnet_sim(a,C.b)>0.5 or
        a and C.b are variants of same root or
        a and C.b have edit distance 0 or 1)
       add a to C and go to next translation
  if not assigned yet, start a new cluster
```

Lin's WordNet similarity measure is:

$$s(t_1, t_2) = \frac{2 \log P(C_0)}{\log P(C_1) + \log P(C_2)} \quad (1)$$

where t_i are terms, C_j are WordNet synsets, $t_1 \in C_1, t_2 \in C_2$, C_0 is the most specific synset that subsumes both C_1 and C_2, and $P(C_j)$ is the probability that a randomly selected term belongs to C_j.

3.3 Implementing Concept Inference

We exploit the local context of a Chinese term in the query to identify promising translation clusters by examining a window of query terms. For each translation cluster associated with a Chinese term, we collect evidence from every translation cluster of all other Chinese terms in the same window. We combine two sources of evidence:

- the WordNet similarity described in equation (1), with values between zero and one.
- log-likelihood cooccurrence statistics, computed using a window size of ± 3 words on an English corpus of 78 million words from Wall Street Journal, Associated Press and San Jose Mercury News stories in the TREC collection. We normalized these values by applying a scaling factor (0.05 for $\log_2$) and then limiting the result to fall between zero and one. Additional details of the computation are provided in [16]

A score for each cluster was computed within a window as follows:

```
for each Chinese term t1 in the window
 for each term t2 right of t1 in the window
  for each English cluster C1 of t1
   for each English cluster C2 of t2
    accumulate WordNet_sim(C1,C2) to C1 and C2
    accumulate cooccur(C1,C2) to C1 and C2
```

For relatively short queries, a single window containing all Chinese terms was used. For longer queries, a window containing 15 terms was stepped across the query in 12-term increments so that two successive windows will have 3 terms in common. We accumulate evidence for WordNet similarity and cooccurrence separately within a window as if they were probabilities, computing $a + b$ as $a + (1 - a)b$. For each English cluster, the overall score score is then computed as a linear combination of WordNet similarity (0.3) and cooccurrence (0.7). If an unstructured bilingual dictionary were used, the computational complexity of the algorithm would be $O(m^2n^2)$, where m is the average number of translations per term and n is the number of terms in the window for which multiple translations are known. For the experiments reported in this paper we instead used only the English term in each cluster that appeared most frequently in the Brown corpus, a balanced corpus of English. Using a structured dictionary in this way thus reduces m to the average number of clusters per query-language term, a substantial savings. At the end of the computation, each dictionary cluster has a numeric score that describes the likelihood of it representing the correct English meaning in the given context. The resulting scores can be used to select the best cluster or to weigh each cluster appropriately.

4. EXPERIMENT DESIGN

Experience has shown that several factors make it hard to obtain statistically significant results in CLIR evaluations. The translation process introduces a source of "noise" that can increase the sample variance of measures such as average precision, so the largest practical number of evaluation topics would be desirable. Paradoxically, CLIR test collections typically include fewer topics than their monolingual counterparts because the cost of developing those collections can only be amortized over the relatively few participants that have access to the needed linguistic resources. The use of small relevance judgment pools can exacerbate the problem, introducing undersampling effects that could contribute a systematic bias that statistical significance testing cannot reveal. In order to avoid these limitations, we chose to use a monolingual test collection for which translated queries are available, and to base our evaluation on the largest possible number of topics.

We used one hundred topic descriptions from the TREC-6 and TREC-7 ad-hoc evaluations (TREC topics 301-400) for the experiments. The full topic descriptions were manually translated into Chinese by native speakers at two different facilities. Topics 301-350 were translated at National Taiwan University, and topics 351-400 were translated at the University of Maryland. The Chinese topic descriptions were automatically segmented using software from New Mexico State University that is freely available for research use.[4] We formed two sets of queries from the automatically segmented topic descriptions by taking every character from the indicated fields: very short (title), long (title, description and narrative).

For query translation, we merged a bilingual term list provided by the Linguistic Data Consortium (LDC)[5] with a bilingual term list derived automatically from the Chinese-English Translation Assistance (CETA) dictionaries. The CETA group, started in 1965 and continuing into the 1990's, was a project to collect, evaluate, edit, and revise one single reference source for translating all kinds of Chinese documents into English.[6] The 230,000 entries in the CETA dictionaries were compiled from 250 dictionaries, some general purpose, others domain-specific. Because they were originally designed for manual use, explanatory definitions and examples of usage are present in addition to simple translations. We selected 20 CETA dictionaries with an emphasis on modern broad coverage dictionaries and dictionaries specialized for economics and politics. We first removed English entries that contained more than 10 words, which are almost certainly definitions. For remaining English entries that contained several words, we used Lin's freely available MINIPAR parser [6] and some simple pattern matching to identify a headword and retained the smallest coherent phrase that included that headword. Finally, we automatically mapped the resulting English terms into WordNet. Preprocessing for the LDC dictionary was considerably simpler, consisting mainly of removal of target language forms that were descriptions of function where automatically identifiable as such. When merging the term lists we sought to automatically remove duplicate entries. The alternate English translations for a Chinese term were then ranked as follows: first all single word entries are ordered by decreasing target language unigram frequency calculated according to the Brown corpus, followed by all multi-word translations, and finally single word entries with zero unigram Brown corpus frequency. The resulting bilingual term list contained 195,078 unique Chinese terms, with an average of 1.9 known English translations per Chinese term.

We indexed the TREC-7 collection, which contained 210,158 Financial Times stories from 1991–1994, 131,896 Los Angeles Times stories from 1989–1990, 55,632 Federal Register documents from 1994, and 130,471 Foreign Broadcast Information Service stories from 1996. Retrieval effectiveness measures were computed using TREC relevance judgments, which were developed at TREC-6 and TREC-7 using a pooled assessment methodology, and the trec_eval program.[7] Automatically constructed English queries were then submitted to the Inquery retrieval system (version 3.1p1), which is available for research use from the University of Massachusetts. The following Inquery operators were used:

#sum Belief values associated with each term or node are averaged, each term contributes equally.

#wsum Belief values associated with each term or node are averaged, with possibly unequal relative contributions from each term specified by the integer that precedes that term or node. The initial integer is the sum of the scaling factors.

#max The largest belief value associated with a term or node is selected.

#or The belief values are accumulated as if they were probabilities, computing $a + b$ as $a + (1 - a)b$.

#uwN Belief values associated with each term or node are used if the terms are found (in any order) within a window of N words.

#syn A belief value for the node is computed using the Inquery term weighting formula rather than

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

[5]Available at http://www.ldc.upenn.edu/TDT/

[6]A machine-readable version of the CETA dictionaries is available from MRM Corp., Kensington, MD.

[7]Information about the documents, queries and relevance judgments is available at http://trec.nist.gov/ and the trec_eval program is available at ftp://ftp.cs.cornell.edu/pub/smart/.

by combining belief values. The term frequency of each term is summed to provide the term frequency for this computation, and the number of documents in which any of the terms appears is determined with reference to the index. The computation is described in [5].

5. RESULTS

Our hypothesis was that within-cluster term selection should be handled differently from across-cluster term selection. For our initial experiments, we compared four techniques for using the belief values of alternative translations within a cluster. To do this, we treated every known translation as a single cluster and formed queries in five ways. Examples for each of are shown using a the very short query for TREC Topic 358 "blood-alcohol fatalities" that was manually translated to "xue3ye4jiu3jing1si3wang1shi4gu4" and then automatically segmented to "xue3ye4 jiu3jing1 si3wang1 shi4gu4." The corresponding dictionary entries are {blood bloodstream}, {ethanol alcohol spiritus}, {death demise deadly doom (to die) abosis}, and {trouble accident mishap}.

SU Single Unstructured. For each Chinese term, include every known translation. #sum(blood bloodstream ethanol alcohol spiritus death demise deadly doom #uw2(to die) abosis trouble accident mishap);

SP Single Pirloka. The query is structured using Pirkola's method [14]. #sum(#syn(blood bloodstream) #syn(ethanol alcohol spiritus) #syn(death demise deadly doom #uw2(to die) abosis) #syn(trouble accident mishap));

SO Single Or. For each Chinese term, combine belief values for each alternative translation using a probabilistic or operator. #sum(#or(blood bloodstream) #or(ethanol alcohol spiritus) #or(death demise deadly doom #uw2(to die) abosis) #or(trouble accident mishap));

SW Single Weighted. For each Chinese term, weight alternative translations equally. #sum(#wsum(2 1 blood 1 bloodstream) #wsum(3 1 ethanol 1 alcohol 1 spiritus) #wsum(6 1 death 1 demise 1 deadly 1 doom 1 #uw2(to die) 1 abosis) #wsum(3 1 trouble 1 accident 1 mishap));

BU Best Unstructured. For each Chinese term, select the most frequent English translation based on the word unigram statistics of the Brown corpus. #sum(blood ethanol death trouble);

Figure 3 shows the mean average precision for each technique, averaged over 100 very short or long queries. Ellipses in that figure enclose values that are statistically indistinguishable (for $p < 0.05$) using a two-tailed paired t- test. SP clearly outperformed all other methods, so we adopted Pirkola's method for within-cluster combination of evidence and compared SP with four techniques based on the use of structured dictionaries:

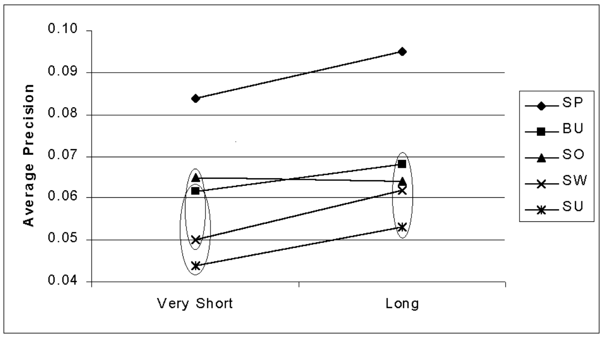

Figure 3: Within-cluster combination of evidence, averaged over 100 queries. Ellipses indicate statistically indistinguishable results.

BP Best Pirkola. Select the single cluster with the highest log-likelihood. #sum(#syn(blood bloodstream) #syn(ethanol alcohol) #syn(death demise) #syn(trouble accident mishap));

MP Maximum Pirkola. Select the cluster that produces the maximum belief value. #sum(#syn(blood bloodstream) #or(#syn(ethanol alcohol) spiritus) #or(#syn(death demise) deadly doom #uw2(to die) abosis) #syn(trouble accident mishap));

OP Or Pirkola. Combine the belief values for each alternative translation cluster using a probabilistic *or* operator. #sum(#syn(blood bloodstream) #max(#syn(ethanol alcohol) spiritus) #max(#syn(death demise) deadly doom #uw2(to die) abosis) #syn(trouble accident mishap));

WP Weighted Pirkola. Weight alternative translation clusters based on the available evidence. #sum(#syn(blood bloodstream) #wsum(3 3 #syn(ethanol alcohol) 0 spiritus #wsum(11 6 #syn(death demise) 1 deadly 2 doom 2 #uw2(to die) 0 abosis) #syn(trouble accident mishap));

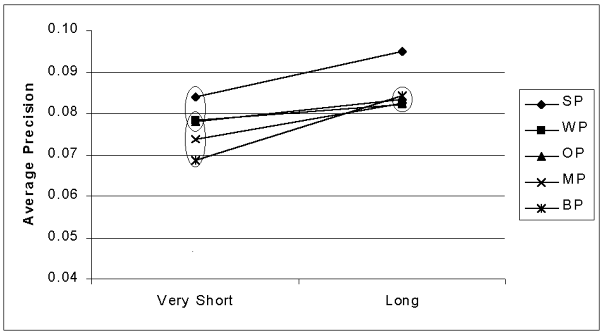

Figure 4: Across-cluster combination of evidence, averaged over 100 queries.

Figure 4 shows the mean average precision for each technique. For very short queries, SP, WP and OP are statistically indistinguishable, but SP outperforms all other techniques on long queries. Examination of the cases in which SP outperformed WP suggested that the clusters in our dictionary might be overly fine-grained for the purpose of retrieval, so one of us (a native speaker of English that does not know Chinese) reclustered a subset of the dictionary by hand. We started from the automatically clustered entries for the approximately 200 unique terms used in very short queries 351-400 and aggressively joined clusters unless we felt that it was clear that the terms were intended to represent different concepts. We did this once in the lexicographic order of the Chinese terms, and made no reference to the queries themselves during this process. Ultimately, we made changes to the clustering for 71 Chinese terms, reducing the average number of clusters per Chinese term over that set from 3.7 to 1.6, and more than half of those 71 reclustered Chinese terms retained only a single translation cluster. As Figure 5 illustrates, the performance of WP improved when the manually reclustered dictionary was used, achieving results identical to SP over those 50 queries. The slight decline shown for SP in that figure appears to be the result of the result of the manual removal of a few duplicate translations that our automatic dictionary cleanup had missed.

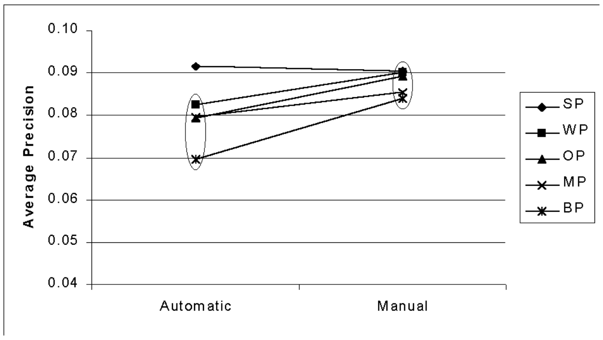

Figure 5: Manual clustering results, averaged over 50 queries. WP and OP overlap for very short queries.

We were surprised by our inability to outperform SP using structured translation. The relatively poor performance of BP points to one possible explanation for this—it appears that our somewhat *ad hoc* algorithm for assigning scores to each cluster would need to be improved. WP seems to be less sensitive to query length than SP, an observation that is consistent with the design of each. Terms that have several translations may have some that are frequent and others that are distinctive. Because ranked retrieval systems such as Inquery depend in part on the collection frequencies of each term to compute belief values, distinctive terms typically dominate the belief value for a document, with more common terms making a far smaller contribution. The document frequency recalculation that is built into SP is essentially a disaster-avoidance strategy, preventing individual translations from generating inappropriately high belief values by computing an aggregate document frequency for every known translation. Terms with a lower degree of translation ambiguity then have the opportunity to dominate the retrieval results. The larger vocabulary used in longer queries naturally provides more opportunities of the use of a term with only fairly selective translations, so better results are observed when long queries are used. WP uses a different strategy, seeking to limit the adverse effect on the document frequency computations on a term by term basis. While the use of better cluster weights might lead to better performance for WP on relatively short queries, the greater effect query length effect observed with SP may wash out any benefit WP can offer on longer queries.

It is interesting to note that the performance of OP is almost indistinguishable from that of WP in our experiments. We feel that this is of little consequence, however, since we can see no clear way of improving OP. The poor performance of MP should not be surprising when considered from the same light as SP. Not only does MP lack SP's disaster-mitigation mechanism, it is actually biased in favor of disaster enhancement! The presence of a single cluster of highly selective translations would be enough to capture the #max operator, thus dominating the belief value computations regardless of whether there is any reason to believe that cluster is an appropriate choice.

6. CONCLUSION AND FUTURE WORK

We have introduced the idea of structured translation and explored ways in which structured translation can be applied in a CLIR system. Although our experiments did not demonstrate better retrieval effectiveness than Pirkola's method, we believe that this may result from weaknesses in our present cluster weighting algorithm. Our experience with manual dictionary clustering has clarified the need for task-based measures of clustering effectiveness, and we plan to explore that issue further through user studies and additional retrieval experiments. The clustering algorithm used in our experiments builds on existing work in monolingual applications, but bilingual dictionaries also offer the potential for iterative refinement, first clustering in one language and then the other. We are interested in exploring that possibility as a way of improving cluster assignments. We also plan to explore clustering techniques that exploit additional sources of evidence, such as the use of derivational morphology and the cross-part-of-speech links in EuroWordNet.

Although we introduced structured translation from the perspective of query translation, the retrieval problem is symmetric—it is equally useful to think of selecting an indexing vocabulary for documents that searchers are likely to choose. Incorporation of these ideas in a CLIR approach based on document translation should thus be quite straightforward. The results that we have focused on in our analysis have generally been statistically significant, but differences of the magnitude we have observed (typically less than 20%) might be of little consequence in a real application. This automated processing is, however, generally only a part of some larger cross-language search process. As we mentioned in Section 2, structured translation might find its most important application in other parts of that process. For interactive retrieval, Oard and Resnik suggested that displaying as many as three translations, one for each possible meanings of a term, might help users recognize promising documents through a "pop out" effect in which the context reinforces the user's perception of the appropriate translation [12]. User-assisted query translation might also benefit from structured translation. Ogden et al., for example, displayed all known back-translations for each candidate translation to provide monolingual users with a basis for making such a selection [13]. Fewer such decisions would be needed if alternate translations were grouped by meaning. Structured translation back into the query language might also be helpful since alternatives could then be displayed in semantically related groups.

Ultimately, the value of structured translation rests on the degree to which the model depicted in Figure 1 represents the structure of the task at hand. The common feature of the tasks that we have identified—cross-language retrieval, gloss translation for browsing, and query refinement using retranslation—is that techniques are known that can tolerate some degree unresolvable ambiguity. Structured translation offers a principled way of using the available evidence in an appropriate manner without making inappropriate choices in cases where sufficient evidence is not available.

Acknowledgments

The authors would like to thank Philip Resnik for his advice and WordNet similarity computation routines, Dekang Lin for the use of MINIPAR and for precomputed co-occurrence data, Dagobert Soergel and Keith Cogdill for pointers into the information access literature, and Jianqiang Wang and Hsin-Hsi Chen for providing the translated queries. This work has been supported in part by DARPA contract N6600197C8540 and a Shared University Research equipment grant from IBM.

7. REFERENCES

[1] M. J. Bates. Subject access in online catalogs: A design model. *Journal of the American Society for Information Science*, 37(6):357–376, 1986.

[2] H.-H. Chen, G.-W. Bian, and W.-C. Lin. Resolving translation ambiguity and target polysemy in cross-language information retrieval. In *ACL 99*, pages 215–222, June 1999.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[4] N. Ide and J. Veronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.

[5] J. Kekäläinen and K. Järvelin. The impact of query structure and query expansion on retrieval performance. In *SIGIR 98*, pages 130–137, Aug. 1998.

[6] D. Lin. PRINCIPAR – an efficient, broad coverage, principle-based parser. In *COLING 94*, pages 42–48, 1994.

[7] D. Lin. An information-theoretic definition of similarity. In *Fifteenth International Conference on Machine Learning*, 1998.

[8] R. Mandala, T. Takenobu, and T. Hozumi. The use of WordNet in information retrieval. In *COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[9] J. S. McCarley. Should we translate the documents or the queries in cross-language information retrieval. In *ACL 99*, pages 208–214, June 1999.

[10] S. W. McRoy. Using Multiple Knowledge Sources for Word Sense Discrimination. *Computational Linguistics*, 18(1):1–32, 1992.

[11] D. W. Oard and A. R. Diekema. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, volume 33. ASIS, 1998.

[12] D. W. Oard and P. Resnik. Support for interactive document selection in cross-language information retrieval. *Information Processing and Management*, 35(3):363–379, July 1999.

[13] W. Ogden et al. Getting information from documents you cannot read: An interactive cross-language text retrieval and summarization system. In *SIGIR/DL Workshop on Multilingual Information Discovery and Access*, Aug. 1999.

[14] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *SIGIR 98*, pages 55–63, Aug. 1998.

[15] P. Sheridan and J. P. Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *SIGIR 96*, Aug. 1996.

[16] R. Sperer. Translation ambiguity reduction in translingual information filtering using WordNet. Master's thesis, University of Maryland, College Park, 1999.

[17] B. M. Wildemuth, K. Cogdill, and C. P. Friedman. The transition from formalized to compromised need in the context of clinical problem solving. In *Second International Conference on Research in Information Needs, Seeking and Use in Different Contexts*, pages 290–303, Aug. 1998.