LAMP-TR-097 February 2003
CS-TR-4452
UMIACS-TR-2003-22

The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval

Dina Demner-Fushman, Douglas W. Oard

Language and Media Processing Labratory
Instititue for Advanced Computer Studies
College Park, MD 20742

Abstract

Bilingual term lists are extensively used as a resource for dictionary-based Cross-Language Information Retrieval (CLIR), in which the goal is to find documents written in one natural language based on queries that are expressed in another. This paper identifies eight types of terms that affect retrieval effectiveness in CLIR applications through their coverage by general-purpose bilingual term lists, and reports results from an experimental evaluation of the coverage of 35 bilingual term lists in news retrieval application. Retrieval effectiveness was found to be strongly influenced by term list size for lists that contain between 3,000 and 30,000 unique terms per language. Supplemental techniques for named entity translation were found to be useful with even the largest lexicons. The contribution of named entity translation was evaluated in a cross-language experiment involving English and Chinese. Smaller effects were observed from deficiencies in the coverage of domainspecific terminology when searching news stories.

***The support of the LAMP Technical Report Series and the partial support of this research by the National Science Foundation under grant EIA0130422 and the Department of Defense under contract MDA9049-C6-1250 is gratefully acknowledged.

The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval

Dina Demner-Fushman
Department of Computer Science and UMIACS
University of Maryland
College Park, MD 20742
301-405-6746
demner@umiacs.umd.edu

Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland
College Park, MD 20742
301-405-7590
oard@glue.umd.edu

Abstract

Bilingual term lists are extensively used as a resource for dictionary-based Cross-Language Information Retrieval (CLIR), in which the goal is to find documents written in one natural language based on queries that are expressed in another. This paper identifies eight types of terms that affect retrieval effectiveness in CLIR applications through their coverage by general-purpose bilingual term lists, and reports results from an experimental evaluation of the coverage of 35 bilingual term lists in news retrieval application. Retrieval effectiveness was found to be strongly influenced by term list size for lists that contain between 3,000 and 30,000 unique terms per language. Supplemental techniques for named entity translation were found to be useful with even the largest lexicons. The contribution of named entity translation was evaluated in a cross-language experiment involving English and Chinese. Smaller effects were observed from deficiencies in the coverage of domain-specific terminology when searching news stories.

1 Introduction

The goal of Cross-Language Information Retrieval (CLIR) is to support the task of searching multilingual collections by allowing users to enter queries in a language that might be different from that in which the documents are written. In dictionary-based CLIR techniques, the principal source of translation knowledge is a translation lexicon (which often contains information extracted from machine-readable dictionaries). Very simple translation lexicons, bilingual term lists with no information about selectional preference, are widely used for this purpose. Bilingual term lists are widely available, but our experience suggests that retrieval effectiveness can vary substantially from one language pair to another and even within a language pair, depending on the size and quality of the bilingual term list. The causes of this variation are not yet well understood, and our goal in this paper is to explore one possible cause—term list coverage—using real bilingual term lists.

The translation component of dictionary-based CLIR techniques depend on a successful cascade of three processes: (1) selection of the terms to be translated, (2) generation of a set of candidate translations, and (3) use of that set of candidate translations in the retrieval process. For the first stage, the best results are typically obtained by translating multiword expressions when possible, backing off to individual words when necessary, and further backing off to morphological roots when the surface form cannot be found [17]. In the second stage, algorithms for choosing among alternative translations have been extensively studied, and older techniques based on averaging weights computed for each translation can benefit significantly from translation selection based on term co-occurrence within the target corpora [12]. The focus on the third stage has been somewhat more recent, with the best presently known technique based on accumulating term frequency and document frequency evidence separately in the document language, then combining that evidence to create query-language term weights [15, 13]. What is less well understood, however, is the effect of term list coverage. If a possible translation is not known, it cannot be selected, and therefore cannot be used. What effect will this have on retrieval? That question is the focus of this paper.

We begin with a review of prior controlled studies on this topic in which the effect of systematically altering term list coverage on retrieval effectiveness has been characterized. Each of these studies depended on artificially ablating the coverage in some manner, and to the best of our knowledge none of these approaches have been validated through comparison with naturally occurring bilingual term lists of

various sizes. In Section 3 we address this concern by assessing the effect of coverage deficiencies in actual bilingual term lists obtained from several sources. Our results indicate that retrieval effectiveness is indeed positively correlated with term list size, but that term lists with similar sizes sometimes yield substantially different retrieval effectiveness. We therefore examine one term list in greater detail in Section 4, identifying eight types of missing terms. Named entities are found to be by far the most common type of missing term in news stories, so we examine the effect of named entity translation in detail in Section 5. With this as background, we then describe the known techniques for accommodating coverage deficiencies when building CLIR systems and recommend topics for further research that seem particularly promising in light of the insights that we have obtained through the experiments reported in this paper.

2 Background

We began our exploration of this question several years ago with a simple experiment in which we used two bilingual term lists from different sources to measure the effect of the linguistic resource on the effectiveness of cross-language information retrieval [8]. We obtained what seemed like a counterintuitive result: the smaller term list (with 30,322 unique English terms) actually did somewhat better than the larger list (with 89,003 unique English terms), although the difference was neither large nor statistically significant. We tried combining the two lists (resulting in 97,603 unique English terms), obtaining an effectiveness measure between the two that of the small and the large list. Clearly, the size of the list did not tell the whole story and the key question was not whether you know a lot of translations, but whether you know the right ones!

This first study of ours had been inspired in part by a paper in which Grefenstette had suggested several alternative measures for the coverage measures for assessing the utility of a translation lexicon in CLIR applications. For example, he had calculated that the English portion of the relatively large (37,600 entry) ELRA Basic Multilingual Lexicon covered common terms quite well, with 97% of the 1,000 most common English words being found (after splitting multiword expressions, conflating inflectional variants, and excluding proper names) [6]. Less common words seemed more problematic, however, with only 51% of the most common 50,000 English words found in the lexicon. We therefore developed some alternative coverage measures based on different ways of looking at term importance. We were, however, unable to find a measure that was positively correlated with retrieval effectiveness.

A subsequent ablation study by Xu and Weischedel began to shed some light on this counterintuitive result [20]. Starting with an 80,000-word English-Chinese bilingual term list, they simulated the effect of smaller term lists by progressively removing terms from the list in order of increasing frequency of occurrence in a large collection of English documents. They found that mean average precision increased with the size of the lexicon, but reached a plateau once translations for the most common 20,000 English words were included. If this ablation process accurately models the way in which real bilingual term lists are formed, then a plausible explanation for the results we observed in our first study would be that the added terms in the larger bilingual term list were rare, and thus not likely to be observed in any particular set of queries.

That was the point of departure for the work reported in this paper—Xu and Weischadel's approach seemed to offer a useful insight into coverage effects, but before we could generalize from that work we needed some insight into whether their ablation model captured what really happened when people built bilingual term lists. After completing our work, we learned of a concurrent study by McNamee and Mayfield that shed additional light on this question [10]. They tried an approach similar to that of Xu and Weischadel, but with two key differences: (1) the selection of terms for which translation was suppressed was made randomly with a uniform distribution, and (2) corpus-based pre-translation expansion was used to enrich the set of terms to be translated. With no pre-translation expansions, they observed fairly consistent declines in retrieval effectiveness with coverage ablation, even with relatively large conditions. This tends to confirm our intuition that the way in which coverage ablation is modeled is consequential. McNamee and Mayfield's most important result, however, is that much of the lost effectiveness can be regained using corpus-based pre-translation expansion, because the effect of expansion increases markedly at higher ablation levels. We did not use pre-translation expansion for the experiments reported in this paper, and it is now quite clear that this will be an important area for further work.

3 Characterizing Term List Coverage

We obtained from the Internet 34 freely distributed bilingual term lists that each pair English with one of 24 other languages, and we extracted a 35th bilingual term list from a large machine-readable bilingual dictionary (see Appendix A for a listing). The smallest of these term lists (English-Eskimo) contains 700 unique English terms—the largest (an English-Chinese term list extracted from the machine-readable dictionary) contains 193,297 unique English terms. Although that set contains a good spread of sizes (measured as the number of unique English terms), no single language pair is represented by more than four term lists. Gaining the sort of insight that we sought using multi-

ple language pairs would be difficult, however, because coverage effects might well be masked by other factors (e.g., differential importance and/or effectiveness of compound splitting and morphological normalization). We therefore chose to focus only on the English side of each term list.

There are three basic approaches to CLIR: translate the query into the language of the document collection [7]; translate the documents into the language of the queries [9]; or create a language-neutral representation of both the queries and the documents (c.f. [4]). We chose to model a query translation process, and to suppress effects other than coverage by simulating translation from English to English in a way that was sensitive to the English-language coverage of each term list.

For our experiments, we used an information retrieval test collection from the Cross-Language Evaluation Forum (CLEF 2000). The collection contains 113,000 English news stories from the Los Angeles Times (about 435 MB of text), 33 English topic descriptions,[1] and binary (yes-no) relevance judgments for topic-document pairs.

We used this monolingual test collection with each specific bilingual term list to simulate CLIR in the following manner:

- English queries are formed using every word in the title and description fields of the topic description. This is typically a sentence or two, representative of how an information need might initially be expressed to a human intermediary that is helping with the search. We repeated our experiments with shorter queries built from the title field alone (which are designed to be representative of what a searcher might type into a Web search engine), obtaining similar results. Because the observed coverage effects were very similar for both sets of queries, we present results only for the title+description queries in this paper.

- Any English word that does not appear on the English side of the bilingual term list was removed from the query. We refer to this process as "filtering" the query using the term list. Resnik et al. observed that bilingual term lists found on the Internet often contain an eclectic mix of root forms and morphological variants, and proposed a backoff translation strategy in which English words with the same stems were conflated prior to translation [17]. This achieves an effect similar to McNamee and Mayfield's pre-translation expansion, but the key difference is that the conflation is performed only if the surface form of the word to be translated is not found—this limits the introduction of spurious translations when good ones are already known. We modeled backoff translation by including an alternate condition in which English terms in the bilingual term list were reduced to their stems using the Porter stemmer and added to the original list before matching. Because our English-English coverage assessment includes no actual translation, we modeled this as a single step with no actual backoff, but we nonetheless refer to it as the "backoff" condition.

- We use English as a surrogate for the second language, so we do not actually translate the filtered query in this experiment. This can be thought of as modeling a case in which translation of known terms is perfect and in which the handling of target-language terms is consistent. Since this will obviously sometimes not be the case in real applications, our method clearly computes only upper bounds on retrieval effectiveness—we would expect the results in actual analogous end-to-end CLIR applications to be lower. But by holding these other factors constant, we are able to focus more sharply on coverage effects.

- We search the English document collection using the InQuery text retrieval system and the filtered query. InQuery is a state-of-the-art system that ranks documents in decreasing order of likely relevance. We then compute mean average precision, a commonly used measure of comparative retrieval effectiveness that reflects the expected density of relevant documents near the top of the ranked list [19]. Mean average precision assumes values between zero and one (with higher values preferred). We performed a second set of experiments using a vector-space text retrieval system (MG) with similar results, so we are reasonably confident that the results reported in this paper are not overly dependent on the details of the design of the retrieval system.

In our experiments, the principal independent variable is the number of unique English terms in the bilingual term list (which we plot on the X axis), and the principal dependent variable is mean average precision (on the Y axis). We used a two-tailed t-test, pairing observations of mean average precision values by bilingual term list size, to test the statistical significance of observed differences.

As Figure 1 shows, the mean average precision for bilingual term lists of similar size exhibits quite a lot of variation. Simulating backoff translation typically results in greater retrieval effectiveness for any size term list (statistically significant at $p < 0.001$), and in less variation in retrieval effectiveness across the set of relatively large term lists. This clearly indicates that proper handling of morphological variants is an important issue for dictionary-based CLIR. Most of the smallest bilingual term lists—up to about

[1] The CLEF 2000 collection contains 40 topics, but no relevant English documents are known for topics 2, 6, 8, 23, 25, 27, and 35, so they were excluded from our experiments because they cannot distinguish between the conditions that we wished to explore.

3,000 unique English terms—are of little use for CLIR. Approximately linear growth in mean average precision is evident between about 3,000 and 20,000 terms, with little further improvement observed beyond that range. These observations are consistent with Xu and Weischadel's assumption that small term lists predominantly contain very common terms (which InQuery gives little weight to) and that the additional terms present in the largest term lists are so rarely used that they are very unlikely to be present in a query, and therefore tend to support their model.

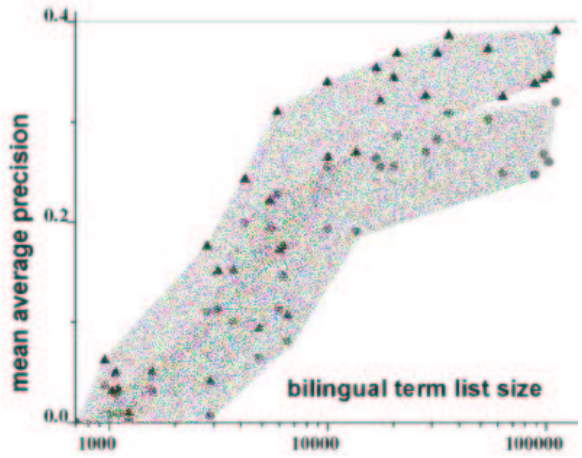

Figure 1. Effects of lexicon size. Upper area (triangles) with backoff translation, lower area (circles) without.

In this case, the original queries with no terms removed achieve a mean average precision of about 0.4. Interestingly, even when backoff translation is simulated, many of the largest bilingual term lists yield a mean average precision value that is 15-20% below that. Moreover, mean average precision aggregates the effect over many queries—the effect on average precision for individual queries is even more dramatic. For example, topic 17 (*Bush fire near Sydney*) achieves an average precision of 0.67 with one term list, but 0.32 with another of similar size. In the next section, we explore the question of what is missing from bilingual term lists found on the Internet that might be important in CLIR applications.

4 The Missing Terms

Our goal in this section is to examine a representative set of untranslatable terms that might appear in queries in order to focus our future efforts to improve dictionary coverage. When assessing coverage effects using a test collection, we can only see the effect of terms that happened to be present in some query. But we can gain access to a far larger set of terms that might be included in a query by examining the documents rather than the queries. We therefore chose to examine terms from two document collections in order to explore the reasons for coverage failures. We chose to study our largest non-Asian bilingual term list (German-English term list number 33 in Appendix A) because comparable collections were available for both languages. For the English analysis, we used the CLEF 2000 English collection (described above). For German, we used the CLEF 2000 German collection, which contains 153,499 stories (301 MB) published by the magazine *Der Spiegel* and the *Frankfurter Rundshau* newspaper in 1994.

The English collection was stemmed using Porter stemmer. The English collection contains over 60 million tokens after stemming and stopword removal, approximately 7% of which do not appear in the term list. German compounds were split using the greedy left-to-right longest substring match matching using the German side of bilingual term list. The German collection contains approximately 50 million tokens, approximately 17% of which do not appear in the term list. In each case, the missing terms were collected in a list and duplicates were removed. The first author of this paper then examined a randomly selected sample of 1,000 words from each list, and grouped the terms into categories. Figure 2 illustrates the observed distribution of terms in the following categories:

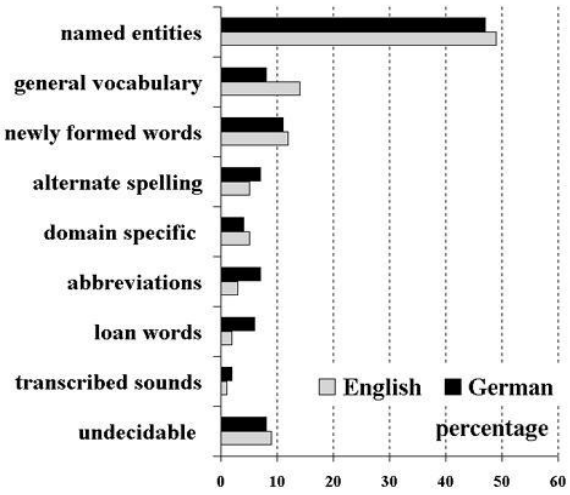

Figure 2. Distribution of the out-of-vocabulary words in the CLEF 2000 collection.

Named entities, which include proper nouns, locations, brand names, etc. For example, the absence of *Sydney* from some terms lists is what resulted in the extreme variability in average precision for the *Bush fire near*

Sydney topic. Named entities comprise almost half the missing terms.

General vocabulary, defined as words that would be expected to be found in a comprehensive printed monolingual dictionary without an annotation that its usage is restricted to a particular domain. For example, the general vocabulary term *firefighter* was missing from the English side of most bilingual term lists. Because missing German words were aggressively split to find known words but missing English words were not, general vocabulary is more often found to be missing in English.

Newly formed words, which are built from other words (typically as compound terms). For example, *cyberwalk* was found in the English collection, but that term would not likely appear in any dictionary. Some German compounds could not be split because one constituent was missing from the term list (e.g. in the compound German term *gruselstory* (horror story), story is a loan word from English that does not appear in the German side of the bilingual term list.

Alternate spellings, some of which are typographical errors, and others of which result from regular variations within a language. For example, the British spelling *privatisation* appears in one query, but the Los Angeles Times news stories contain only the American spelling *privatization*. We chose to group these two cases because we expected at the time of this study that similar methods (e.g., fuzzy matching) might be used to deal with them.

Domain-specific terminology, which would not be expected to be present in broad-coverage lexical resources. For example, the term *thoracoscope* would be expected to appear only in term lists specialized to the medical domain.

Abbreviations, which might either be acronyms or shortened forms of a term. For example, the abbreviation *url* did not appear on either side of the bilingual term list.

Loan words, which are adopted from another language with no change in meaning, but perhaps with minor variations in spelling. For example, the Russian term *glasnost* appeared in the English collection. Loan words were considerably more common in German than in English, perhaps reflecting the pervasive influence of American media on adoption of terms in other cultures.

Transcribed sounds, which are used to convey colloquialisms or to imitate sounds. For example the term *aaaarrff* was found in the German collection.

Undecidable, a category that was used to code any term that could not be reliably placed in another category. Terms were normalized to lower case and removed from their context before being examined, and sometimes this precluded accurate categorization. For example *simeone* might have been a misspelling of *someone*, or it might have been the name of a person.

As figure 2 clearly shows, named entities are by far the most common type of missing term. We therefore elected to study their impact on retrieval effectiveness in more detail in the next section.

5 The Impact of Named Entities

Early work on dictionary-based approaches to CLIR in European languages generally showed relatively little adverse effect from omission of named entities. When performing CLIR among European languages, the usual approach is to retain untranslatable terms unchanged (perhaps with the omission of accents or other diacritics), and named entities such as the names of persons were often written the same way in both languages. Once experimentation extended to language pairs with different character sets, however, the magnitude of the problem became clear. Since our goal is to characterize coverage effects, we have chosen to explore three conditions: (1) retain all named entities, (2) retain only those named entities that appear in the English side of the bilingual term list, and (3) retain no named entities (even if they appear in the term list). In each case, we retain all terms that are not named entities if and only if they are present in the term list. The first condition simulates the case in which named entity translation is perfect (e.g., when string matching between European languages works). The second simulates the cross-character set condition (with no augmentation from transliteration), and is identical to the condition reported in Section 3. The third is a contrastive condition that is designed to provide a reference point for the other two.

We hand tagged each named entity in the 33 title+description queries (22 queries actually contained named entities) and then repeated the experiments described in Section 3. As before, we performed this experiment using the English side of all of the term lists and only the CLEF 2000 English test collection. We obtained similar results with and without simulating backoff translation, so we show only the backoff condition in Figure 3. The trend in the first and third conditions is quite clear, with term lists that contain more than 30,000 unique English terms almost always achieving maximal retrieval effectiveness. The solid horizontal line at 0.4 shows the results of a full monolingual query, and the dashed line at 0.225 shows the results that would be achieved by removing all (and only) named

entities from the queries. Figure 3 reveals a sigmoidal shape, with little benefit from term lists that contain fewer than 3,000 unique English terms, nearly linear improvement with increasing coverage between 3,000 and 30,000 unique terms, and little benefit from further increases beyond that size. From examining the middle condition, it seems clear that much of the observed variation reported in Section 3 resulted from differences in the coverage of named entities.

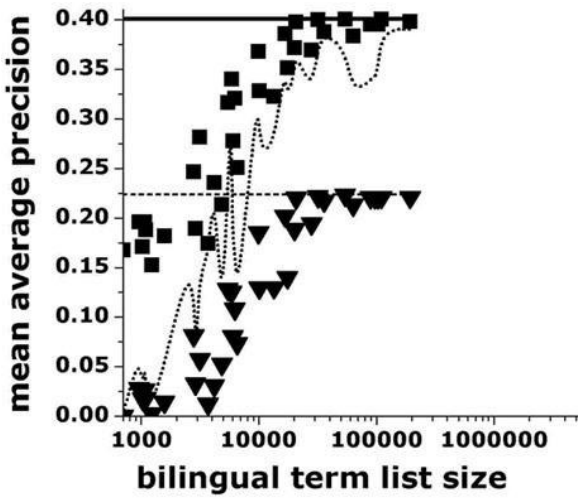

Figure 3. Impact of named entities in simulated CLIR. Squares: always used; Triangles: never used; Dotted line: used if present in term list.

We also ran one full CLIR experiment to see the effect of named entity handling in actual practice. We chose the English-Chinese language pair for those experiments because that allowed us to use our largest available bilingual term list (Number 35 in Appendix A). For this experiment, we used the TREC 5-6 Mandarin Corpus of approximately 170 megabytes of articles drawn from the People's Daily newspaper and the Xinhua newswire. That test collection contains 54 Chinese topics, for which both English translations and relevance judgments are available. We hand tagged all named entities in the English translations, and each was then manually translated into Chinese by a native speaker of Chinese. The same native speaker also hand-segmented the Chinese query terms for use in a contrastive monolingual Chinese run.

We used Pirkola's method [15]) to structure the translated queries and used the InQuery text retrieval system. This has the effect of estimating the within-document term frequency for each query term in each document as the sum of the frequencies of any translation of the query term in that document, and the collection-wide "document frequency" of a term as the number of documents in which any translation of that term occurs. The weight for each query term is thus computed in the document language—this is now widely accepted as a good approach when translation probability information is not available. The rest of the experiment design followed the monolingual experiment described in Section 3, with three conditions: (1) all named entities manually translated, (2) named entities translated only if present in the term list, and (3) named entities never translated. In each case, terms other than named entities were translated in whatever way the bilingual term list specified. For contrast, we also performed an ordinary (all terms retained) monolingual Chinese run and a run with no bilingual term list but all named entities manually translated.

Figure 4 shows the results. In this case, the bilingual term list contained many named entities, and suppressing their translation hurt substantially (over 60% reduction in mean average precision, statistically significant). Further improvement appeared to result from manually translating all named entities, but the advantage over using the term list alone was not found to be statistically significant. Taken together, these results seem reasonable, since the bilingual term list that we chose for this experiment is relatively rich in named entities. Interestingly, manually translating only the named entities (with no bilingual term list) did nearly as well as using the bilingual term list alone (with no manual translation of named entities). From this we conclude that the results of our one actual cross-language experiment tend to support the results we obtained with single-language coverage measurements using a broader range of bilingual term lists.

6 Accommodating Coverage Deficiencies

In this section, we briefly review techniques that have been used to overcome deficiencies in the coverage of bilingual term lists in CLIR systems. With that as background, we suggest further work on one additional technique, translation extraction from comparable corpora, that has been explored as an abstract problem in computational linguistics but not yet applied to CLIR.

Two broad classes of approaches have been tried to handle named entities that do not appear in bilingual term lists. The first, widely used when both languages are expressed in the same writing system, is to retain the untranslated term (or perhaps to strip accents, if accents are commonly used in only one of the two languages). In addition to matching names that are written identically, this also generates felicitous matches on loan words, which can be an important factor if cultural factors have resulted in significant sharing over time within the language pair. When the languages use different writing systems, the second approach, pho-

netic transliteration, provides a useful way to achieve similar effects (c.f., [1]).

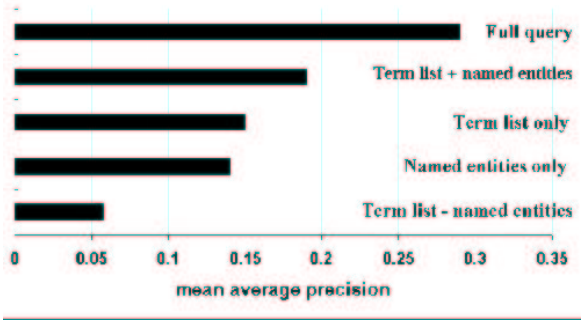

Figure 4. The effect of named entities in actual CLIR.

Our results suggest that three other relatively straightforward techniques should receive more attention than they have to date. The first is decompounding, which is clearly of value in a freely compounding language such as German, but which might also be helpful in some cases in languages such as English where the technique is rarely applied. Since decompounding is designed to enhance recall, possibly at the expense of precision, a backoff strategy in which decompounding is only tried when the original term is not found is probably a good idea. The second technique is normalization of alternative spellings, which is a variant of the spelling correction problem (c.f., [3]). The beneficial effect of normalizing alternative spellings may not be easy to see using standard test collections because most such collections are built using carefully edited news or journal articles with carefully prepared queries. But many real applications (e.g., Web searching) are considerably messier. In the one case we observed in our collection (*privatisation* vs. *privatization*), normalization was helpful, at least with larger term lists that were likely to contain the normalized form. Third, special handling for abbreviations and transcribed words, also seems to merit consideration. In many cases, we expect that language-specific heuristics will be needed, since the process of abbreviation exhibits considerable variation across languages and the transcription of sounds naturally depends on how people speak.

Of the two remaining categories, general vocabulary was more commonly missing than domain-specific vocabulary. This statistic can be misleading, however, for two reasons. First, it is an artifact of the genre of our test collection—news stories. In a collection of medical journal articles, for example, we would expect domain-specific terminology to be far more prevalent than it is in news stories. The second factor is that domain-specific terminology tends to be quite specific, and thus quite highly weighted by information retrieval systems. Finding larger bilingual term lists (up to approximately 30,000 unique terms) seems to be an effective way of increasing the coverage of general vocabulary, but unless specialized term lists are used, that approach would likely not do much to identify translations for missing domain-specific vocabulary. For this reason, in the remainder of this section, we explore the potential for learning translations of domain-specific terminology in another way.

Techniques for learning translations from parallel text collections (i.e., collections of translation-equivalent document pairs) have been widely studied (c.f., [11]), but domain-specific parallel text collections have proven to be difficult to obtain in many practical applications. For this reason, some researchers in computational linguistics have explored ways in which partial knowledge of possible translations is used to learn translations for additional terms from topically-related text collections (or "comparable corpora") in each language. Comparable corpora are typically easier to obtain than parallel corpora because different sources could provide the collections for each language.

The basic approach to learning new translations in this way is modeled to some extent on the way in which humans acquire vocabulary by reading—the context in which a term is used gives a clue about its meaning. The key idea for using comparable corpora to learn new translations is to start with an incomplete bilingual term list, use the known translation relationships to discover regions in the two collections that have a similar pattern of word use (and hence a similar topical focus), and then hypothesize translation relationships between any terms that lack a known translation relationship but that repeatedly appear together in regions that are paired in this way. The process can then be iterated to further improve coverage. Three variants on this basic idea have been tried by Rapp [16], Fung [5], and Picchi and Peters [14]. Every technique that has been tried relies on some way of estimating the importance of individual terms and then combining those estimates to evaluate term importance. Fung adapted measures used in information retrieval, while both Peters and Picchi and Rapp tried similar ways of comparing observed and expected frequencies to estimate the information content of an observed co-occurrence. Sliding windows are commonly used to limit the extent of the regions that are candidates for alignment.

Although Sheridan and Schauble demonstrated improved performance in cross-language retrieval when domain specific similarity thesauri obtained from context alignment on document level were used, we are not aware of any case in which the idea of sliding windows for context alignment have been applied in a CLIR application [18].

We have started to explore the potential value of translations learned from comparable collections to CLIR. Picchi and Peters observed that although homonymy (terms with different meanings that are written identically) precluded

effective use of the technique with news collections, it could be quite useful in domain-specific applications. We do not yet have an appropriate evaluation collection that is domain-specific, so we have started by checking to see whether any benefit might be found in the collections that we have been working with. We implemented a technique that aligns unknown terms according to similarity of their translated contexts within windows of 3,4, and 5 tokens immediately surrounding an unknown term for which we desired to learn a translation and measured the degree of association between that term and all terms in its context. We call our normalized association measure [2] 'affinity,' and calculate it as follows:

$$A(w_1, w_2) = \frac{count(w_1, w_2)}{count(w_1) + count(w_2) - count(w_1, w_2)}$$

In our initial experiments with news stories, we have found that this measure tends to associate domain-specific terms with synonyms and related terms in the other language, in other words, it appears that the learned "translations" might be useful for information retrieval. Moreover, we observed that the set of "translations" associated with relatively specific terms tends to be relatively short. For example the set of English terms with the strongest affinity to the German word for *tumor* are *cancer, oncology, diagnose, risk, treatment, cause, cell, surgery, radiation, chemotherapy*, and *leukemia*. The term *tumor* does appear in one query, so we computed the average precision for that query without that term and with that term, finding some benefit.

With general vocabulary, we observed that terms associated with particular events in our collection tended to have a high affinity with the terms that describe that event. For example *Estline, Baltic*, and *Estonia* have the highest affinity with the word *ferryboat* (the collections contain stories about a ferryboat that sank in the Baltic Sea). At this point our results are merely suggestive, of course, but it appears that this is a promising direction for further work.

7 Conclusion

We have shown that the pattern previously observed in ablation studies of lexicon size on retrieval effectiveness is discernible in large number of bilingual terms lists that we obtained from the Internet. Bilingual term lists containing at least 30,000 unique terms in the query language were found to optimize the coverage of general vocabulary, although the coverage of named entities was found to be highly variable, resulting in substantial variations in retrieval effectiveness for lexicons of similar size. We found that named entities make important contributions to retrieval effectiveness when searching news, but we noted that proper handling of domain-specific terms may be more important in other applications. We therefore also began to explored a strategy for learning translations of domain-specific terminology from comparable corpora.

Our work raises some interesting new questions, perhaps the most important of which is whether comparable corpora can be shown to be useful in domain-specific CLIR applications. In addition to questions of retrieval effectiveness, important questions about computational complexity and data sparseness remain to be explored. The National Institute of Informatics (Japan) has created a Japanese-English test collection of scientific paper abstracts that might prove to be a useful tool for exploring this question. Another question raised by Mayfield and MacNamee's recent work is the effect of using pre-translation query expansion in conjunction with the techniques that we are exploring. By exploring questions such as these, we hope to push frontier in CLIR research, expanding beyond our roots in retrieval from collections of news stories to a broad range of applications that reflect the rich potential of this technology.

Acknowledgments

The authors would like to thank Terry Zhao for tagging named entities in Chinese queries. This work has been supported in part by DARPA cooperative agreement N660010028910.

References

[1] Y. Al-Onaizan and K. Knight. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL02)*, 2002.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, New York, 1999.

[3] Michele Banko and Eric Brill. Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In *Proceedings of the Conference on Human Language Technology*, San Diego, 2001.

[4] Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence, March 1997.

[5] Pascale Fung. A statistical view of bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In David Farwell, Laurie Gerber, and Eduard

Hovy, editors, *Third Conference of the Association for Machine Translation in the Americas*, pages 1–16. Springer, October 1998.

[6] Gregory Grefenstette. Evaluating the adequacy of a multilingual transfer dictionary for the cross language information retrieval. In *First International Conference on Language Resources and Evaluation*, pages 755–758, May 1998.

[7] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[8] Gina-Anne Levow and Douglas W. Oard. Evaluating lexicon coverage for cross-language information retrieval. In *Workshop on Multilingual Information Processing and Asian Language Processing*, pages 69–74, November 1999.

[9] J. Scott McCarley. Should we translate the documents or the queries in cross-language information retrieval. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 208–214, June 1999.

[10] Paul McNamee and James Mayfield. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceeding of the 25th annual international conference on Research and development in information retrieval*, pages 159–166. ACM Press, 2002.

[11] I. Dan Melamed. Empirical methods for exploiting parallel texts. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Third Conference of the Association for Machine Translation in the Americas*, pages 18–30, October 1998.

[12] Douglas W. Oard and Anne R. Diekema. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, volume 33. American Society for Information Science, 1998.

[13] Douglas W. Oard and Funda Ertunc. Translation-based indexing for cross-language information retrieval. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR*, pages 324–333, March 2002.

[14] Eugenio Picchi and Carol Peters. Cross language information retrieval: A system for comparable corpus querying. In *Cross-Language Information Retrieval*. Kluwer Academic, 1998.

[15] Ari Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–63, August 1998.

[16] Reinhard Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Meeting of ACM*, pages 320–322. Cambridge, MA, 1995.

[17] Philip Resnik, Douglas Oard, and Gina Levow. Improved cross-language retrieval using backoff translation. In *First International Conference on Human Language Technologies*, 2001.

[18] Páraic Sheridan and Peter Schäuble. Cross-language information retrieval in a multilingual legal domain. In *First European Conference on Research and Advanced Technology for Digital Libraries*, September 1997.

[19] Ellen M. Voorhees. Variations in relevance judgments and the mesaurement of retrieval effectiveness. In C.J. Van Rijsbergen W. Bruce Croft, Alistair Moffat, editor, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323. ACM Press, August 1998.

[20] Jinxi Xu and Ralph Weischedel. Cross-lingual information retrieval using hidden markov models. In *Proceedings of the 2000 Joint SIGDAT Conference*, pages 95–103, Hong Kong, October 2000.

Appendix A. Bilingual term lists used in the experiment.

	language	*lexicon size*	*available from*
1.	Eskimo	700	http://www.pageweb.com/kleekai/
2.	Swahili	957	http://www.freelang.com
3.	Old English	1,026	http://www.freelang.com
4.	Indonesian	1,070	http://www.freelang.com
5.	Welsh	1096	http://www.freelang.com
6.	Portuguese	1,228	http://www.june29.com/IDP
7.	Latin	1,568	http://www.freelang.com
8.	Finnish	2,804	http://www.freelang.com
9.	French	2,898	http://www.june29.com/IDP
10.	Icelandic	3,148	http://www.freelang.com
11.	Danish	3,703	http://www.freelang.com
12.	Afrikaans	4.185	http://www.freelang.com
13.	Italian	4,860	http://www.june29.com/IDP
14.	Greek	5,437	http://www.freelang.com
15.	Portuguese	5,868	http://www.freelang.com
16.	Norwegian	6,027	http://www.freelang.com
17.	German	6,265	http://www.june29.com/IDP
18.	Spanish	6,545	http://www.june29.com/IDP
19.	Dutch	9,959	http://www.freelang.com
20.	Swedish	10,052	http://www.freelang.com
21.	Italian	13,475	http://www.wordgumbo.com
22.	Esperanto	16.710	http://www.freelang.com
23.	French	17,466	http://www.freelang.com
24.	French	20,078	http://www.wordgumbo.com
25.	Spanish	20,761	http://www.wordgumbo.com
26.	Italian	28,087	http://www.freelang.com
27.	Russian	31,725	http://www.freelang.com
28.	Spanish	35,752	http://www.freelang.com
29.	Japanese	54,112	http://www.freelang.com
30.	Hungarian	63,164	http://www.freelang.com
31.	German	89,046	http://www.freelang.com
32.	German	97,038	http://www.quickdic.de/
33.	German	103,166	http://www.tu-chemnitz.de/dict
34.	Chinese (LDC Version 2)	110,831	http://morph.ldc.upenn.edu
35.	Chinese (CETA)	193,297	MRM corporation, Kensington, MD