Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 475–487
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Empirical studies on the impact of lexical resources on CLIR performance☆, ☆☆

Jinxi Xu *, Ralph Weischedel

BBN Technologies, 10 Moulton Street, Cambridge, MA 02138, USA

Received 10 June 2004; accepted 14 June 2004
Available online 20 August 2004

Abstract

In this paper, we compile and review several experiments measuring cross-lingual information retrieval (CLIR) performance as a function of the following resources: bilingual term lists, parallel corpora, machine translation (MT), and stemmers. Our CLIR system uses a simple probabilistic language model; the studies used TREC test corpora over Chinese, Spanish and Arabic. Our findings include:

- One can achieve an acceptable CLIR performance using only a bilingual term list (70–80% on Chinese and Arabic corpora).
- However, if a bilingual term list and parallel corpora are available, CLIR performance can rival monolingual performance.
- If no parallel corpus is available, pseudo-parallel texts produced by an MT system can partially overcome the lack of parallel text.
- While stemming is useful normally, with a very large parallel corpus for Arabic–English, stemming hurt performance in our empirical studies with Arabic, a highly inflected language.

☆ This research is sponsored by the Defense Advanced Research Projects Agency and managed by SPAWAR under contract N66001-00-C-8008. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency, SPAWAR, or the United States Government.

☆☆ Some of the results in this work have appeared in Xu, Weischedel, and Nguyen (2001) and Xu and Weischedel (2003).

* Corresponding author. Tel.: +1 617 873 7674; fax: +1 617 873 2534.
E-mail addresses: jxu@bbn.com (J. Xu), weisched@bbn.com (R. Weischedel).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.06.009

Keywords: Cross-lingual retrieval; Parallel texts; Stemming; Machine translation; Bilingual lexicons

1. Introduction

Research and results in cross-lingual information retrieval (CLIR) have blossomed particularly in the last three years as the availability of corpora, query sets, and relevance judgments have become available, e.g., through TREC (http://trec.nist.gov), CLEF (http://clef.iei.pi.cnr.it/), and NTCIR (http://research.nii.ac.jp/~ntcadm/index-en.html). Our goal is to empirically determine the effect on CLIR performance of the availability of various resources, e.g., the size of a manual bilingual term list, the size of a parallel corpus of translated materials, the availability of a machine translation system, and the availability of a stemmer for the document language.

Our experiments keep constant the statistical retrieval model (Section 2) and the query language (English), but vary the resources available in three fairly disparate languages (Arabic, Chinese and Spanish) as the document language. Arabic and Spanish are highly inflected. By contrast, written Mandarin Chinese has little complexity for stemming; however, word/token boundaries are not marked, but must be induced. We hope that these kinds of experiments will lead to engineering approximations to predict the kind of performance one could expect on a new language, given the resources available; however, we do not claim to have such predictive ability yet.

The relative performance of the underlying statistical retrieval model compared to other approaches is addressed elsewhere (Xu et al., 2001), rather than in this paper. In official TREC evaluations (Voorhees & Harman, 2001–2003), the algorithm has performed among the best in the state of the art; cross-lingual performance has consistently been near or above monolingual performance.

We should point out that blind feedback, a popular technique for improving retrieval performance, is ignored in this work. As a query expansion technique, the success of blind feedback largely depends on the availability of a suitable corpus in the language of the queries. In comparison, implicit query expansion techniques that leverage alternative translations to improve CLIR such as our retrieval model do not suffer from that limitation. McNamee and Mayfield (2002), discussed how to use blind feedback to mitigate the problem of insufficient translation resources.

Section 2 briefly defines the probabilistic retrieval model. Section 3 identifies the TREC corpora used in the experiments. In Section 4, the resources (manual bilingual dictionaries, parallel corpora and stemming algorithms, where appropriate) are identified for Arabic, Chinese, and Spanish. In Section 5, we report a battery of results in CLIR from a Chinese collection, measuring average precision of retrieval as a function of manual bilingual dictionary size and size of parallel corpus for inducing probabilistic term translation. In Section 6, we report an experiment with CLIR from a Spanish TREC collection, this time using a "pseudo-parallel" corpus, automatically generating English translations of the collection by applying a MT system to the collection. Given the pseudo-parallel corpus, we induce a statistical bilingual dictionary. The surprising result is that average precision in CLIR slightly exceeds that of retrieval using machine translation of both the documents and the queries. In Section 7, we report results based on extensive resources for Arabic, including a linguistically motivated morphological analyzer, an Arabic–English parallel corpus including 86 million words of text, and various approaches to stemming. The surprising result is that a statistical bilingual lexicon automatically derived from such a large parallel corpus obviates the need for stemming of the Arabic corpus. Section 8 describes related work. Section 9 draws conclusions. Throughout this paper, we report results in terms of the TREC average non-interpolated precision to measure retrieval performance (Voorhees & Harman, 2002).

2. A CLIR system based on a generative probabilistic model

2.1. The generative model

This cross-lingual retrieval model is an extension of the monolingual retrieval model proposed by Miller, Leek, and Schwartz (1999). It consists of a generative probabilistic model that estimates the probability of generating the query given a document. In monolingual retrieval, a word may be generated either from the document (without change) with some probability or from general English with some probability. In cross-lingual retrieval, the same model may be used; however, to generate an English (the query language) word from a document in a different language. We assume that generating a query word from the document means generating a word in the document language (e.g. Arabic) with some probability and translating that word to English with some probability.

The basic function of an IR system is to rank documents against a query according to relevance. By Bayes' rule,

$$P(D \text{ is rel} \mid Q) = \frac{P(D \text{ is rel})P(Q \mid D \text{ is rel})}{P(Q)}$$

Here D is a document and Q is a query. $P(D \text{ is rel})$ is the prior probability of relevance for D, which we assume to be a constant. [1] $P(Q)$ is the prior probability that Q is generated; since Q is a constant, $P(Q)$ has no effect on document ranking. We can therefore rank documents by $P(Q|D \text{ is rel})$, the probability that query Q is generated given document D.

In this study, queries are in English and documents are in another language. We assume two states, the *general English state* (*GE*) and the *document state* (*D*). In the General English state, an English word for the query is generated; it may or may not describe the content of the document. In the document state, a word from the document is chosen and translated to an English word for the query. The following pseudo-code describes the query generation process:

```
Until all query words are generated
{
Toss a biased coin with probabilities a for heads and 1 − a for tails. Enter the General English state if it is heads and the document state otherwise.
General English state: Pick an English word from the English vocabulary according to a probability distribution.
Document state: Pick a non-English word from the document according to a probability distribution and translate it to an English word according to another probability distribution.
}
```

Fig. 1 illustrates how the query word "computer" might be generated from an Arabic document. With probability a, we could enter the state *GE* and generate "computer" with probability 0.0001 (times a). With probability $1 - a$, we could enter the document state (D), and generate an Arabic word with a given probability (on the link from D to the word), and translate it to "computer" with a probability (on the link from the word to "computer").

To minimize the need for training data, we estimate parameters as follows:

[1] Previous studies show that all documents are not equal. Longer documents in the TREC corpora, for example, are more likely to be relevant than short ones (Singhal, Buckley, & Mitra, 1996). We ignore this issue because it is not a concern in this study.

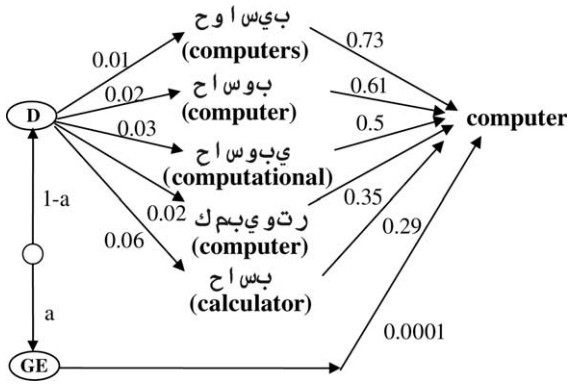

Fig. 1. Graphical view of probability model. The query word "computer" can come from either the document (D) or from general English (GE).

1. The parameter a is a constant. We fix it at 0.3 in this work, based on prior experience.
2. In the general English (GE) state, we estimate the probability distribution as follows:

   $$P(e|GE) = \text{freq}(e, GE)/|GE|$$

   where $\text{freq}(e, GE)$ is the frequency of English word e in an English corpus and $|GE|$ is the size of the English corpus. Any large English corpus can be used for this purpose; we used TREC volumes 1–5 of English data.
3. In the document state (D), we estimate the probability distribution as follows:

   $$P(c|D) = \text{freq}(c, D)/|D|$$

   where $\text{freq}(c, D)$ is the frequency of a non-English word c in D and $|D|$ is the length of the document.
4. The probability of translation to an English word e given a non-English word c, $P(e|c)$, depends on c and e only.

With these assumptions, it is easy to verify that:

$$P(Q|D) = \prod_{e \text{ in } Q} \left(aP(e|GE) + (1-a) \sum_{\text{non-English words } c} P(c|D)P(e|c) \right)$$

In our discussion, we assume that the translation of a term is independent of the document and independent of the query due to data sparseness. The assumption dramatically reduces the number of parameters we need to estimate. If more data (such as a very large parallel corpus) is available for parameter estimation, the independence assumption can be weakened to make the model more powerful. One possible technique is to employ bigram and trigram information to improve term translation.

A potential drawback of the model is that retrieval can be rather slow if it is improperly implemented, due to the involvement of alternative translations in score computing. The problem can be addressed by off-line computation of the probability score for each pair of English term e and foreign language document D at the expense of a higher storage cost.

2.2. Estimating translation probabilities

We use two techniques to estimate translation probabilities. For manual bilingual lexicons, we assume uniform translation probabilities. That is, if a non-English word c has n translations e_1 to e_n, we assume $P(e_i|c) = 1/n$.

Table 1
Statistics about test collections

Corpora	Document language	Number of documents	Number of queries
TREC5S	Spanish	172,952	25
TREC5C	Chinese	164,789	28
TREC6C	Chinese	164,789	26
TREC9X	Chinese	127,938	25
TREC2001X	Arabic	383,872	25
TREC2002X	Arabic	383,872	50

S = Spanish track, C = Chinese track, X = cross-lingual track.

When we have a parallel corpus, we use Brown et al.'s statistical machine translation models (Brown, Della Pietra, Della Pietra, Lafferty, & Mercer, 1993) to automatically induce a probabilistic bilingual lexicon. During the course of our experiments, we used two implementations of those statistical machine translation models: WEAVER (Lafferty, 1999) and GIZA++ (Och & Ney, 2000). Though both implement several of the statistical machine translation models, we used Model 1, the simplest, for its efficiency; thus far, we have not seen an improvement in CLIR performance from the richer models. In order to keep the size of the induced lexicon manageable, a threshold (0.01) was used to discard low probability translations.

3. TEST corpora

Table 1 summarizes the evaluation materials used, including the language of the corpus, the number of documents, and the number of queries.

4. Lexical resources

Throughout this work, the Porter stemmer (Porter, 1980) was used to stem English words.

4.1. Arabic

For Arabic, we used a manual lexicon and a parallel corpus for estimating term translation probabilities. The bilingual lexicon from Buckwalter (2001) has 86,000 word pairs. Uniform translation probabilities are assumed for the English translations in the lexicon.

The parallel corpus was obtained from the United Nations (UN) web site (http://www.un.org), which publishes all UN official documents under a document repository. A special purpose crawler was used to extract documents that have versions in English and Arabic. After a series of clean-ups, we obtained 38,000 document pairs with 86 million English words. Sentence alignment was carried out using an in-house developed algorithm. This corpus is now available through the Linguistic Data Consortium. [2] Translation probabilities were obtained by applying GIZA++ on the UN corpus.

The translation probabilities for the two sources were linearly combined to produce a single probability estimate for each word pair:

[2] Available from LDC as a pre-release, catalog no: LDC2003E11.

$$P(e|a) = 0.8P_{\text{un}}(e|a) + 0.2P_{\text{lexicon}}(e|a)$$

where e is an English word, a is an Arabic word, P_{un} and P_{lexicon} are probabilities from the UN corpus and the manual lexicon respectively. We gave a higher weight to the UN corpus because it appears to be of higher quality.

For Arabic stemming, we used the Buckwalter morphological analyzer. It uses a table-driven algorithm, employing a number of tables that define all valid prefixes, stems, suffixes, and their valid combinations. Given an Arabic word w, the stemmer tries every segmentation of w into three sub-strings, $w = x + y + z$. If x is a valid prefix, y a valid stem and z a valid suffix, and if the combination is valid, then y is considered a stem. If several valid combinations are found, it returns all of the stems. We re-implemented the analyzer to make it faster and compatible with the UTF8 encoding. We also modified it so that if no valid combination of prefix-stem-suffix is found, the word itself is returned as the stem.

4.2. Chinese

Two manual lexicons and one parallel corpus were used for the English and Chinese CLIR experiments:

1. The LDC lexicon. It is available from the Linguistic Data Consortium (LDC).
2. The CETA lexicon. It can be obtained through MRM Corporation, Kensingston, MD.
3. HKNews (Hong Kong SAR News) corpus. This parallel corpus consists of 18,000 pairs of documents in English and Chinese, with about 6 million English words. The corpus is available from LDC.

In order to increase lexicon coverage and to produce more robust probability estimates, different lexicons (including manual and induced) were combined to produce a single lexicon. Translation probabilities from different sources were linearly combined with equal weights:

$$P(e|c) = (P_{\text{ldc}}(e|c) + P_{\text{ceta}}(e|c) + P_{\text{hknews}}(e|c))/3$$

Table 2 shows the statistics about the lexicons.

4.3. Spanish

A co-occurrence based stemmer (Xu & Croft, 1998) was used to stem Spanish words. For Spanish, we did not have a large parallel corpus readily accessible; however, state-of-the-art machine translation systems are readily available. Therefore, we used SYSTRAN version 3.0 (http://www.systransoft.com) over the TREC5S Spanish corpus of approximately 35M words, yielding translations for roughly 400k Spanish word stems. The translated corpus was treated as a pseudo-parallel corpus, from which probabilistic term translations were derived using GIZA++.

Table 2
Lexicon statistics

Lexicon name	Number of English terms	Number of Chinese terms	Number of translation pairs
LDC	86,000	137,000	240,000
CETA	35,000	202,000	517,000
HKNews	21,000	75,000	1,266,000
Combination	105,000	305,000	1,490,000

Combination = a combination of all three sources.

5. Effect of size of bilingual lexicons and parallel corpora on CLIR performance

Availability of bilingual resources varies from one pair of languages to another. While such resources abound for high-density language pairs (e.g. English and Chinese), they are scarce for many so-called low-density language pairs. Considering this variability, it would be very useful if we can determine what level of CLIR performance is achievable given the existing bilingual resources. Furthermore, if existing data cannot meet our goal for CLIR, we would like to know how much more data has to be created. In this section, we will empirically measure CLIR performance as a function of two variables, the size of the manual lexicon and the size of the parallel corpus.

In Fig. 2, each curve corresponds to a parallel corpus of a certain size. The experiments were carried out on TREC5C and TREC6C. Only the title and description fields of the TREC topics were used in query formulation. A parallel corpus of *n words* was created by using the first *m* sentences pairs in the HKNews corpus that contain *n* English words. The *X* axis corresponds to the size of the manual lexicon. A lexicon of *n* words contains the most frequent *n* English words with their Chinese translations in the combined LDC–CETA lexicon. The frequency of the English words were compiled from TREC English volumes 1–5. The *Y* axis shows retrieval performance. The curve labeled 0 reflects the manual lexicon alone while the leftmost point in each curve reflects the parallel corpus alone. For reference, monolingual performance was 0.420.

Several observations can be made. First, as expected, the combination of a parallel corpus and a manual lexicon is always better than either resource alone, regardless of their sizes.

Second, without any parallel text, there is a limit on the CLIR performance when we increase the size of the manual lexicon. To achieve beyond that limit, parallel text has to be used. Although Fig. 2 shows that the value of parallel text is modest when we have a large manual lexicon, this is largely due to the dialect mismatch between the HKNews parallel corpus (Cantonese) and the TREC5&6C collections (Mandarin). The results suggest that manual lexicons alone are not sufficient for high performance CLIR; parallel text has to be used together with manual bilingual dictionaries. With no parallel corpus, after the dictionary exceeds 20,000, performance levels off. An examination of the translated queries shows that words not appearing in the 20,000-word lexicon usually do not appear in the larger lexicons either. Thus, increases in the general lexicon beyond 20,000 words did not result in a substantial increase in the coverage of the

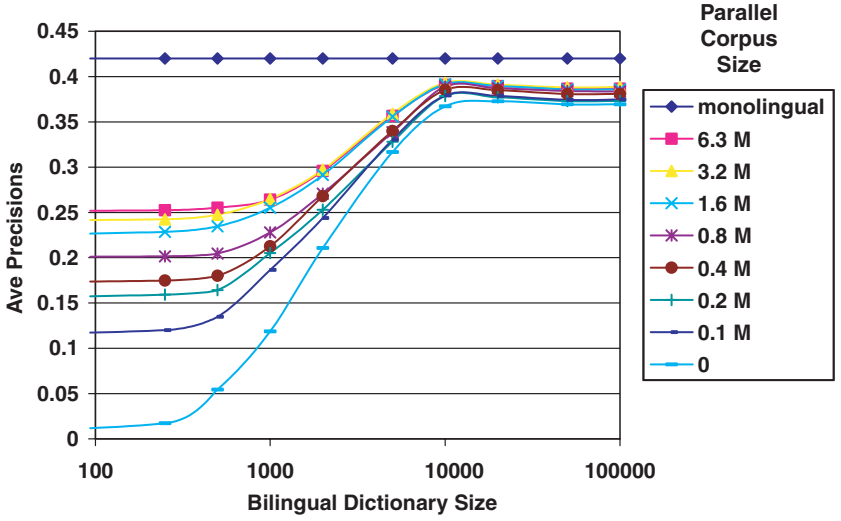

Fig. 2. English–Chinese CLIR performance as a function of sizes of manual lexicon and parallel corpus.

query terms. Another problem with manual lexicons is that they provide no probability information about possible translations. That forces us to naïvely assume all possible translations are equally likely, improperly penalizing good translations and rewarding bad ones.

Third, different combinations of lexicon and parallel corpus sizes can produce the same performance. This gives us some freedom in choosing what types of resources to acquire if our goal is to achieve decent but not the best possible CLIR. For example, to achieve 85% of monolingual IR, Fig. 2 shows that we could either have a 10,000 word lexicon with no parallel text or have a 5000 word lexicon and a medium size parallel corpus of a few million words.

Fourth, we categorized the missing terms and found that most of them are proper nouns (especially locations and person names), highly technical terms, or numbers. Such words understandably do not normally appear in traditional lexicons. Translation of numbers can be solved using simple rules. Transliteration, a technique that guesses the likely translations of a word based on pronunciation, can be readily used in translating proper nouns. Prior work (Demner-Fushman & Oard, 2003) demonstrated that name-entity translations can help dictionary-based CLIR.

Fig. 3 shows the results of a similar set of experiments on the TREC 2001 & 2002, using 75 English queries against an Arabic collection. The curve labeled 0 reflects the manual lexicon alone while the leftmost point in each curve reflects the parallel corpus alone. Overall, the patterns of the results are similar to Fig. 2. One difference is that the manual lexicon (Buckwalter lexicon) is less crucial when a large parallel corpus (the UN corpus with 86 million words) is available. But even with such a large parallel corpus, the manual lexicon still contributes somewhat to the combined performance, improving the score from 0.2981 to 0.3229. A monolingual baseline for the combined TREC 2001 & 2002 query set is not available.

We should point out that our results are based on a few dozen queries and are by no means conclusive. Nonetheless, we hope that the results can at least give us some insight into realistic cross-lingual search situations. In Xu and Weischedel (2003) we show a technique of more intelligent sentence selection from the parallel corpus to reduce the amount of training required for the bilingual dictionary, achieving the same average precision with only twenty five percent of the parallel sentences.

Fig. 4 shows the relative contribution of the two manual bilingual dictionaries, of the HKNews parallel corpus, and of all resources combined on the TREC5C and TREC9X test materials. On both test sets, performance of the combined resource run is noticeably better than that of individual resources. The *t*-test (Hull, 1993) indicates the improvement in all cases is significant. That is, the *p*-value in all cases is smaller than 0.05. Our monolingual results were obtained using Miller et al.'s (1999) HMM monolingual retrieval system. The monolingual results form a strong baseline; they are better than the best official monolingual

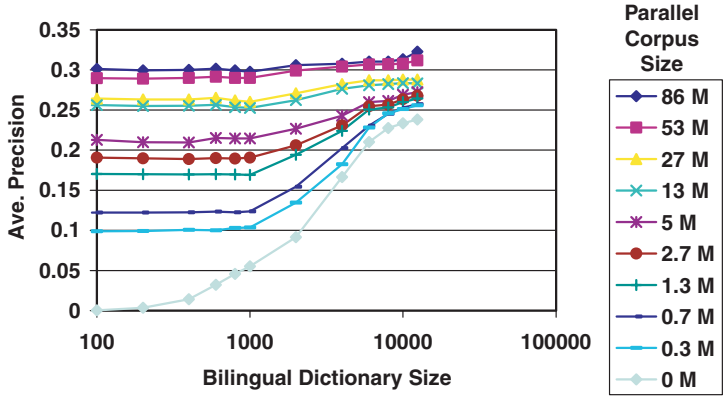

Fig. 3. English–Arabic CLIR performance as a function of sizes of a manual lexicon (Buckwalter) and a parallel corpus (UN).

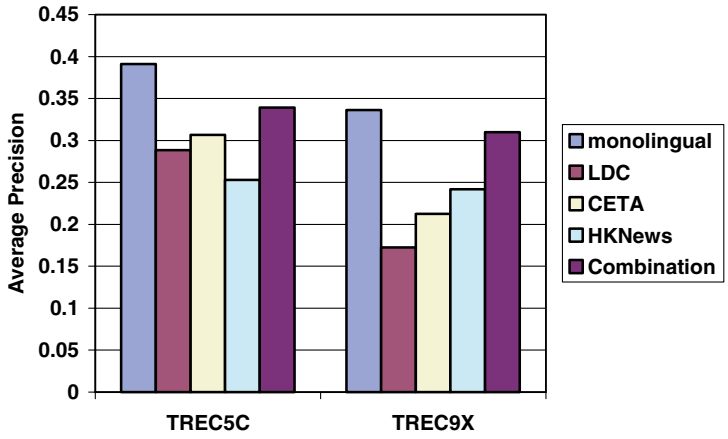

Fig. 4. Resource contributions to English–Chinese CLIR performance on TREC5C and TREC9X.

results in the TREC Proceedings (Voorhees & Harman, 1997, 2001). Both parallel corpora and manual lexicons have pros and cons. Parallel corpora can produce reliable probability estimates for frequent terms but not for infrequent ones due to data sparseness. In contrast, the flat probability distribution from manual lexicons is unreliable for frequent terms, which tend to have many translations, but is better for infrequent terms, which tend to have few translations. The complementary properties make the combination of the two types of resources essential to good CLIR. Given the strong baseline, the cross-lingual results are very impressive because they are around 90% of monolingual results (87% on TREC5C and 92% on TREC9X).

The results show that dialect similarity can also affect retrieval performance. Both the TREC9X corpus and the HKNews parallel corpus are in Cantonese (a Chinese dialect). Therefore, HKNews is more effective on TREC9X than LDC and CETA, which have a strong bias toward Mandarin (standard Chinese). On the other hand, since TREC5C is a Mandarin corpus, LDC and CETA are better than HKNews on TREC5C.

6. The usefulness of a machine translation (MT) system for CLIR

The major difference between MT-based CLIR and our approach is that the former uses one translation per term and the latter uses multiple translations. It has been suggested that CLIR can potentially utilize the multiple useful translations in a bilingual lexicon to improve retrieval performance (Klavans & Hovy, 1999). In our experiments, we used SYSTRAN version 3.0 (http://www.systransoft.com) for query and document translation. SYSTRAN is generally accepted as one of the best commercial MT systems for Spanish–English translation.

We performed four retrieval runs on the TREC5S corpus:

1. Query translation. English queries are translated to Spanish via SYSTRAN. Retrieval was performed using the translated queries on the Spanish corpus.
2. Document translation. The Spanish corpus is translated to English via SYSTRAN. Retrieval was performed using English queries on the translated corpus.
3. Combined run. The two retrieval scores for each document obtained in 1 and 2 were multiplied to produce a combined score for that document. Documents were then ranked based on the combined scores. Previous studies (McCarley, 1999) suggested that such a combination can improve CLIR performance.

Table 3
Comparing our English–Spanish CLIR system and MT-based CLIR

Monolingual	0.4275
Query translation	0.2943
Doc translation	0.3197
Doc and query translation	0.3466
Probabilistic CLIR	0.3615

4. Probabilistic CLIR. We induced a bilingual lexicon from the translated corpus by treating the translated corpus as a pseudo-parallel corpus. WEAVER was used to induce a bilingual lexicon for our approach to CLIR.

Table 3 shows that probabilistic CLIR using our system outperforms the three runs using SYSTRAN, but the improvement over the combined MT run is very small. Its performance is around 85% of monolingual retrieval. Please note that the induced lexicon is probably a trimmed version of the true lexicon in SYSTRAN. Had we had direct access to the relevant linguistic knowledge (including lexicon and disambiguation knowledge) in the MT system, we could probably make a better probabilistic bilingual lexicon than the one induced from a pseudo-parallel corpus. As a result, we could produce better retrieval performance. On the other hand, the test set has only 25 queries and the difference between our system and the combined MT run is very small. The *t*-test indicates the difference is not statistically significant. Therefore, we cannot draw a firm conclusion about the retrieval advantage of probabilistic CLIR without further study.

Nonetheless, the results suggest that a simple dictionary-based approach can be as effective as a sophisticated MT system for CLIR. This is particularly important for languages where MT may not be available, but where some parallel corpus and bilingual word lists may have been compiled. The results also suggest that pseudo-parallel texts generated by a MT system can mitigate the problem caused by the lack of a true parallel corpus. By treating MT systems as tools for generating parallel texts, the integration of seemingly disparate translation resources (i.e. manual lexicons, parallel corpora and MT systems) becomes seamless under a single retrieval model.

The goal of our experiments is not to dismiss the MT-based approach; it is viable for at least two reasons. First, it is 10 times as fast as our CLIR system in the above experiments. Even though pre-computation can improve the efficiency of our system we expect MT-based CLIR would still be faster due to a sparser term-document matrix. Second, the retrieved documents are readable by end users. These properties make it the ideal search strategy in an interactive CLIR environment. The advantage of the dictionary-based approach is also twofold. It is relatively inexpensive to build and it can potentially produce better retrieval results by using more than one translation per term.

7. Stemming and parallel corpora

For a highly inflected language such as Arabic, stemming is an active area of research. Theoretically however, if one had a large enough parallel corpus, one might see enough important inflected forms of important stems that stemming of Arabic for cross-lingual IR might not be crucial. We therefore tried to determine what strategy would be most effective empirically. All results are based on the TREC2001X corpus.

7.1. Effect of Arabic stemming in inducing a bilingual lexicon from a parallel corpus

We have compared three modes of learning term translations from the UN corpus. The first did not stem Arabic words. The second and third use the Buckwalter morphological analyzer for Arabic (Buckwalter,

Table 4
Impact of three modes of GIZA++ training on average precision for English–Arabic CLIR

Monolingual	No-stem	Sure-stem	All-stems
0.3682	0.3106	0.2994	0.2895

Table 5
Impact of resource combination on average precision for English–Arabic CLIR

Monolingual	Buckwalter only	UN only	Both
0.3682	0.2695	0.2994	0.3604

2001). The second strategy is "sure-stem", where a word is stemmed if the Buckwalter analyzer yields only one stem for the word. In the third strategy ("all-stems"), if the Buckwalter analyzer yields n stems for a word, all are used, but each has probability $1/n$. All three have pros and cons. The first keeps the maximum amount of word distinction, but requires more training data. The third requires less training data due to the reduced dimensionality, but increases word ambiguity, and the probability estimates are affected due to the one-to-many mapping from words to stems. The second is a compromise. The Buckwalter lexicon was not used in the experiments.

The retrieval scores in Table 4 shows that no-stem is slightly better than sure-stem, which is slightly better than all-stems. While the differences are too small to make firm conclusions, they suggest that Arabic stemming is not an important issue in CLIR, if a very large parallel corpus is available to induce a statistical bilingual dictionary. The t-test indicates the differences between the different stemming methods are not statistically significant.

7.2. Impact of resource combination

Table 5 shows the retrieval scores when:

- Only the Buckwalter lexicon was used for term translation.
- Only the UN corpus was used.
- Both resources were used.

Sure-stem stemming was used in the experiments. The scores indicate that the combination of the UN and the manual lexicon significantly outperforms either resource alone, suggesting that the word ambiguity problem in Arabic is satisfactorily handled by complementing a manual lexicon with a parallel corpus. The improvement over Buckwalter-only is statistically significant (t-test, p-value 0.02), but the improvement over UN-only is not. The score for the combined lexicon approaches that of monolingual. The results are consistent with the Chinese results in Section 5.

8. Related studies

There are a number of studies of CLIR performance from the perspective of available translation resources (Demner-Fushman & Oard, 2003; Franz, McCarley, Ward, & Zhu, 2001; Grefenstette, 1998; Kraaij, 2001; Kwok, Grunfeld, Dinstl, & Chan, 2001; Levow & Oard, 1999; Xu & Weischedel, 2000, 2003). A number of previous studies (Demner-Fushman & Oard, 2003; Grefenstette, 1998; Levow & Oard, 1999) explored the impact of lexicon coverage on CLIR performance. Franz et al. (2001) studied how the

quantity of parallel texts affects CLIR performance. Kraaij (2001) compared the relative utilities of manual lexicons, parallel corpora and machine translation for query translation in CLIR. This work complements existing studies in two ways. First, a more complete set of resources (e.g. bilingual lexicons, parallel texts, MT and stemming) is studied. Second, our results are based on a number of disparate languages (i.e. Spanish, Chinese and Arabic) that have little in common.

The effect of stemming on monolingual retrieval is well documented in the literature. For example, Harman (1991) studied the effect of stemming on monolingual English Retrieval. In comparison, there have been very few studies on the value of stemming for CLIR (Larkey, Ballesteros, & Connel, 2002; Xu, Fraser, & Weischedel, 2002).

There are many prior studies using Machine Translation for CLIR (e.g. Ballesteros & Croft, 1998; Gey, He, & Chen, 1999; Oard, 1998). McCarley (1999) studied both query and document translations and concluded the combination of the two translations can improve retrieval performance.

The use of statistical language modeling techniques for IR and CLIR has appeared in a number of studies (Berger & Lafferty, 1999; Hiemstra & de Jong, 1999; Miller et al., 1999; Ponte & Croft, 1998). In particular, our CLIR model is similar to the one proposed by Hiemstra and de Jong (1999). A difference is that our model makes use of corpus statistics of the query language while Hiemstra's uses the corpus statistics of the document terms.

9. Conclusions and future work

This research has explored how CLIR performance depends on the availability of linguistic resources, including manual bilingual term lists, parallel texts, MT systems and stemming. Our approach makes use of all of these resources when available. Parallel corpora are used as data for estimating term translation probabilities based on models originally published by Brown et al. (1993). Empirical results on a number of TREC test corpora show that it is possible to achieve an acceptable CLIR performance (70–80% of monolingual performance) with a sufficiently large manual lexicon. When parallel texts are used in addition to a manual lexicon, CLIR performance can rival monolingual performance. While MT systems are not necessary for effective CLIR, pseudo-parallel corpora produced by an MT system can mitigate the problem caused by the lack of true parallel texts. Our results also show that while stemmers are useful tools for monolingual IR, they are not necessary for CLIR if a large amount of parallel texts is available for learning term translations. One area to extend this work is to incorporate more types of resources, such as comparable corpora (Fung & Mckeown, 1997) and named entity translations (Al-Onaizan & Knight, 2002).

References

Al-Onaizan, Y., & Knight, K. (2002). Translating named entities using monolingual and bilingual resources. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 400–408).

Ballesteros, L., & Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the ACM SIGIR conference* (pp. 64–71).

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the ACM SIGIR conference* (pp. 222–229).

Brown, P., Della Pietra, S., Della Pietra, V., Lafferty, J., & Mercer, R. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics, 19*(2), 263–311.

Buckwalter, T. (2001). Personal communications.

Demner-Fushman, D., & Oard, D. (2003). Effect of bilingual term list size on dictionary-based cross-lingual information retrieval. In *Proceedings of Hawaii international conference on systems sciences.*

Franz, M., McCarley, J., Ward, T., & Zhu, W. (2001). Quantifying the utility of parallel corpora. In *Proceedings of the ACM SIGIR conference* (pp. 398–399).

Fung, P., & Mckeown, K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the fifth annual workshop on very large corpora, Hong Kong* (pp. 192–202).

Gey, F., He, J., & Chen, A. (1999). Manual queries and machine translation in cross-language retrieval at TREC-7. In E. M. Voorhees & D. K. Harman (Eds.), *The seventh text retrieval conference (TREC7). NIST special publication 500-242* (pp. 527–540).

Grefenstette, G. (1998). Evaluating the adequacy of a multilingual transfer dictionary for cross language information retrieval. In *Proceedings of the first international conference on language resources and evaluation* (pp. 755–758).

Harman, D. (1991). How effective is suffixing?. *Journal of the American Society for Information Science, 42*(1), 7–15.

Hiemstra, D., & de Jong, F. (1999). Disambiguation strategies for cross-language information retrieval. In *Proceedings of the third European conference on research and advanced technology for digital libraries* (pp. 274–293).

Hull, D. (1993). Using statistical testing in evaluation of retrieval experiments. In *Proceedings of the ACM SIGIR conference* (pp. 329–338).

Klavans, J., & Hovy, E. (1999). Multilingual (or cross-lingual) information retrieval. In E. Hovy, N. Ide, R. Frederking, J. Mariani & A. Zampolli (Eds.), *Linguistica Computazional: Vol. XIV–XV* (pp. 35–56). Pisa, Italy.

Kraaij, W. (2001). TNO at CLEF 2001, comparing translation resources. In *Proceedings of CLEF-2001* (pp. 78–93).

Kwok, K., Grunfeld, L., Dinstl, N., & Chan, M. (2001). TREC-9 cross language, web and question–answering track experiments using PIRCS. In E. M. Voorhees & D. K. Harman (Eds.), *The ninth text retrieval conference (TREC9). NIST special publication 500-249* (pp. 417–426).

Lafferty, J. (1999). Personal communications.

Larkey, L., Ballesteros, L., & Connel, M. (2002). Improving retrieval for arabic information retrieval: light stemming and co-occurrence analysis. In *Proceedings of the ACM SIGIR conference* (pp. 275–282).

Levow, G. A., & Oard, D. (1999). Evaluating lexical coverage for cross-language information retrieval. In *Proceedings of workshop on multilingual information processing and Asian language processing, Beijing*.

McCarley, J. (1999). Should we translate the documents or the queries in cross-language information retrieval. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 208–214).

McNamee, P., & Mayfield, J. (2002). Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the ACM SIGIR conference* (pp. 159–166).

Miller, D., Leek, T., & Schwartz, R. (1999). A hidden Markov model information retrieval system. In *Proceedings of the ACM SIGIR conference* (pp. 214–221).

Oard, D. (1998). 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 America* (pp. 472–483).

Och, F., & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the annual meeting of the Association for Computational Linguistics* (pp. 440–447).

Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR conference* (pp. 275–281).

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the ACM SIGIR conference* (pp. 21–29).

Voorhees, E., & Harman, D. (1997). In *The fifth text retrieval conference (TREC5). NIST special publication 500-238*.

Voorhess, E., & Harman, D. (2001). In *The ninth text retrieval conference (TREC-9). NIST special publication 500-249*.

Voorhees, E., & Harman, D. (2002). In *The tenth text retrieval conference (TREC 2001). NIST special publication 500-250*.

Voorhees, E., & Harman, D. (2003). In *The eleventh text retrieval conference (TREC 2002). NIST special publication 500-251*.

Xu, J., & Croft, W. B. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems, 16*(1), 61–81.

Xu, J., Fraser, A., & Weischedel, R. (2002). TREC 2001 cross-lingual retrieval at BBN. In E. M. Voorhees & D. K. Harman (Eds.), *The tenth text retrieval conference (TREC 2001). NIST special publication 500-250* (pp. 102–106).

Xu, J., & Weischedel, R. (2000). Cross-lingual information retrieval using hidden Markov models. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP), Hong Kong*.

Xu, J., & Weischedel, R. (2003). A probabilistic approach to term translation for cross-lingual retrieval. In X. Croft & X. Lafferty (Eds.), *Language modeling for information retrieval* (pp. 125–140). Kluwer.

Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a probabilistic model for crosslingual retrieval. In *Proceedings of the ACM SIGIR conference* (pp. 105–110).