CLIS-TR-98-01 LAMP-TR-014 CS-TR-3897 UMIACS-TR-98-27 April, 1998

A Comparative Study of Knowledge-Based Approaches for Cross-Language Information Retrieval*

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

Bonnie J. Dorr
Institute for Advanced Computer Studies and
Department of Computer Science
University of Maryland, College Park, MD 20742
bonnie@umiacs.umd.edu

Paul G. Hackett
College of Library and Information Services
University of Maryland, College Park, MD 20742
pghtwoz@glue.umd.edu

Maria Katsova
Department of Computer Science
University of Maryland, College Park, MD 20742
katsova@umiacs.umd.edu

Abstract

Cross-language retrieval systems seek to use queries in one natural language to guide the retrieval of documents that might be written in another. Acquisition and representation of translation knowledge plays a central role in this process. This paper explores the utility of two sources of manually encoded translation knowledge, bilingual dictionaries and translation lexicons, for cross-language retrieval. We have implemented six query translation techniques that use bilingual dictionaries, one based on lexical-semantic analysis, and one based on direct use of the translation output from an existing machine translation system; these are compared with a document translation technique that uses output from the same existing translation system. Average precision measures on portions of the TREC collection suggest that arbitrarily selecting a single translation from a bilingual dictionary is typically no less effective than using every translation in the dictionary, that query translation using an existing machine translation system can achieve somewhat better effectiveness than simple dictionary-based techniques, and that performing document translation rather than query translation may result in further improvements in retrieval effectiveness under some conditions.

*This work has been supported in part by DARPA contract N6600197C8540, the Logos Corporation, and NSF PFF IRI-9629108.

1 Introduction

As international markets and rapidly expanding trans-national information networks interact, an imperative for access to information written in many languages is becoming increasingly apparent. Cross-Language Information Retrieval (CLIR), the detection of relevant documents in one natural language using queries expressed in another, provides an important capability that can help meet that challenge [13]. Two principal lines of CLIR research have emerged, approaches which exploit explicit representations of translation knowledge (such as bilingual dictionaries or machine translation lexicons) and those which seek to extract useful translation knowledge from training corpora using representations such as cooccurrence matrices that are not designed for direct human interpretation. We refer to the approaches in the first group as "knowledge-based" and those in the second group as "corpus-based." Carbonell, et al. have reported excellent results when corpus-based techniques are evaluated on a held-back portion of the corpus from which the translation knowledge was extracted [4]. On the other hand, we have previously investigated the retrieval effectiveness of corpus based CLIR and found that domain shift effects can adversely affect retrieval effectiveness when translation knowledge is acquired from one corpus and then used for retrieval from a different collection [11]. Knowledge-based techniques that exploit resources such as dictionaries and machine translation lexicons are less sensitive to this effect, so their use may be preferred when domain-specific training corpora are not available. Rather than focus further on corpus-based techniques, in this paper we explore the performance of several knowledge-based CLIR approaches.

There are four fundamental strategies for knowledge-based CLIR: direct matching of terms in different languages without translation, translation of each query into every document language, translation of each document into every possible query language, and translation of each query and each document into a single common language. Cognate matching, in which knowledge about related word forms in a pair of languages is encoded directly into the query-document matching algorithm, is an example of the first strategy (c.f., [3]), and controlled vocabulary retrieval in indexing and search terms are chosen from a domain-specific multilingual thesaurus is an example of the last strategy (c.f. [15]). Cognate matching depends on semantically meaningful lexical regularities that are presently known in only a few language pairs; thus, achieving broad coverage with fully automated controlled vocabulary techniques has proven to be difficult. So we have chosen to focus this study only on the query translation and document translation strategies.

Over the past several years, query translation has emerged as the most popular strategy for fully automatic broad coverage CLIR [12]. Query translation can be quite efficient when short queries are presented, but simple query translation approaches suffer a severe penalty in effectiveness, usually achieving about half of the retrieval effectiveness of corresponding monolingual techniques when typical measures such as average precision are used. A number of studies have reported that simple linguistic processing such as limiting candidate translations for query terms to those with the same part of speech, or indexing phrases as well as individual words, can raise this performance to perhaps 75% of the monolingual effectiveness (c.f., [5, 10]). In this paper we describe a new query translation technique based on Lexical Conceptual Structures (LCS) and compare it to an alternative technique based on the output of an existing Machine Translation (MT) system; these two are furthermore compared to some more efficient dictionary-based query translation techniques.

A document translation strategy in which fully automatic MT is used to translate each document into a single language (the query language) at indexing time may be attractive for interactive applications if the users need to rapidly skim retrieved documents in their preferred language. This is a requirement that query translation strategies could not presently support (translation rates are a minute or so per page on typical workstations), but a document translation strategy in which the full text of the translations is retained would be well suited for this. Document translation may also improve retrieval effectiveness if the MT system is able to exploit linguistic context to choose correct translations more often in documents than in queries. Since queries are sometimes quite short and are often not well formed sentences, there is some reason to believe that this improvement may be achievable. We have tested this hypothesis by implementing a document translation technique and comparing it with several query translation techniques.

In addition to the query translation and document translation techniques, we also implemented two baseline techniques without any translation component: query construction in the same language as the documents, and the presentation of queries in a language different from that of the documents. We expect the first to provide an upper bound for CLIR effectiveness and the second to provide a lower bound. The lower bound is important because proper names, foreign language terms embedded in the documents, and words with the same written form in the each language can result in fortuitous cognate matches that might, if undetected, produce the impression of better performance from a CLIR technique than would be justified.

The next section presents our experiment design and describes the techniques we have implemented in detail. We have learned that arbitrarily selecting a single translation from a bilingual dictionary can be as effective as more commonly implemented techniques based on retaining every possible translation, that techniques based on loosely coupling machine translation and information retrieval perform somewhat better than simple dictionary based techniques, that there is reason to believe more tightly coupled techniques could perform even better, and that document translation can outperform query translation under some conditions. Section 3 describes these results in detail, and the paper concludes with a discussion of the implications for further work on cross-language information retrieval.

2 Experiment Design

Earlier CLIR evaluations have been hampered by inadequate test collections, but the Text REtrieval Conference (TREC-6) recently developed the first large-scale multilanguage collection that is designed specifically to support CLIR experiments. We have used the German documents ("SDA/NZZ") from that collection for the majority of our experiments, supplemented where necessary by the corresponding English collection ("AP") and an earlier Spanish collection from TREC-4 ("El Norte") for which English queries are available.

The TREC-6 CLIR SDA/NZZ collection contains 251,840 German newswire articles from two Swiss news agencies. The SDA documents are from 1988, 1989 and 1990, and the NZZ documents are from 1994. Standard topics are described in German, English, and three other languages, and relevance judgments were made by at the National Institutes of Standards and Technology (NIST) using a pooled assessment methodology in which each of the top one hundred documents from four different monolingual German retrieval systems were evaluated for relevance to a topic description similar to that in Figure 1. Documents not in that set were presumed not to be relevant for the purpose of computing recall and precision. The process was repeated for 22 topics and relevant documents were discovered for 21 of those topics.[1] Three standard sources for query terms were defined at TREC-6: "title queries" are formed from the one to three words in the "title" field, "short queries" are formed from only the one or two sentences or sentence fragments in the "desc" field, and "long queries" are formed using every word in the "title," "desc," and "narr" fields. We report results below for title queries and long queries.[2]

The TREC-6 CLIR AP collection contains 242,918 articles from the Associated Press newswire service in the United States that were generated in 1988, 1989 and 1990. The collection has been assessed at NIST for the same 22 topics using a pooled assessment methodology based on the top one hundred documents from five different monolingual English retrieval systems. Relevant documents are known in the AP collection for the same 21 topics as for the SDA/NZZ collection.

The TREC-4 El Norte collection contains 57,780 Spanish newswire articles from a Mexican news service that were generated in 1994. The collection has been assessed at NIST for 25 topics using a pooled assessment methodology based on the top one hundred documents from 10 different monolingual Spanish retrieval systems. Topic descriptions in TREC-4 lacked "title" and "narr" fields, so only "short" queries can be constructed for the TREC-4 El Norte collection. The original topic descriptions are in Spanish and human-prepared English translations of the topic descriptions are available. When two English translations were provided with the TREC-4 El Norte collection, we chose the first one which was generally the more direct (although frequently somewhat awkward) translation.

For text retrieval we ran version 3.1p1 of the Inquery system from the University of Massachusetts on a single SPARC 20 under the Solaris 2.5 operating system. The Inquery "kstem" stemmer and the standard English Inquery stopword list were used when processing the AP documents and when processing English translations of the SDA/NZZ documents. The Inquery Spanish stemmer and Spanish stopword list were used

[1]The TREC-6 CLIR evaluation originally included 25 topics. No relevant documents were discovered in the SDA/NZZ collection for topic CL22, and relevance judgments are not yet available for topics CL03, CL15 and CL25.

[2]We omit results for short queries on the SDA/NZZ collection because we discovered at TREC-6 that the "desc" field often fails to perform well in monolingual German evaluations, presumably because the topic descriptions were constructed by amplifying rather than repeating earlier information [13].

```
<top>
<num> Number:  CL1

<E-title> Waldheim Affair

<E-desc> Description:  Reasons for controversy surrounding Waldheim's World
War II actions.

<E-narr> Narrative:  Revelations about Austrian President Kurt Waldheim's
participation in Nazi crimes during World War II are argued on both sides.
Relevant documents are those that express doubts about the truth of these
revelations.  Documents that just discuss the affair are not relevant.
</top>
```

Figure 1: The English version of TREC-6 topic CL01.

Approach	Test Collection (Document Language)		
	SDA/NZZ (German)	AP (English)	El Norte (Spanish)
Same Language Query	X	X	X
Dictionary-based Query Translation	X	X	X
MT-based Query Translation	X	X	X
LCS-based Query Translation			X
MT-based Document Translation	X		
Foreign Language Query	X	X	X

Table 1: Summary of CLIR approaches.

when processing the Spanish El Norte documents. No stemmer or stopword list was used when processing SDA/NZZ documents in the original German, and no techniques for splitting German compounds were implemented.[3]

Table 1 summarizes the six CLIR approaches that we have implemented. The "X" marks identify the cases for which experimental results are reported in Section 3.

2.1 Same Language Query (SLQ)

To approximate an upper bound for the performance of any CLIR system, we compared the retrieval effectiveness of our four experimental approaches with the retrieval effectiveness achieved by using queries that are given in the same language as the documents. For example, the CL01 "title query" would be presented as `[waldheim affair]` when retrieving English AP documents and as `[die affaire waldheim]` when retrieving German SDA/NZZ documents.

2.2 Dictionary-Based Query Translation (DQT)

By far the most commonly used query translation approach is to replace each query term with appropriate translations that are automatically extracted from an online bilingual dictionary (c.f., [10, 2]). For translating queries from English into German for retrieval from the SDA/NZZ collection we used an online bilingual

[3]We tried a small German stopword list in our TREC-6 experiments and found that it hurt average precision somewhat in most cases [13].

dictionary developed by Stefan Büdenbender.[4] That dictionary contains 131,274 bilingual pairs in which each pair consists of one word or phrase in English and the corresponding word or phrase in German. The number of unique words in the dictionary is far smaller than 131,274 because many words appear in several bilingual pairs and the number of unique stems is smaller still because the dictionary contains multiple morphological variants for many of the words. The pairs were initially sorted in lexicographic order based on the English terms and we used the same dictionary to translate queries from German into English for retrieval from the AP collection after resorting the pairs by the German terms.

For translating queries from English into Spanish we used a Spanish-English bilingual dictionary that was produced specifically for this evaluation from a lexicon that had originally been developed for a foreign language tutoring application [7, 16]. The original lexicon contained 12,885 unique Spanish stems corresponding to 171,164 morphological variants and 29,360 bilingual pairs. We used a two-level Kimmo-based morphology system [1] to generate all morphological variants of terms matching the English terms (stemmed and unstemmed) for the subset of the topics that we processed in the El Norte collection.

It is common for a single word to have several translations, some with very different meanings. Bilingual dictionaries typically seek to help users select appropriate translations of individual words by embedding the word in a representative phrase. It is not at all clear how one should design an algorithm to extract only the "appropriate" translations using this information, so we have implemented six simple dictionary-based query translation techniques that together explore the effects of winner-take-all, word-match and stem-match approaches. We illustrate the effect of each technique with a German translation of the English CL01 title query given above.

Single Word (SW) The first exact single whole-word match in the dictionary.[5]
`[waldheim affäre]`

Single Word, Stemmed (SWS) The first exact single whole-word match if present, otherwise the first exact single stem match.[6]
`[waldheim affäre]`

Every Word (EW) Every exact single whole-word match in the dictionary.
`[waldheim affäre angelegenheit ereigneis geschäft handlung sache]`

Every Word, Stemmed (EWS) Every exact single stem match in the dictionary.
`[waldheim affäre angelegenheit angelegenheiten ereigneis geschäft handlung sache]`

Every Phrase (EP) Every exact whole-word match in the dictionary, regardless of whether the word appears alone or as part of a phrase.

```
[waldheim affäre angelegenheit ereigneis geschäft handlung sache ehrensache
familienagelegenheit liebesglück es war eine abgekartete sache es ging heiss
her liebesaffäre liebeserlebnis techtelmechtel staatsangelegenheit das ist
meine sache]
```

Every Phrase, Stemmed (EPS) Every exact stem match in the dictionary, regardless of whether the stemmed word appears alone or as part of a phrase.

```
[waldheim affäre angelegenheit angelegenheiten ereigneis geschäft handlung
sache ehrensache familienagelegenheit liebesglück es war eine abgekartete sache
es ging heiss her liebesaffäre liebeserlebnis techtelmechtel mein
privatangelegenheiten staatsangelegenheit staatsangelegenheiten bescherung das ist
meine sache seine angelegenheiten in ordnung bringen geschäfte abwickeln]
```

In every case we replace each word in the query with the corresponding word or phrase in every matching bilingual pair to produce a version of the query that can be compared with the documents in the collection.

[4] The dictionary we used is freely available at http://www.bg.bib.de/~a2h6bu/, and our query translation code will be available shortly at http://www.glue.umd.edu/~oard/research.html

[5] An "exact" match is one in which the two character strings are the same length and each character in the two strings matches and a "whole word" is a string of characters that appear in the document.

[6] We used the Porter stemmer for English that is available from ftp://ftp.vt.edu/pub/reuse/IR.code/ for this purpose.

Words that appear in the standard English Inquery stopword list are not translated and thus do not affect the translated query, but words that do not match any dictionary entry are included unchanged in the translated query. In addition to simple word-to-word mappings, word-to-phrase mappings are possible (and, in fact, common), so translated queries are typically longer than untranslated queries and they sometimes contain repeated words. Because our dictionaries are sorted in alphabetical order rather than with regard to the predominance of a give translation within a known domain, the semantic effect of techniques SW and SWS are likely to be close to that achieved by random selection of a single translation from the sets produced in techniques EW and EWS respectively.

2.3 MT-Based Query Translation (MTQT)

Machine translation systems seek to translate documents from one language to another, either as an aid for human translators or for direct use as a fairly rapid and inexpensive rough translation. This provides an obvious approach to query translation, but we are aware of only one prior experiment to use such a technique [14]. In that experiment, Radwan and Fluhr compared the retrieval effectiveness of queries translated from French into English by the SYSTRAN machine translation system with the effectiveness of their EMIR dictionary-based query translation system using a version of the small Cranfield collection for which French queries were available. In that study they found that the EMIR was more effective than their MT-based query translation technique using SYSTRAN. Our experiments offer some insight into the performance of a MT-based query translation approach on larger test collections.

The Logos machine translation system that we used for our experiments is a commercial product that is designed to assist human translators by automatically preparing fairly good translations of individual documents.[7] The system is typically used by translation bureaus and other organizations as the first stage of a machine-assisted translation process, and we have previously used it for cross-language routing experiments [11]. The Logos system includes extensive facilities for adding domain-specific technical terminology and new linguistic constructs, but for the experiments reported here we used only the machine readable dictionaries and semantic rules that are delivered as standard components of the product.

We used the Logos system to translate English queries into German for use with the SDA/NZZ collection, to translate German queries into English for use with the AP collection, and to translate English queries into Spanish for use with the El Norte collection. Since the Logos system is designed to generate readable translation, it generates only a single "best guess" translation for any input. Thus MTQT is most similar to the DQT-SW technique in which a single candidate translations is retained.

2.4 LCS-Based Query Translation (LCSQT)

Lexical conceptual structures are automatically constructed linguistic representations that are based on lexicalized regularities that reveal meaningful semantic relationships. Our LCS-Based query translation approach involves the construction of disambiguated (target-language) queries from event-based entries in our lexicon. The first stage of this approach involves a sentence analysis component that builds a syntactic structure produced by a parser called REAP (Right Edge Adjunction Parser) [16]. For example, the parse tree produced for the sentence "What are Mexico's attitudes toward press censorship" has the following structure:

```
[CP What_i
  [S are
    [NP mexico [N attitudes [PP toward [NP press censorship]]]]
    [VP e_i]]]
```

The next stage of query translation involves the construction of a language-independent, compositional representation called Lexical Conceptual Structure (LCS) [6, 8]. For example, the LCS representation for the verb "be" is:

```
(be ident (* thing x) (at ident (thing x) (* thing y)))
```

[7] Logos Corporation, 111 Howard Boulevard, Suite 214, Mount Arlington, NJ 07856 USA

This LCS is uninstantiated, i.e., it has unfilled argument positions (as indicated by the * marker). During the process of LCS composition, argument positions are filled. For example, the sentence above would correspond to the following composed representation:

```
(be ident (attitude (mexico (toward (censorship (press)))))
          (at ident (attitude (mexico (toward (censorship (press)))))
                    (wh-thing)))
```

We have developed a technique for representing instantiated LCS forms as queries in the Parka-DB knowledge representation system [9]. Parka-DB provides an efficient technique for matching graph structures that we use to generate the terms for the target-language query. The system produces a collection of terms in the target language based on the structure of the composed LCS. The scalability of the Parka-DB system allows us to represent large lexicons for the languages of interest. The generation of target-language terms entails lexical selection from the composed LCS associated with each event-based term. We have not yet implemented some of the components necessary to produce LCS representations for German; thus, we performed a preliminary evaluation of LCSQT using topics SP34, SP35 and from the TREC-4 El Norte collection. For example, the English short query for topic SP45 is:

`Mexico's attitudes toward press censorship`

The LCS for this query would be:

`(attitude (mexico (toward (censorship (press)))))`

and the Spanish terms generated for this LCS are:

`[actitud méxico hacia censura pulse prensa]`

For comparison, the official Spanish version of the SP45 short query is:

`Actitudes en México sobre la censura de la prensa`

2.5 MT-Based Document Translation (MTDT)

Our MT-based document translation approach parallels the design of our MTQT design. We have selected English as a query language and translated each SDA/NZZ document into English as a preprocessing step. We then indexed the translated document collection and used English queries for the retrieval experiments. Essentially the preprocessing step reduces cross-language retrieval to a (possibly degraded) monolingual case. We used four SPARC 20 workstations and a fifth workstation that was upgraded from a SPARC 5 to a SPARC Ultra 1 after about three quarters of the documents had been translated.[8] Translation of the 48 months of newswire stories contained in the SDA and NZZ collections using these machines required approximately 10 machine-months, and successful translations were obtained for 251,572 documents. The remaining 268 documents were omitted from the translated collection.

2.6 Foreign Language Query (FLQ)

Monolingual information retrieval systems sometimes produce useful results because of fortuitous matches between words in different languages, proper names that are rendered in the same way in different languages, and foreign language terms in the documents that happen to be in the query language. For example, the English version of the CL01 title query shown above contains the proper name "Waldheim" which also often appears in relevant German documents. In order to establish a practical lower bound on retrieval effectiveness we have used both untranslated queries and untranslated documents to reveal the effect of these cognate matches.

3 Results

Table 2 summarizes the non-interpolated average precision results for the SDA/NZZ collection, using every technique except LCSQT, averaged over the 21 topics for which relevant documents are known. For title

[8] The translated documents are available to TREC participants from NIST.

Technique	Query Length	
	Title	Long
SLQ	0.2480	0.2396
DQT-SW	0.1749	0.1342
DQT-SWS	0.1542	0.0969
DQT-EW	0.1778	0.1312
DQT-EWS	0.1363	0.0827
DQT-EP	0.1152	0.0165
DQT-EPS	0.1172	0.0182
MTQT	0.1668	0.1561
MTDT	0.1761	0.2171
FLQ	0.0307	0.0117

Table 2: Non-interpolated average precision for the SDA/NZZ collection, averaged over 21 topics.

Technique	Query Length	
	Title	Long
SLQ	0.3449	0.3958
DQT-SW	0.1982	0.1154
DQT-EW	0.1805	0.0710
MTQT	0.1928	0.2455
FLQ	0.0105	0.0132

Table 3: Non-interpolated average precision for the AP collection, averaged over 21 topics.

queries the advantage of SLQ over four of the eight CLIR techniques is statistically significant (with 95% confidence), as is the difference between 3 of the CLIR techniques and FLQ, but the available 21 queries are not sufficient to detect statistically significant differences among the CLIR techniques that we have tested. It does appear, however, that DQT-SW is no worse than the more commonly implemented DQT-EW technique, and that the same pattern is evident in the stemmed variant of each technique and with long query as well. MTQT and MTDT also appear to work well in this experiment, with a slight edge perhaps going to MTDT.

In order to seek confirmation for these results we applied three of our CLIR techniques to the English AP collection. Table 3 summarizes the non-interpolated average precision results for that collection, using DQT-SW, DQT-EW and MTQT. We observe the same trends on the AP collection as on the SDA/NZZ collection, finding that DQT-SW is no worse than DQT-EW and that MTQT performs somewhat better than either of those techniques. Thus although we have not obtained statistically significant results we have strong reason to believe that our most important observations are repeatable.

Table 4 summarizes the non-interpolated average precision results for the El Norte collection, using every technique except MTDT. Our English parser is still under development, so we processed only the three (of 25) topics in this collection that our parser was able to handle. While this is an inadequate sample to obtain statistically significant results, we did obtain average precision better than that achieved by any DQT technique and comparable to that achieved by MTQT on both queries for which any retrieval technique produced credible results.[9] From this we conclude that further investigation of the LCSQT approach is justified.[10]

[9] Very few relevant documents are known for topic SP35.

[10] Interestingly, the average precision of LCSQT surpassed that of even SLQ on topic SP34. In this case the difference

Technique	Topic			
	SP34	SP35	SP45	Average
SLQ	0.1762	0.0114	0.1875	0.1250
DQT-SW	0.1270	0.0001	0.1015	0.0762
DQT-SWS	0.0887	0.0000	0.0944	0.0611
DQT-EW	0.1981	0.0002	0.0081	0.0688
DQT-EWS	0.0373	0.0003	0.0309	0.0152
DQT-EP	0.1905	0.0002	0.0081	0.0663
DQT-EPS	0.0365	0.0001	0.0079	0.0148
MTQT	0.1288	0.0000	0.1699	0.0996
LCSQT	0.2398	0.0086	0.1448	0.1310
FLQ	0.0000	0.0000	0.0000	0.0000

Table 4: Non-interpolated average precision for three queries in the El Norte collection.

4 Conclusions

We have conducted an extensive evaluation of eight cross-language information retrieval techniques and found some interesting results. It is clearly possible to exploit these same resources that we have used for DQT in these experiments it is clearly possible to craft more sophisticated techniques. For example, we could take advantage of redundancy in the dictionary to improve our translation choices in the DQT-SW method. And in MTQT and MTDT we could preserve some additional terms in the face of unresolvable ambiguity by coupling the translation and retrieval systems more tightly. But we have shown that document translation is a practical approach for cross-language text retrieval on moderately large collections, that MT-based query translation performs well, and that arbitrary translation selection appears to work as well as any other technique for dictionary-based query translation. As cross-language test collections improve these results should provide a sound basis for further research on knowledge-based techniques for cross-language information retrieval.

Acknowledgments

The authors are grateful to Wade Shen for his help with parsing and LCS composition, Scott Bennett and Harriet Leventhal for their assistance with the Logos translation system, the University of Massachusetts for the use of Inquery, James Allan for help with Inquery configuration, and Fred Gey for making us aware of the German dictionary that we used.

References

[1] E.L. Antworth. *PC–KIMMO: A Two-Level Processor for Morphological Analysis.* Dallas Summer Institute of Linguistics, 1990.

[2] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 1997.

[3] Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie. Using clustering and SuperConcepts within SMART: TREC 6. In *The Sixth Text REtrieval Conference (TREC-6).* National Institutes of Standards and Technology, November 1997. To appear.

appears to result from LCSQT using the more common term “fuerza” in place of “fortalezas” and two terms, “tierra” (land) and “ejército” (armed forces) as the translation of “army,” rather than thee single, more general term “ejército.”

[4] Jamie Carbonell, Yimying Yang, Robert Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. Translingual information retrieval: A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, August 1997.

[5] Mark Davis and William C. Ogden. Quilt: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 1997.

[6] Bonnie J. Dorr. *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge, MA, 1993.

[7] Bonnie J. Dorr. Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring. In *Proceedings of the ACL Fifth Conference on Applied Natural Language Processing (ANLP)*, pages 139–146, Washington, DC, 1997.

[8] Bonnie J. Dorr and Mari Broman Olsen. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 151–158, Madrid, Spain, July 7-12 1997.

[9] M. Evett, J. Hendler, and L. Spector. Parallel knowledge representation on the connection machine. *International Journal of Parallel and Distributed Computing*, 22, 1994.

[10] David A. Hull and Gregory Grefenstette. Experiments in multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996. http://www.xerox.fr/people/grenoble/hull/papers/sigir96.ps.

[11] Douglas W. Oard. Adaptive filtering of multilingual document streams. In *Fifth RIAO Conference on Computer Assisted Information Searching on the Internet*, June 1997. http://www.glue.umd.edu/~oard/research.html.

[12] Douglas W. Oard. Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence, March 1997. http://www.glue.umd.edu/~oard/research.html.

[13] Douglas W. Oard. Serving users in many languages: Cross-language information retrieval for digital libraries. *D-Lib Magazine*, December 1997. http://www.dlib.org.

[14] Khaled Radwan and Christian Fluhr. Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 121–136, April 1995.

[15] Dagobert Soergel. Multilingual thesauri in cross-language text and speech retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence, March 1997. http://www.clis.umd.edu/dlrg/filter/sss/papers/.

[16] Amy Weinberg, Joseph Garman, Jeffery Martin, and Paola Merlo. Principle-Based Parser for Foreign Language Training in German and Arabic. In Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors, *Intelligent Language Tutors: Theory Shaping Technology*, pages 23–44. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.