Evaluating Resources for Query Translation in Cross-Language Information Retrieval

Bonnie J. Dorr and Douglas W. Oard
Department of Computer Science and Institute for Advanced Computer Studies (first author)
College of Library and Information Services (second author)
University of Maryland, College Park, MD, USA 20742
{bonnie,oard}@umiacs.umd.edu

Abstract

Our goal is to evaluate the utility of a lexical resource containing Lexical Conceptual Structures (LCS) for use in cross-language information retrieval. Our evaluation makes use of a combination of techniques from interlingual machine translation with conventional information retrieval techniques. Given a query in one language, we transform the query into the corresponding terms in a second language. Our focus is on the construction of disambiguated target-language queries by using verb-based entries in our lexicon to construct Lexical Conceptual Structures. The main innovation of LCS-based query translation is that it provides a principled framework for controlling translation ambiguity in cross-language information retrieval applications. We evaluate this approach by comparing the resulting retrieval effectiveness with alternative techniques based that use off-the-shelf machine translation or translation knowledge extracted from machine readable dictionaries. We view this work as an important demonstration of techniques for evaluating the utility of alternative structures for multilingual lexicons in cross-language IR applications. In addition, this work provides a basis for measuring the extent to which disambiguation can enhance cross-language information retrieval effectiveness.

1 Introduction

We have constructed a large database of lexical entries for English, Spanish, and Arabic using a combination of automatic and semi-automatic techniques. This lexicon presently contains approximately 60,000 entries per language, and each entry consists of a linguistically-motivated representation called Lexical Conceptual Structure (LCS). We have previously used this lexicon for experiments in interlingual Machine Translation (MT) [4] and foreign language tutoring [5, 17]. Our goal in this paper is to evaluate the utility of this representation for Cross-Language Information Retrieval (CLIR), an information search task in which the query may be posed in a natural language that is different from that used in the documents [10, 13]. We view these preliminary results as a useful step toward establishing the utility of our LCS-based lexicon as a large-scale lexical resource for a variety of tasks that involve more than one language.

Query translation has emerged as a popular strategy for fully automatic broad coverage CLIR [12]. The key idea is to leave the documents in their original language and to transform each query into every language at run time. Query translation is efficient when short queries are presented, but unsophisticated techniques based on simple word substitution adversely effect retrieval effectiveness when compared with monolingual information retrieval scenarios. This adverse impact appears to result from translation ambiguity, limited lexical coverage, and a failure to correctly translate noncompositional phrases [12]. Fairly 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, has been shown to improve retrieval effectiveness somewhat (c.f., [3, 9]), but there appears to be room for further improvement.

We have developed an LCS-based Query Translation (LQT) technique that uses LCS representations as a basis for selectional restrictions and applied that technique to perform query translation. The main innovation of this technique is that it provides a framework for dealing with the translation ambiguity problem. The focus of our initial ex-

periments has been construction of disambiguated (target-language) queries from verb-based entries in our lexicon. We have evaluated the effectiveness of our approach by translating sixteen queries from English into Spanish, using an off-the-shelf text retrieval system to develop a ranked list of the documents best matching each translated query, comparing that list with the set of documents that have been judged to be relevant to that query, and reporting standard effectiveness measures such as recall and precision. In order to establish a basis for comparison, we implemented two additional approaches: MT-based query translation (MQT) and dictionary-based query translation (DQT). For MQT we translated each query into Spanish using an off-the-shelf MT system. DQT was accomplished by replacing each query word with appropriate translation(s) from a simple bilingual term list.

2 Experiment Design

The Text REtrieval Conferences (TREC) have developed large-scale collections that can support CLIR experiments. We have evaluated our LCS database as lexical resource for CLIR using the Spanish collection from TREC-4 for which 25 Spanish queries and manually prepared English translations of those queries are available. More than one group prepared English translations for these queries, and when two English translations were provided with the collection, we chose the second one which was typically closer to that which a native speaker of English might have constructed in our opinion. The collection contains 57,780 Spanish articles from the Mexican newspaper "El Norte" that appeared in 1994. Relevance judgments were constructed at the U.S. National Institutes of Standards and Technology (NIST) using a pooled assessment methodology in which the top one hundred documents for each query from each of 10 different monolingual Spanish retrieval systems were judged for relevance that query. For text retrieval we ran version 3.1p1 of the Inquery system from the University of Massachusetts on a single Sun SPARCstation 20 using the Solaris 2.5 operating system. The Spanish stemmer and stopword list delivered with Inquery were applied to both the collection and the queries.

The next sections describe the techniques that we implemented.

3 LCS-Based Query Translation (LQT)

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) [17]. 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) [4, 6]. For example, the LCS representation for the verb "be" is:

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

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 [7]. 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.

Our evaluation of LQT is based on topics SP26-50 from the TREC-4 El Norte collection.[1] For example, the English 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
```

3.1 MT-Based Query Translation (MQT)

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 [15]. 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.[2] 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 the 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.

3.2 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., [9, 2]). We implemented a similar process using a Spanish-English bilingual term list produced specifically for this evaluation from a more sophisticated lexicon that had originally been developed for a foreign language tutoring application [5, 17]. The original lexicon contained 12,885 unique Spanish stems corresponding to 30,555 bilingual pairs, where each pair contains a morphological variant of a Spanish word coupled with its English counterpart. We used a two-level Kimmo-based morphology system [1] to generate all morphological variants of every Spanish term. Our generation of English morphology is not yet fully automated, so we manually changed the form of each English term to match the morphological variant present in the English queries, duplicating the set of translations as necessary if more than one form was present in the queries.

It is common for a single word to have several translations, some with very different meanings. It is not yet clear how one should design an algorithm to extract "appropriate" translations from a simple bilingual term list. In recent experiments between English and German, for example, we learned that selecting the first exact translation (in lexicographic order) was as effective as selecting every translation of an exactly matching word, but that applying the Porter stemmer to the query terms and the English side of the bilingual term list typically reduced retrieval ef-

[1] Of these queries, we were able to achieve full syntactic and semantic analysis for 16 cases. The remaining 9 cases were not analyzable by the REAP parser. Modifications to REAP are currently underway to accommodate the syntactic phenomena that occur in those sentences. The queries that were handled were: SP26, SP27, SP28, SP29, SP33, SP34, SP35, SP36, SP37, SP40, SP41, SP44, SP45, SP48, and SP49.

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

fectiveness [14]. We took this opportunity to rerun that experiment between English and Spanish, implementing four DQT techniques.[3] We illustrate the effect of each technique with a Spanish translation of query SP45:

Single Word (DQT-SW) The lexicographically first exact whole-word match in the dictionary.[4]
`[mexicos actitud acosar censura]`

Every Word (DQT-EW) Every exact whole-word match in the dictionary.
`[ mexicos actitud adelantamiento adelantamientos adelanto adelantos acosar apinarse apurar importunar planchar prensa prensar pulsar pulse urgir censura excentricidades]`

Every Word, Stemmed (DQT-EWS) Every exact stem match in the dictionary.[5]
`[ mexico actitud adelantamiento adelantamientos adelanto adelantos actitud adelantamiento adelantamientos adelanto adelantos acosar apinarse apurar importunar planchar prensa prensar pulsar pulse urgir apretad urgente censura excentricidades ]`

In every case we replaced each word in the query with the corresponding word or words in the selected bilingual pair(s) to produce a version of the query that can be compared with the documents in the collection. Words that appear in the standard English Inquery stopword list were not translated and thus did not affect the translated query. Ten query words failed to match in our initial bilingual term list, but we added those words.[6] Our results should thus be interpreted as an upper bound on the performance of the DQT techniques that we have implemented.

[3] Bilingual term lists that are constructed using bilingual dictionaries often include phrase translations as well. We were limited in the current experiment to single-word entries.

[4] An "exact" match is one in which the two character strings are the same length and each character in the two strings matches. A "whole word" is any white-space delimited string of characters. Words were sorted in lexicographic order using the Latin-1 (ISO 8859-1) character set.

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

[6] With the exception of *folklore* and *ballet*, all missing query words were proper names (e.g., *Aztec*) that were specific to the document collection.

3.3 Upper Bound: Same Language Query (SLQ)

To approximate an upper bound on CLIR effectiveness, we compared the retrieval effectiveness of our three experimental approaches with the that achieved using the original Spanish queries. For example, query SP45 would be presented as `[Actitudes en México sobre la censura de la prensa]`. Because particularly fortuitous word choices can dramatically improve retrieval results, CLIR techniques do occasionally outperform SLQ. But when averaged over several queries, SLQ provides a fairly reliable upper bound on CLIR performance.

3.4 Lower Bound: 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 query SP28 contains the proper name "China," which also could be expected to appear in relevant Spanish documents. In order to establish a practical lower bound on retrieval effectiveness we have used English queries directly to retrieve Spanish documents to reveal the performance that could have been achieved by relying solely on these cognate matches. For example, query SP45 would be presented as `[Mexico's attitudes toward press censorship]`.

4 Results

Precision is defined as the fraction of the documents in a retrieved set (e.g., the first 10 retrieved documents) that were judged to be relevant by the assessor, and recall is defined as the fraction of the documents that were judged by the assessor as relevant that are included in the set. Ranked retrieval systems such as Inquery seek to place the documents with the greatest probability of relevance towards the top of the ranked list, so larger retrieved sets typically achieve greater recall at the expense of lower precision. Since some users require high precision while others require high recall, it is common to report the precision that is achieved at several levels of recall[16]. Figure 1 depicts the performance of the techniques

that we implemented. [7]

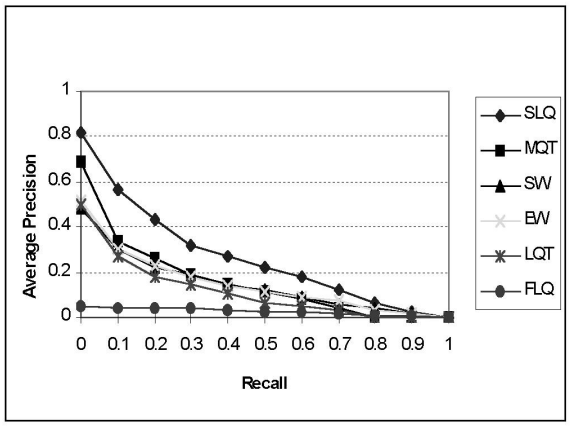

Figure 1: Precision-recall graph for the El Norte collection, averaged over 16 TREC-4 queries.

	Average Precision
SLQ	0.2494
MQT	0.1416
DQT-EW	0.1366
DQT-SW	0.1364
LQT	0.1050
DQT-EWS	0.0926
FLQ	0.0232

Table 1: Non-interpolated average precision doe the El Norte collection, averaged over 16 TREC-4 queries.

As expected, Figure 1 shows that SLQ outperforms every other technique and that every other technique outperforms FLQ. MQT appears to do slightly better than any other CLIR technique at lower levels of recall (i.e., it achieves a greater density of relevant documents near the top of the ranked list), and LQT appears to do slightly worse at moderate recall. When a representative set of queries are used, statistical significance tests offer some insight into whether such differences in retrieval results can be expected to extend reliably to other similarly constructed queries [8]. It is convenient to characterize the performance of each technique using a single number when performing statistical significance tests, and the average precision over all values of recall is commonly used for this purpose. This is equivalent to the area under the curve that would have been obtained without interpolation at the 11 recall points shown in Figure 1. Table 1 shows the average precision figures obtained in this way, averaged over the 16 queries that we used. A student-T test of these means revealed that among the CLIR techniques only the difference between DQT-EW and DQT-EWS was significant at the 0.05 level, a result that is consistent with our earlier observations between English and German [14]. While not statistically significant at that level, a student-T test also revealed that there was less than a 13% chance that the differences observed between MQT and LQT effectiveness reflect the specific set of queries that were chosen rather than the typical performance of the two techniques on this collection with any set of similarly constructed queries.

Another way to visualize the results is to depict the relative performance of two techniques for each query. Figure 2 shows a bar graph in which the average precision achieved by DQT-EW on each query is used as a reference point to define the horizontal axis and the difference between that value and the average precision achieved by LQT (or MQT) on that query is used to define the height of the associated bar. Queries for which no difference was observed result in no bar, and those for which DQT-EW outperformed LQT (or MQT) are shown below the baseline. It appears from Figure 2 that four of the queries (SP39, SP40, SP41 and SP44) are particularly challenging for both LQT and MQT. This sort of failure analysis can reveal deficiencies in the lexical resources that would not be detectable using metrics that are averaged over a set of queries.

We believe that the poor performance of LQT in these early experiments can be explained by the nature of the queries in the collection that we chose. The LCS representation is an event-oriented structure—one that would allow strong selectional restrictions to be applied to translations of terms in event-based queries. The queries in the collection we used, however, contain primarily non-event noun phrases and non-content words such as *be*). The comparison between DQT and LQT thus reveals more about the relative coverage of the two lexicons than it does about the efficacy of LCS-based selectional restrictions. The LQT technique depends on both the English and Spanish LCS lexicons, so the effective size of the overall LCS lexicon is best thought of as the intersection of the two separate lexicons. Although our Spanish LCS lexicon has coverage similar to the DQT bilingual term list, the English LCS

[7] DQT-EWS was omitted form the precision recall graph in order to minimize clutter because the average precision of DQT-EW was statistically significantly better at the 0.05 level.

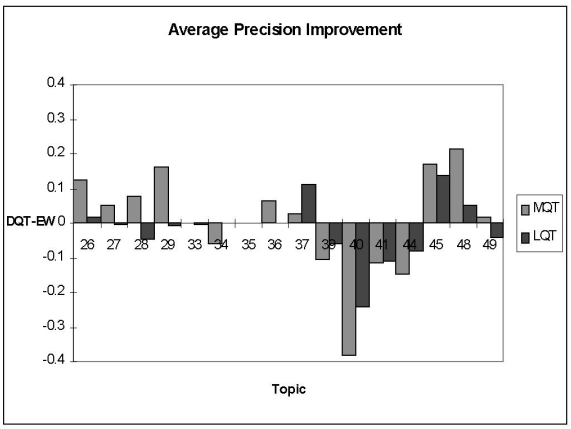

Figure 2: Improvement of MQT and LQT over DQT-EW on the El Norte Collection for 16 TREC-4 queries.

lexicon that we used for these experiments was considerably smaller.

5 Conclusions

We have presented a technique for evaluating the utility of lexical resources for cross-language information retrieval and applied that technique to compare several ways of using three different resources. We found that our newly developed technique based on Lexical Conceptual Structures is not as effective as a technique based on the use of an off-the shelf machine translation system. Further examination revealed that the queries in the test collection that we selected were not well matched with the expected strengths of our new technique. Although our investigations of linguistically sophisticated techniques for cross-language information retrieval are still at an early stage, we believe that these results illustrate a useful way to evaluate the utility of alternative linguistic resources for this application.

5.1 Acknowledgments

The authors are grateful to Maria Katsova and Wade Shen for their help with parsing, LCS composition, and generation of target-language terms; Paul Hackett for implementation of the information retrieval techniques; Scott Bennett and Harriet Leventhal for their assistance with the Logos translation system; the University of Massachusetts for the use of Inquery; and James Allan for help with Inquery configuration. The authors have been supported, in part, by Army Research Laboratory contract DAAL01-97-C-0042 and LETTER11097, NSF PFF IRI-9629108 and Logos Corporation, NSF CNRS INT-9314583, DARPA/ITO Contract N66001-97-C-8540, NSA Contract MDA904-96-C-1250, and Alfred P. Sloan Research Fellowship Award BR3336.

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] 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.

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

[5] 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.

[6] 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.

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

[8] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th ACM/SIGIR Conference*, pages 329–338, 1993.

[9] 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.

[10] Douglas W. Oard. *Multilingual Text Filtering Techniques for High-Volume Broad-Domain Sources*. PhD thesis, University of Maryland, 1996.

[11] Douglas W. Oard. Adaptive filtering of multilingual document streams. In *Fifth RIAO Conference on Computer Assisted Information Searching on the Internet*, June 1997.

[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.

[13] Douglas W. Oard and Bonnie J. Dorr. Evaluating cross-language text filtering effectiveness. In Gregory Grefenstette, editor, *Cross-Language Information retrieval*, chapter 12. Kluwer Academic, Boston, 1998.

[14] Douglas W. Oard, Bonnie J. Dorr, Paul G. Hackett, and Maria Katsova. A comparative study of knowledge-based approaches for cross-language information retrieval. Technical Report CS-TR-3897, University of Maryland, Institute for Advanced Computer Studies, April 1998.

[15] 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.

[16] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[17] 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.