Evaluating Resources for Query Translation in Cross-Language Information Retrieval

Bonnie J. Dorr and Douglas W. Oard

Department of Computer Science /UMIACS
and College of Library and Information Services
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 (Dorr, 1993) with conventional information retrieval techniques (Oard, 1996; Oard and Dorr, 1996). Given a query in one language, we transform the query into the corresponding terms in a second language. We evaluate this approach by comparing the resulting retrieval effectiveness with two translation-based techniques as well as two techniques for determining the lower and upper bound. The main innovation of this work is that it provides a principled framework for controlling translation ambiguity in cross-language information retrieval applications. Our focus is on the construction of disambiguated (target-language) queries by using verb-based entries in our lexicon to construct Lexical Conceptual Structures (LCS). We view our approach as a crucial step toward the evaluation of the utility of our LCS-based multilingual lexicon. 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. The database presently contains approximately 60,000 entries per language. Each entry consists of a linguistically-motivated representation called Lexical Conceptual Structure (LCS). Our goal is to evaluate the utility of this representation in the context of cross-language information retrieval (CLIR), using techniques from both interlingual machine translation (MT) (Dorr, 1993) and conventional IR (Oard, 1996; Oard and Dorr, 1996). As such, we view this endeavor as a first step toward establishing our LCS-based lexicon as a large-scale lexical resource that has applicability to multilingual problems other than machine translation.

We adopt an approach to CLIR that uses NLP techniques based on the LCS representation to transform a query in one language into the corresponding terms for document retrieval in a second language. Query translation has emerged as the most popular strategy for fully automatic broad coverage CLIR (Oard, 1997b). This technique 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. Research on CLIR has shown that translation ambiguity compounds the problem, producing a significant adverse effect on retrieval effectiveness (Oard, 1997c). 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., (Davis and Ogden, 1997; Hull and Grefenstette, 1996)).

We describe a new LCS-based query translation technique that takes advantage of this key insight. The main innovation of this technique is that it provides a framework for dealing with the translation ambiguity problem. Our focus for initial experimentation is on construction of disambiguated (target-language) queries from verb-based entries in our lexicon. We evaluate the effectiveness of our approach translating 17 queries from English into Spanish, using the Inquery text retrieval system to retrieve a ranked list of the documents best matching each translated query, comparing the top ranked documents to the set of documents that have been judged to be relevant to that query, and then computing standard IR effectiveness measures such as recall and precision.

In order to establish a sound basis for comparison, we implemented two additional approaches: MT-based query translation (translation of each query into the document language using an existing fully automatic MT system) and dictionary-based query translation (replacement of each query term with approrpiate translation from an online bilingual dictionary). Finally, we 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 take 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.

2 Experiment Design

The Text REtrieval Conference (TREC) has developed large-scale multilanguage collections designed specifically to support CLIR experiments. We have evaluated our LCS database as lexical resource for CLIR using the Spanish collection from TREC-4 (El Norte) for which English queries are available. This 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 elds, 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 rst 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 when processing the Spanish El Norte documents.

The next ve sections summarize the ve CLIR approaches that we have implemented.

3 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 rst stage of this approach involves a sentence analysis component that builds a syntactic structure produced by a parser called REAP (Right Edge Adjunction Parser) (Weinberg et al., 1995). For example, the parse tree produced for the sentence What are Mexico's attitudes toward press censorship has the following structure:

```
[CP Whati
 [S are
  [NP mexico
   [N attitudes
    [PP toward
     [NP press censorship]]]]
  [VP ei]]]
```

The next stage of query translation involves the construction of a language-independent, compositional representation called Lexical Conceptual Structure (LCS) (Dorr, 1993; Dorr and Olsen, 1997). 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 un lled argument positions (as indicated by the * marker). During the process of LCS composition, argument positions are lled. 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 (Evett, Hendler, and Spector, 1994). Parka-DB provides an ef cient 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 LCSQT is based on topics SP26-50 from the TREC-4 El Norte collection.[1] 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 mexico hacia]
censura pulse prensa]
```

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

```
Actitudes en Mexico sobre
la censura de la prensa
```

3.1 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 (Radwan and Fluhr, 1995). 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 Cran eld 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

[1] Of these queries, we were able to achieve full syntactic and semantic analysis for 17 case. The remaining 8 cases were not analyzable by the REAP parser. Modi cations 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, SP30, SP33, SP34, SP35, SP36, SP37, SP40, SP41, SP44, SP45, SP48, and SP49.

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 (Oard, 1997a). 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 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 translation is retained.

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., (Hull and Grefenstette, 1996; Ballesteros and Croft, 1997)). 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 (Dorr, 1997; Weinberg et al., 1995). 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 (Antworth, 1990) 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. 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 four simple dictionary-based query translation techniques that together explore the effects of winner-take-all, word-match and stem-match approaches.[3] We illustrate the effect of each technique with a Spanish translation of query SP45 above:

Single Word (SW) The first exact single whole-word match in the dictionary.[4]
[ ]

Single Word, Stemmed (SWS) The first exact single whole-word match if present, otherwise the first exact single stem match.[5]
[ ]

Every Word (EW) Every exact single whole-word match in the dictionary.
[ ]

Every Word, Stemmed (EWS) Every exact single stem match in the dictionary.
[ ]

In every case we replace each word in the query with the corresponding word in every matching bilingual pair 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 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. 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.

3.3 Upper Bound: Same Language Query (SLQ)

To approximate an upper bound for the performance of any CLIR system, we compared the retrieval effectiveness of our three experimental approaches with the retrieval effectiveness achieved by using queries that are given in the same language as the documents. For example, query SP45 would be presented as [Actitudes en Mexico sobre la censura de la prensa] when retrieving El Norte documents and as [Mexico's attitudes toward press censorship] when retrieving English documents.

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 SP45 shown above contains the proper name "Mexico" which also often appears in relevant Spanish 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.

4 Results

Table 1 summarizes the non-interpolated average precision results for the El Norte collection. The average precision for the LCSQT was better than that achieved by any DQT technique and comparable to that achieved by MTQT. From this we conclude that further investigation of the LCSQT approach is justified.[6]

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

[3] Bilingual dictionaries typically provide phrasal level entries as well. We were limited in the current experiment to single-word entries. Subsequent experiments will make use of an enhanced lexicon that contains phrasal level translations entries for both the source and target languages.

[4] 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.

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

[6] Interestingly, the average precision of LCSQT surpassed that of even SLQ on topic SP34. A closer examination reveals that

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

5 Conclusions

We have examined the question of whether LCS-based lexicons are a useful lexical resource for CLIR. Our evaluation of the utility of the LCS representation consisted of a comparison of the LCS-based query translation technique with several alternative CLIR techniques. 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 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.

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.

this was brought about by two distinctions between the LCSQT terms [senal fuerza nervio debilidad ejercito tierra mexican] and the SLQ terms [indicaciones fortalezas debilidades ejercito mexicano]. First, LCSQT used the more commonly used term fuerza in place of fortalezas. Second, LCSQT used two terms, tierra (land) and ejercito (armed forces) to produce the translation of army, instead of just the single, more general term ejercito.

References

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

Ballesteros, Lisa and W. Bruce Croft. 1997. 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.

Davis, Mark and William C. Ogden. 1997. 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.

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

Dorr, Bonnie J. 1997. 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.

Dorr, Bonnie J. and Mari Broman Olsen. 1997. 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.

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

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

Oard, Douglas W. 1996. Multilingual Text Filtering Techniques for High-Volume Broad-Domain Sources. Ph.D. thesis, University of Maryland.

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

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

Oard, Douglas W. 1997c. Alternative Approaches for Cross-Language Text Retrieval. In AAAI Technical Report SS-97-05, Stanford, CA.

Oard, Douglas W. and Bonnie J. Dorr. 1996. Evaluating Cross-Language Text Filtering Effectiveness. In Proceedings of the Cross-Linguistic Multilingual Information Retrieval Workshop, pages 8 14, Zurich, Switzerland.

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

Weinberg, Amy, Joseph Garman, Jeffery Martin, and Paola Merlo. 1995. 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. Lawrence Erlbaum Associates, Hillsdale, NJ, pages 23 44.