Implementing Cross-Language Text Retrieval Systems for Large-scale Text Collections and the World Wide Web

Mark W. Davis and William C. Ogden

From: AAAI Technical Report SS-97-05. Compilation copyright © 1997, AAAI (www.aaai.org). All rights reserved.

Computing Research Laboratory
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003
{madavis, ogden}@crl.nmsu.edu

Abstract

QUILT (Query User Interface with Light Translations) is a prototype implementation of a complete cross-language text retrieval system that takes English queries and produces English gloss translations of Spanish documents. The system indexes the Spanish documents in Spanish, but converts the English query into a Spanish equivalent set through a novel combination of lexical methods and parallel-corpus disambiguation. Similar methods are applied to the returned document to produce a simple translation that can be examined by non-Spanish speakers to gauge the relevance of the document to the original English query. The system integrates traditional, glossary-based machine translation technology with information retrieval approaches and demonstrates that relatively simple term substitution and disambiguation approaches can be viable for cross-language text retrieval. Components of QUILT have been used to build a CLTR interface to WWW-based search services.

Introduction

A Cross-language Text Retrieval (CLTR) system retrieves documents in a language that is different from the query language. A user of a CLTR system can enter a query in one language, but the returned documents will be in the language of the document collection. Some have speculated that one conceivable scenario is that a user might have some knowledge of the document language, but have difficulty formulating effective queries. This sort of user might very well be able to distinguish good documents from bad documents based on their limited knowledge. Such a user could then send the documents that they have judged relevant on to a translation service bureau or machine translation system. The more pressing scenario, however, is a user who does not have the time or the ability to judge documents in languages other than their own. A user with no knowledge of a language must pass all of the retrieved documents to other resources, which could constitute a significant load on an organization.

The Query User Interface with Light Translations (QUILT) system is designed to address the need for a CLTR system that takes queries in one language and returns documents in that same language, despite the fact that the underlying text retrieval system has indexed documents in a different language. The prototype system makes use of a bilingual lexicon to generate translation equivalents, then chooses among those equivalents based on a combination of corpus-based methods and lexical decision-making. The current prototype accepts queries in English for document collections in Spanish. The entire system is presented to the user as a GUI that allows entering queries, viewing, saving and printing queries and document lists, and viewing documents both in Spanish and as a simple English translation. Search terms are highlighted in the documents and the user can quickly examine alternative translations of terms in the English translation.

The potential value of a system like QUILT is that analysts and researchers with little or no knowledge of a language can nevertheless access textual databases in the language. They can also potentially make sense of the documents they discover without having to burden additional translation resources within their organization. The Spanish documents are still available for examination in QUILT, however, and so those with better bilingual skills can compare the original with the translation. The system thus has the potential for assisting in language learning in addition to retrieval.

This paper presents an overview of the design of QUILT. The translation methodologies employed in the system are simple, but appear to provide an adequate basis for further research and development of "full-route" CLTR systems. Parts of the system have been tested on a very large collection of Spanish documents in the context of the Text REtrieval Evaluation Conference (TREC). This collection consists of around 165,000 documents in over 300 Mb of Spanish text from AFP news wire. Standard queries with known relevance judgments were used to evaluate the performance of the query translation and document retrieval component. Experimental determination of the added value of

the GUI in allowing users with varying degrees of bilingual knowledge to distinguish good documents from bad ones is currently ongoing.

Query Translation

The first step in the QUILT approach to CLTR is to translate the English user query into Spanish. Other approaches in recent literature are not reliant on query translation (Dumais, 1996). For QUILT, the ability to perform query translation means that the translated queries can potentially be used to query text retrieval systems other than the current one implemented in this prototype.

The query translation process contains numerous steps:

1. Tag the English query with a statistical parts-of-speech tagger.
2. Reduce the English query terms to their stems and eliminate kill words, preserving the tagger markup from (1).
3. Look-up the stemmed English query terms in a bilingual lexicon, selecting Spanish equivalents that match the part-of-speech of the query term.
4. Retrieve a document set using the English query on the English side of a parallel, aligned corpus.
5. Speculatively retrieve document sets for each Spanish term on the Spanish side of the parallel, aligned corpus.
6. Choose the Spanish equivalent terms that produce retrieved documents that most resemble the English retrieval results.
7. Examine the remaining Spanish equivalents to determine if they occur in the target Spanish document database. If they do not occur, see if the English term, stemmed with a Spanish stemmer, occurs in the target Spanish database.
8. Construct a Spanish query based on the collected Spanish equivalents from (6) and (7).

The final query is thus constructed from terms in a bilingual lexicon that have been disambiguated on a parallel document corpus. Terms that are not in the lexicon are handled specially in step (7) because they may be proper nouns, Spanish terms or loan-words that occurred in the English query.

These steps are described in more detail in the following sections.

0.1 Tag English query

The first step in QUILT query translation is tagging the original query with parts-of-speech markup. The English query is passed to the English POS tagger from MITRE Corp. after the inclusion of SGML markup to delineate the start and ends of sentences. As an example, the following TREC-5 query (English description portion):

```
How has the threat of swine fever affected
international trade?
```

tags to:

```
<lex pos=WRB> How </lex> <lex pos=VBZ> has </
lex> <lex pos=DT> the </lex> <lex os=NN> threat
</lex> <lex pos=IN> of </lex> <lex pos=NNS>
swine </lex> <lex pos=NN> fever </lex> <lex
pos=VBD> affected </lex> <lex pos=JJ> interna-
tional </lex> <lex pos=NN> trade? </lex>
```

A further filtering step is performed by collapsing the spectrum of noun and verb tags generated by the MITRE POS tagger to JJ, NN,VB, CD, IN, DTD, PRP and FW, and then prefixing each query term with the POS tag. Other tags are discarded. For the query above, this results in:

```
VB_has NN_threat NN_swine NN_fever VB_affected
JJ_international NN_trade
```

Overall, the performance of the MITRE tagger appears to be very good. On the 25 TREC queries, there were 8 errors by the MITRE tagger over a total of 222 labelled terms in 25 queries, resulting in a 3.6% error rate on query tagging. Notable errors included: "in" was identified as a Foreign Word (FW), "steps" was incorrectly identified as a Verb (VB) twice, and "extinction" was incorrectly identified as a Verb (VB) once. This evaluation did not examine the CD, IN, DTD and PRP categories.

0.2 Stem the marked-up terms

The second step in the query translation process involves stemming the resulting terms using a standard English stemming algorithm based on the Porter stemmer, but modified to ignore the POS prefix codes. Terms in a 571 term kill list are also removed at this stage. For the query above, this results in:

```
NN_threat NN_swine NN_fever VB_affect JJ_intern
NN_trade
```

where "has" has been removed by the kill procedure and "affected" and "international" have both been reduced to root forms.

0.3 Find equivalents by bilingual lexicon lookup

The current QUILT implementation relies on the Collins Spanish-English and English-Spanish bilingual dictionaries. In a preprocessing stage, the Spanish-English dictionary was parsed from their original format by extracting sense and subsense equivalents along with their POS information. The headwords were then prefixed with the POS tag and stemmed by the correct stemming algorithm. The equiva-

lents were stemmed and duplicates eliminated, and the resulting lists were loaded into a btree-based database for use by the QUILT system. The relevant dictionary entries for the results from 0.2 are shown below in their parsed format:

```
NN_threat achor|amag|amenaz|bravat|conmin|dis-
fuerz|espant|nublar|peligr|ret|ronc

NN_swine
canall|cochin|galduf|jet|malaj|mam|mar-
ran|papal|perr|puerc|rajar|sinvergonz|ver-
gaj|villan

NN_fever calentur|chuch|fiebr|pasm

NN_intern intern

NN_trade comerc|contrat|negoc|ofic|sindi-
cat|tráfag|tráfic|trapich
```

The stemmed Spanish equivalents are separated by vertical bars. Note that the lexicon is highly "noisy" because the English-Spanish version of the lexicon is, in fact, an inversion of the Spanish-English lexicon because of parsing problems with the English-Spanish dictionary. This process of using an inverted dictionary is clearly not the ideal situation, but it does provide a quick path to a reverse lexicon when only one direction is available. The consequences of using this lexicon are described later.

The resulting English-Spanish lexicon contains 33, 360 entries with a mean of 2.11 equivalents per headword. The Spanish-English lexicon contains 40, 252 entries with a mean of 1.39 equivalents per headword.

0.4 Retrieve english documents on parallel text database

The corpus-based disambiguation component of QUILT uses a parallel, aligned database of texts to attempt to choose among the variants produced by 0.3. The corpus-based method is the product of numerous experimental efforts to incorporate aligned corpora into the translation task. Early efforts (Dunning and Davis, 1993) involved solving for a translation matrix of document-term correspondences using training methodologies. Although the results were encouraging, the process was computationally expensive. Davis and Dunning (1995, 1996) applied evolutionary programming methods to attempt to refine Spanish translation of English queries by iteratively comparing the retrieval profiles of English and Spanish queries over a parallel corpus. In Davis and Dunning (1995a), a transfer dictionary was used to create the Spanish queries, but no large-scale retrievals were performed, and the later work (Davis and Dunning, 1995b) used initial Spanish equivalents derived directly from a parallel corpus. Results from the latter were shown to be comparatively poorer than even the full transfer dictionary methods. In both cases, the evolutionary optimization methods were computationally expensive, requiring around 50,000 retrievals per query to achieve acceptable levels of optimization. The combined results of these experiments suggested that if a computationally-tractable alternative to the EP methods could be found that leveraged the value of bilingual lexicons, corpus-based methods might present an opportunity for improving query translation. In Davis (1996), evaluations of the current approach were presented, demonstrating that under the correct circumstances, and especially when combined with POS-disambiguated terms, corpus-based disambiguation can improve query translation.

For the method in QUILT, the first step involves performing a retrieval using the original English query to generate a vector of scored English document numbers. The current implementation of QUILT contains parallel texts from the 1991 United Nations collection of documents. The parallel documents were automatically aligned (Davis, Dunning and Ogden, 1995) resulting in 97,594 alignment pairs at the sentence or double-sentence level. The English documents contained 91,915 unique terms out of a total of 4,483,677. On the Spanish side, there were 122,827 unique terms in a total of 5,259,124. The alignment process has previously been estimated to be 83% correct, although a comprehensive evaluation of the UN alignments has not been performed. The 1991 UN document set was chosen because it was suspected that current issues might be better represented by the most current document set from the UN collection, which includes years 1988 through 1991.

The English set of aligned texts was indexed using the QUILT engine. The Spanish set was similarily indexed simultaneously, with alignment blocks sharing document numbers between the parallel sets. The resulting indexes occupy a total of 77 Mb of disk space, including inverse term token-term dictionaries for testing purposes. The indexing took approximately 20 minutes on a Sparc 5.

0.5 Speculatively retrieve Spanish documents

The next step is to retrieve documents using Spanish equivalents as query terms. For each English headword from 0.3, this involves retrieving a scored document vector for each equivalent term.

0.6 Choose the best equivalent for each English headword

In this step, the normalized dot product of each equivalent's document vector and the original English vector is calculated. For each equivalent set, the equivalent with the highest score is selected. In the case of the headword:

```
NN_fever
```

the dot product calculations indicate that "fiebr" is the best of the four available equivalents. A standard translation of the phrase "swine fever" is, in fact, "fiebre porcino," so that equivalent is a good choice among the four. For this query, the resulting disambiguated query is:

```
amenaz perr fiebr afect intern comerc
```

Compare this with the Spanish version produced by Human translation:

```
¿Qué efecto ha tenido en el comercio internacio-
nal la enfermedad "fiebre porcino?"
```

The comparison is clearer if one examines the stemmed and killed version of the query:

```
efect tenid comerc intern enfermedad fiebr por-
cin
```

Note that except for "afect" instead of "efecto," and the lack of translations "enfermidad" and "tenid", which are modifier terms in the query, all of the significant terms are present in the disambiguated query. The notable exception is the use of "perr" instead of "porcin."

0.7 Check for inclusion of terms that did not translate

Although in this query all relevant terms translated, it is conceivable that proper names, abbreviations or loan-words might not have entries in the bilingual lexicon. An added feature of QUILT is that these terms will be included in the query if they also occur in the target database. The English query terms are thus stemmed with the Spanish stemmer and checked against the target database term list. If they occur, they are included.

0.8 Submit query to Spanish retrieval engine

QUILT takes the resulting query, consisting of Spanish terms from 0.6 and 0.7, and submits them to the Spanish monolingual retrieval engine. The QUILT retrieval engine uses a simple tf-idf document and term weighting scheme similar to the SMART (Salton and Buckley, 1988) system for query-document scoring. As a prototype system, the flexibility of the vector-based tf-idf approach suggested that it was a reasonable approach. Further, a vector model is an inherently linear combination of term weightings, making the substitutions of term equivalents in a CLTR scenario straightforward, with special handling of phrasal components an added option that can be accommodated easily without significant modification of the system.

For QUILT, the tf-idf strategy has some substantial modifications over the Smart system from Cornell. Among these was the development of new Spanish stemmer based on the Porter stemmer model that contains 145 rules for stemming Spanish terminology. The complexity of irregular Spanish verbs was partially handled within this framework, although it was decided to do without specifying irregular verb paradigms precisely to maintain the speed of the stemming algorithm.

The system is capable of indexing at around 200 Mb per hour, Spanish or English, and creates indexes of around 0.5 the size of the original document collection. Posting vectors are incrementally written to B-tree databases to conserve memory and then merged at the end of the process without the necessity of sorting the individual posting sets. Additional options allow for the creation of a database of compressed document signatures which are useful for experimenting with automatic document feedback, although these features are not currently supported by QUILT.

Evaluating Retrieval Performance

The query translation component of QUILT has been evaluated in the context of TREC-4 and TREC-5, using the English description fields of the queries only to produce ranked document lists that could be compared to monolingual Spanish retrieval engines also participating in TREC. The results are encouraging and are reported in Davis (1996), but summarized below for completeness.

The NIST query-relevance judgments were combined with the English portions of the TREC-4 and TREC-5 queries (English description fields only) to assess the performance of the QUILT engine. The English description fields contain a human translation of the Spanish description fields used by the monolingual Spanish systems.

In the discussion that follows, the monolingual Spanish results will be referred to as MONO4 and MONO5. For comparison with the QUILT disambiguation approaches, runs were also done using all of the Spanish bilingual equivalents without any attempt at disambiguation. These runs are labeled ALL4 and ALL5, respectively. The QUILT corpus-based disambiguation method was also used without POS-based disambiguation to ascertain its relative contribution to the disambiguation process. These sets are referred to as CORP4 and CORP5. Similarily, the POS approach was applied independently of the corpus-based method to determine its performance in query translation. This set is referred to as POS5 (the same experiment was not done for TREC-4 queries). The complete QUILT approach is QUILT5, and was also only evaluated for the TREC-5 queries.

The pooled query-relevance judgements (qrels) from NIST were used to evaluate the system for several of these runs. It is possible that the stemming algorithm that was used for Spanish might conflate Spanish terms in a manner not represented in the other systems, so the pooled qrels are probably not a perfect measure of the system's performance. The effect of this probably does not significantly impact on the results presented here, however.

The performance of the various methods is shown in Table 1. The non-interpolated average precision values are

listed by category.

Table 1 Average Precision For All Methods

Method	PR (ND)	% of MONO
MONO5	0.2895	—
MONO4	0.1874	—
QUILT	0.2127	73.5
POS5	0.1949	67.3
ALL5	0.1422	49.1
CORP4	0.1250	66.7
CORP5	0.1153	39.8
ALL4	0.0783	41.8

Automatically translating a query into another language can clearly have a substantial performance penalty, but by performing some simple disambiguation of query term equivalents, the penalty can be reduced substantially.

MONO4, ALL4 and CORPUS4 results are included because they show a slightly different pattern than the corresponding results for TREC-5. In the TREC-4 results, there was a clear performance gain that resulted from corpus-based disambiguation of the translation equivalents. For the TREC-5 results, however, corpus disambiguation decreased performance when used alone but was advantageous when combined with POS-based disambiguation. Exactly why this occurred is not altogether clear. It certainly must be due to substantial disambiguation errors being made over incorrect POS equivalents present in the TREC-5 query equivalent sets. If we consider the monolingual performance of the basic QUILT retrieval engine as a baseline, the POS5 disambiguation component performed 67.3% as well as the monolingual system alone, while adding in the corpus-based component (QUILT) raised the percentage to 73.5%.

The monolingual QUILT retrieval engine is very basic, however, and should be considered as a vehicle for proofing the value of CLTR components, not as an exceptionally capable system in and of itself. It is comparable in design to early SMART systems. Substantial improvements in performance can be expected as the handling of phrases is incorporated into the system and, especially, as Rocchio-style automated feedback methods are added.

The Complete System

The final step in the QUILT process is producing gloss translations of the retrieved documents for the end-user. The current implementation uses the same basic design as the query-translation process to translate the documents to English, but substitutes a frequency-based scheme for the expensive corpus-based disambiguation process.

The steps of the document translation process are:

1. Annotate the document with SGML sentence delimeters.
2. Tag the document around any existing SGMLmarkup.
3. Separate the tagged terms and look them up in the Spanish-English lexicon.
4. Choose the equivalent that has the highest frequency of occurrence in a large English AP collection.

Each of these stages is essentially equivalent to the corresponding stage of the query-translation process, except for some minor variations. For one, the Spanish-English bilingual lexicon does not contain the English equivalents in stemmed form so that the translated document does not contain difficult to read term stems rather than their root forms. Secondly, statistics were collected for 450,000 terms from around 400 Mb of English AP news wires and the English equivalents in the lexicon were rearranged so that the first equivalent is the one with the highest frequency of occurrence in the AP document set. Since the user has access to all equivalents, this arrangement has a high likelihood of choosing the correct equivalent but allows the user to browse all equivalents as well.

The GUI for QUILT contains numerous features:

- A query window
- Menu options for displaying limited numbers of returned documents.
- Options for displaying the query translation terms from the bilingual lexicon.
- Options for saving, loading and printing queries and document lists.
- A document list window.
- Pop-up windows that show the Spanish document with translated Spanish query terms highlighted.
- Options for displaying the English gloss translation of the Spanish document.
- A pop-up window that shows the variant translations of each English term in the gloss translation.

The primary components are shown in Figure 1 on Page 6. In the figure, the main window has a query entry box in the top center. Ranked documents are shown in the lower listing. By clicking on one of the documents in the ranked list, the Spanish document window appears with translated query terms highlighted in red. Pressing the "View Translation" button at the window's base replaces the window contents with the gloss translation. Each translated term is lightly highlighted in brown, and English query terms are shown in red in the translations window. Further, by clicking on any translated term in the window, a small dialog box appears that shows

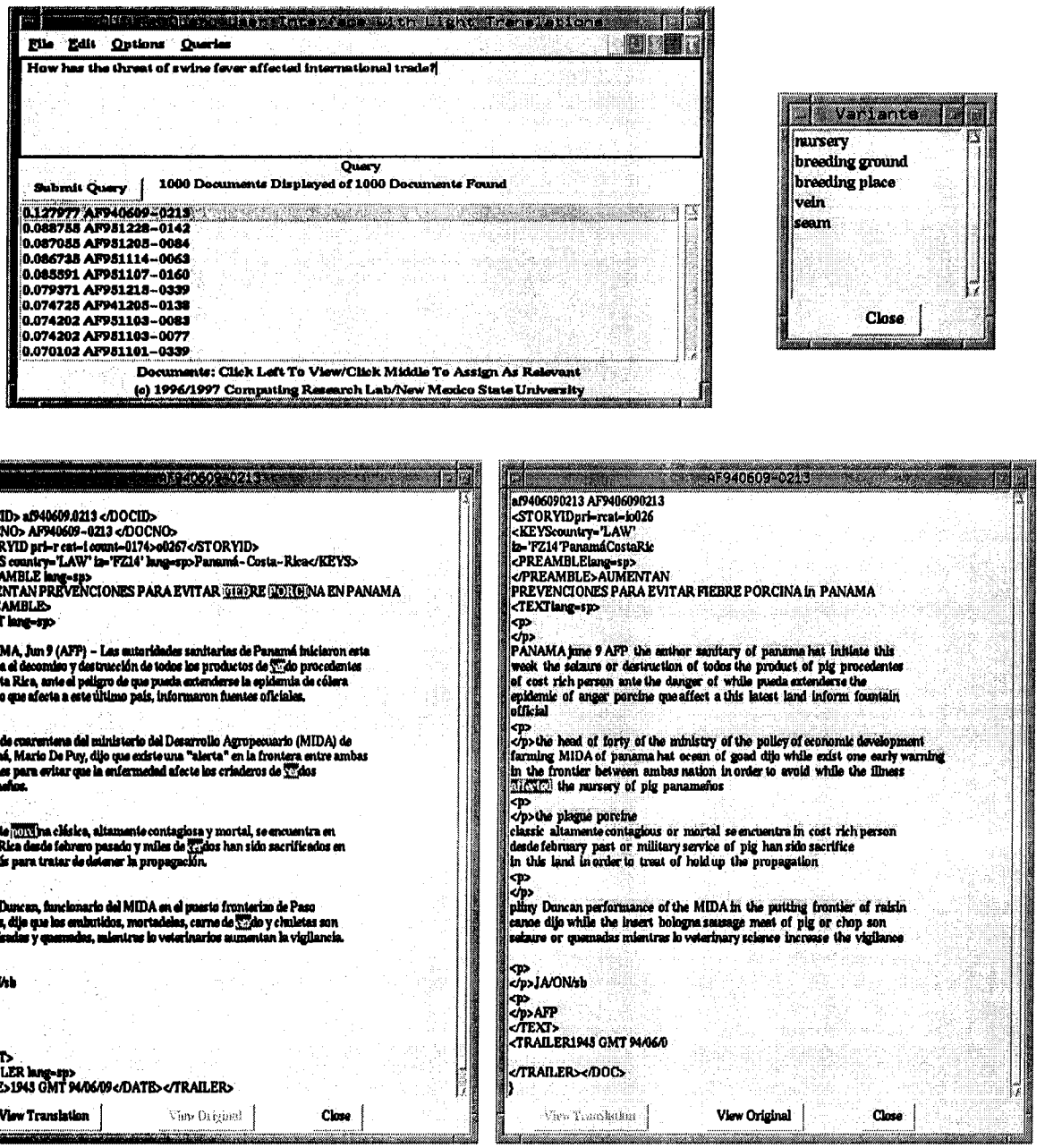

Figure 1. Components of the QUILT graphical user interface, showing the query/document list window (top left), the document views for the top-ranked Spanish document (bottom left) and the gloss translation of the Spanish document (bottom right). The variants of the term "nursery" are shown in the top right window. The variant list was shown by clicking on the term "nursery" in the translation window. Faintly visible are the highlighted query terms and their translations in both windows, allowing the user to quickly locate the query terms regardless of language.

all of the variant equivalents for that term.

For current evaluation purposes, the query window also has the TREC-5 queries built in, and the retrieved document lists are persistent between sessions for each user, being saved to local, hidden files named by user and query number.

Retrieval and translation times are currently somewhat slow, with a mean query time of 56 seconds on the 25 TREC-5 queries. Translation times are highly dependent on the length of the document, with short documents like the one shown in Figure 1 on Page 6 taking 16 seconds to translate, while a longer document from the same set and three times as long requires 31 seconds to translate. Since the POS tagger is called as an external process, the number of segments between native paragraph tags plays a significant role in the amount of time spent on the translation. Incorporating the POS tagger as a linked library to the main QUILT system should substantially decrease the start-up costs associated with tagging operations.

Evaluating QUILT

A complete CLTR system presents special evaluation problems that are not relevant to monolingual retrieval scenarios. We are currently designing experiments to attempt to test the QUILT model.

The first problem is that the standard relevance assessment model used in the ad hoc and routing forums of TREC is not appropriate to the complete CLTR system. One of the goals of QUILT is to assist users in judging the relevance of retrieved documents in non-native languages. The evaluation criteria therefore must be the number of relevant non-native documents correctly identified. This requires assessing the impact of a user's knowledge of the document language with or without the aid of the QUILT system.

Our approach is to pre-test users to ascertain their knowledge of Spanish using a multiple-choice test that has vocabulary specifically chosen over the TREC-5 queries. The users are then randomly assigned to attempt to evaluate the relevance of the Spanish documents to the original query either with or without the gloss translations. The hypothesis is that users with very little Spanish knowledge will be able to better make those judgments using the gloss translations. Better judgments will be determined by comparing the closeness of the relevance judgements by the evaluation users to the TREC relevance judgments provided for those queries.

An alternative methodology that we are currently exploring is the use of user choice analysis to ascertain precisely how users might use various user interfaces tools—including gloss translations—to arrive at an effective strategy for finding and making use of foreign language documents. In user choice analysis, more than one alternative user interface style is made available to a test user and they are free to choose a style after they have become familiar with all alternatives. Tasks types are varied and experimental observations are made of the choices users make to solve the tasks. It may be that the syntactic and functional errors that are in the English gloss translations make the added value of the translations primarily in the area of vocabulary discovery, but that users will rely on the Spanish document for processing the logical flow of the document discussion. By providing the choice for a user to use both together and observing how they actually use the two document representations, it may be possible to determine that additional tools could be effective in aiding the processes of document understanding and making relevance judgments.

Mundial: CLTR on the WWW

Parts of QUILT have been applied to the WWW, providing an interface to existing search engines that are freely accessible. The basic interface is shown in Figure 2 on Page 8. The floating frame at the top of the browser window allows a user to enter search terms, select a search service and choose to translate or not translate a query. In the prototype shown here, a user may also get a gloss translation of the returned document from the search engine. Due to security limitations inherent in Netscape 3.0 and Microsoft Internet Explorer 3.0, only the returned listing from the search engine can be translated. Arbitrary web pages found by "clicking-through" cannot be translated.

A version of Mundial is publicly accessible at:

http://crl.nmsu.edu/users/madavis/mundial.html.

Conclusions

QUILT is a significant step towards fulfilling the need for retrieval in one language over collections in other languages, and allowing the user to assess and understand the retrieved documents in the query language. The close integration of machine translation techniques with retrieval technology in QUILT appears to have increase the success rate of each technology alone in query translation. The same techniques appear to be effective in providing gloss translations of the retrieved documents for the user although significant evaluation efforts remain to be done.

Our ongoing work involves marshaling in-house development of Unicode processing tools and display technology to expand the QUILT model to other languages, and investigating new approaches to rapidly acquiring bilingual lexical information from large corpora. The advances seen in QUILT provide a positive preliminary answer to the basic question of whether a complete CLTR system is possible.

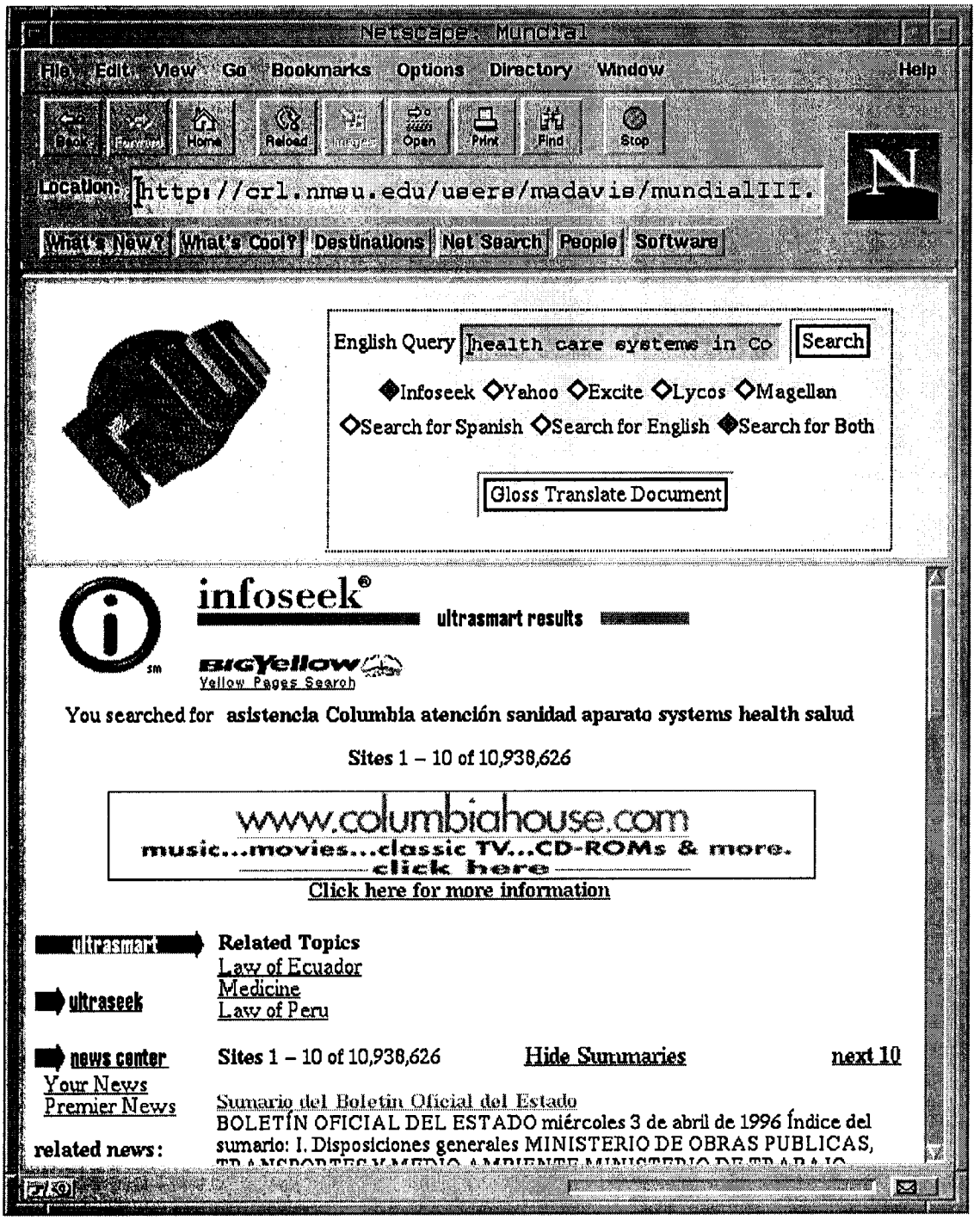

Figure 2. Mundial interface showing an English qeury that has been submitted to the Infoseek search service, retrieving documents in both English and Spanish.

Acknowledgements

We are grateful to MITRE, Corp. for the parts-of-speech taggers for English and Spanish. This work was funded under Tipster Text Phase III by the US Department of Defense.

References

Davis, M. W. (1996) New Experiments in Cross-Language Text Retrieval at New Mexico State University's Computing Research Laboratory. In *Proceedings of the Fifth Text Retrieval Evaluation Conference*, Gaithersburg, MD, National Institute of Standards and Technology..

Davis, M. W. and T. Dunning (1996) Query Translation Using Evolutionary Programming for Multi-Lingual Information Retrieval II. In *Proceedings of the Fifth Annual Conference on Evolutionary Programming*, San Diego, Evolutionary Programming Society.

Davis, M. W. and T. Dunning (1995a) Query Translation Using Evolutionary Programming for Multi-Lingual Information Retrieval. In *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, San Diego, Evolutionary Programming Society.

Davis, M. W. and T. Dunning (1995b) A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval.. In *Proceedings of the Fourth Text Retrieval Evaluation Conference*, , Gaithersburg, MD, National Institute of Standards and Technology..

Davis, M. W., T. Dunning, and W. Ogden (1995) Text Alignment in the Real World: Improving Alignments of Noisy Translations Using Common Lexical Features, String Matching Strategies and N-Gram Comparisons. In *Proceedings of the Conference of the European Chapter of the Association of Computational Linguistics*. University College Dublin. March.

Dunning, T. E., and M. Davis (1993) Multi-Lingual Information Retrieval. *Memoranda in Computer and Cognitive Science*, MCCS-93-252, Computing Research Laboratory, New Mexico State University.

Dumais, S.T., T. Landaurer and M. Littman (1995) Automatic Cross-Linguistic Information Retrieval Using Latent Semantic Indexing. In *Proceedings of the Workshop on Cross-Linguistic Information Retrieval*, SIGIR'96, Zurich.

Salton, G. and C. Buckley (1988) Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 5(24): 513-523.