ADVANCES IN MULTILINGUAL TEXT RETRIEVAL

Mark Davis
Computing Research Lab
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003
madavis@crl.nmsu.edu
(505) 646-1148

INTRODUCTION

Multilingual text retrieval extends the basic monolingual detection task to include retrieving relevant documents in languages other than the query language. The task therefore merges efforts in machine translation with efforts in text retrieval, but the machine translation component may be substantially simplified due to some basic assumptions about the design and implementation of high-performance text retrieval systems. A primary consideration is that most modern text retrieval systems regard queries and documents as unordered "bags" of words. The translation of an unordered set of terms is therefore approximately the translation of the terms themselves. Although a linearity assumption such as this breaks down when considering phrasal elements in most languages, it is reasonably accurate for many terms and becomes increasingly accurate at the sentence level and above.

A second consideration in multilingual text retrieval is where the translation is done. It is possible to translate every document at index time, for example, but the resource costs are substantially higher than translating the query at retrieval time. An added benefit of translating only the query is that queries can be prepared with no special weighting scheme applied to the terms. The queries are then available to any natural language text retrieval system.

The range of translation techniques that are available to a query translation system is greater than in standard machine translation systems. Previously translated document corpora can be made available for exploiting domain-specific terminology by direct comparison of the retrieval results for the query and target document languages. No special heuristics are needed for using this "example-based" translation approach; the query can be optimized by adding or deleting terms until the target language retrieval results are approximately the same as the source language retrieval results. Lexical-transfer techniques can also be used in the same context, providing wide coverage of term senses.

CRL evaluated five methods for query translation in Tipster II. The results were then evaluated in TREC by hand-translating the TREC Spanish monolingual queries into English and applying the automatic query translation methods to produce new Spanish queries. Ongoing work is focusing on improving the performance of query translation techniques while expanding the techniques to work with new languages and search engines, including WWW search services.

MLTR IN TREC

Starting with TREC- 3, Spanish corpora and query sets have been available for evaluating text retrieval engines. The queries and corpus are monolingual, however, so testing a multilingual system is only possible if the query set or the corpus is translated into a different language. We chose to translate the queries since they were very short. With translated queries, a query translation system that produces Spanish queries from hand-translated English versions of original Spanish queries can then be compared against the original queries. The differences between the two results are then a reasonable measure of the effectiveness of the translation process in preserving the characteristics of the original query that contribute to retrieval. Several of the Spanish TREC queries and their hand-translated versions are shown in Table 1, below.

The query translation methods that we applied to produce new Spanish queries were of two major types: methods that used a prepared lexicon and methods that used a parallel training corpus. While a lexicon tends to produce translations that are shallow but comprehensive, covering all possible senses of a term but limited in the range of synonyms that are produced for each term, corpus methods tend to produce translations that are deep but narrow, with enormous repetition of domain-related senses of terminology. This justified an examination of the comparative merits of both approaches.

As is often the case, our parallel corpus was not precisely of the same domain as the TREC document collection for the ultimate evaluation. The corpus itself was extremely large, however, which we hoped would offset the difficulties of using a distinctly different type of text. The corpus was 1.6 Gb of Spanish and English translations from the United Nations, containing proceedings of meetings, policy documents and notes on UN activities in member countries. The documents were automatically aligned [1] at the sentence level using a procedure that is conservatively estimated to have an 83% accuracy over grossly noisy document pairs (which the UN documents were not). This produced a parallel corpus of around 680,000 aligned sentence pairs.

Lexical Transfer

The first method was to perform term-by-term translation with the Collins English-Spanish bilingual dictionary. Individual terms in the English query were reduced to their morphological roots and lookup was performed. The resulting set of Spanish terms became the Spanish query. Some repetition of terms is apparent in the resulting queries because all senses of each term were used with no attempt to disambiguate the contextual usage of the English terms. For example, Query 28 is transformed from

```
Indicators of economic and business
relations between Mexico and Asian
countries, such as Japan, China and
Korea.
```

to

```
indicador indicador ayuda expansión
previsiones crecimiento comercio com-
ercio narración relación parentesco
México Ciudad gripe patria campo
región amor semejante parecido tanto
el laca China Mar té porcelana vitrina
coalín Corea Corea Corea mexicana mex-
icano México
```

Note that "China" has been replaced with both "China" and "porcelana" as a result of this simple lexical substitution scheme, and that "relations" has included the familial sense "parentesco". Lexicon-generated Spanish Queries

The lexical-transfer approach produced Spanish queries rapidly, requiring only a simple database lookup procedure. This process is shown in Figure 1 (a).

High-Frequency Terms from Parallel Text

In text, the terms that occur with the highest frequency are rarely of statistical significance, and are more often than not merely redundant. Yet the terms that occur with moderate frequency are sometimes significant. In order to evaluate other corpus-based methods, we wanted to establish a baseline for queries formed from these moderate frequency term sets. Using a vector-based text retrieval system with no term spreading or other modifications, the English queries were translated by performing a lookup on the English side of the parallel corpus, collecting the Spanish sentences that were parallels to the top 100 retrieved documents, filtering the remaining terms to eliminate the top 500 most frequent Spanish terms, and collecting the next 100 most frequent Spanish terms to create the new query. This process is shown in Figure 1 (b):

Several of the resulting queries are given in Table 2. Some formatting codes from the UN documents have been eliminated in some of the queries, reducing the count to below 100 terms in those queries. For brevity, only the first two queries are shown in Table 2.

Statistically Significant Terms

Whereas the high-frequency terms extracted in the previous method provide a baseline for examining improved methods, high-frequency terms are themselves not necessarily the best terms for discriminating the significant features involved in text retrieval. A better approach is to extract the terms which are statistically significant in the retrieved segments of parallel text in comparison to the corpus as a whole. Various methods are possible for testing statistical significance, but the method we applied is based on a log-likelihood ratio test that assumes a χ^2 distribution is an accurate model of the term distributions in text [2].

The method begins by extracting all of the terms from the sentences that are parallels to the top 100 retrieved English sentences. The counts of the pooled terms are then compared with the counts for the entire UN training corpus to evaluate their statistical significance. The top 100 most-significant terms are then extracted and become the new Spanish query. Figure 1 (c) diagrams the process. The resulting queries are in Table 3, below.s

Evolutionary Optimization of Queries

If we could make a set of derived Spanish queries retrieve documents in a manner that is similar to the English queries over a training corpus, then the Spanish query could conceivably produce similar results on a novel corpus. One way to change Spanish queries is to add and remove terms. The number of possible unique deletions that can be performed on a 70 word query is quite large, however, making the direct examination of all possible modified queries effectively impossible.

We applied an evolutionary programming (EP) [3] approach to modify a population of 50 queries. In an EP approach, an initial population of queries is needed along with a mutation strategy to modify queries. Optimization then proceeds by evaluating the comparative fitnesses of the queries, mutating a selected sub-population of the queries to produce "offspring" solutions and re-evaluating the queries iteratively until a suitable number of generations have passed. Our EP approach considered the comparative evaluation of document score vectors as an objective measure of the relative fitness of a query to the collection. This process is diagrammed in Figure 1 (d).

The initial queries for this test were the queries from the high-frequency lookup strategy discussed above. Previously, we have used a lexicon to generate initial queries [4]. The mutation strategy applied between one and ten modification operations to each of the 50 queries per generation and collected only the best 10% of the queries to propagate into the next generation. Optimization proceeded for 50 generations, resulting in a wide range of changes to each query.

The types of queries produced by this system typically showed the repetition of key terminology combined with the elimination of irrelevant terms. The fitness judgment for a query was based on comparative retrieval results using a training corpus of only 80,000 aligned sentences. Table 4, below, shows two of the resulting queries from the EP method.

Singular Value Decomposition and the Translation Matrix

The final query translation method was a radical departure from the others, but is derived from earlier work by [5] and [6]. This method is at heart a numerical approach to derive a translation matrix from parallel texts.

In this effort, we applied a *QR*-decomposition technique to reduce the complexity of calculating the singular value decomposition, resulting in query translation that took only a matter of seconds on a SPARC 10. Several of the generated queries are given in Table 6. Figure 1 (e) diagrams the process.

OVERVIEW OF RESULTS

The resulting queries were given to University of Massachusetts, Amherst, who ran them against the Spanish TREC document collection using Spanish Inquery. The original Spanish TREC queries were also evaluated to establish a reference baseline. The results were as follows:

1. On average, the dictionary-based queries produced performance which was about 50% worse than the reference queries.
2. The EP-derived queries produced performance which was 60-70% worse than the reference queries, except at higher recall levels (.6-1.0), at which they performed better than the Method 1 queries.
3. The other methods performed even more poorly.
4. On at least two queries, performance of the lexical methods was as good or better than the reference queries.
5. On two queries, performance of the EP approach was as good as the reference queries, although they tended to have better precision at higher recall.

These modest results demonstrate that lexical and corpus methods can be applied to query translation in a large-scale multilingual text retrieval scenario, although at a fair penalty in performance. Each of these methods was purposely limited to as simple a scheme as possible, however, so there is plenty of room for improvement and further experimentation.The average precision-recall curve for all 25 queries is shown in Figure 2.

RECENT AND ONGOING WORK

Current work is focusing on improving the performance of MLTR methods, applying the methods to new languages and making use of new retrieval engines.

An example of the latter is shown in Figure 3. *Mundial* is a query interface to Infoseek and Yahoo that takes queries in English, translates them to Spanish and submits the resulting queries to the Infoseek and Yahoo search engines directly. Figure 4 shows the completed search for Spanish documents on Infoseek. The *Mundial* demo uses a bilingual dictionary combined with several heuristics to limit the terminological expansion of the input query. Limiting query size is important because most search engines, like Infoseek, restrict the size of a query to around 80 characters. Overgeneration in the translation process is handled by using the longest terms (in character count) in Mundial. Although in some cases this may be in error, the hope is that automatic stemming of query terms at the search engine will reduce long terms to stems common to many of the keywords that might have been substituted if the entire definition was transferred. The second motivation was that long terms tend to be more precise than short terms, and content words should be as precise as possible.

Mundial may be accessed at:

`http://crl.nmsu.edu/ANG/ML/ml.html.`

REFERENCES

[1] Davis, M. W., T. E. Dunning, and W. C. 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 1995.

[2] Dunning, T. E. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, 19, 1: 61-74.

[3] Fogel, D. B. (1992), "A Brief History of Simulated Evolution," In *Proc. of the First Annual Conference on Evolutionary Programming,* ed. D.B. Fogel and J.W. Atmar, 1-16. San Diego: Evolutionary Programming Society.

[4] Davis, M. W. and T.E. Dunning (1995) "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, 1995.

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

[6] Landauer, T. K. and M. L. Littman (1990). "Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing," In *Proceedings of the 6th Conference of UW Centre for the New Oxford English Dictionary and Text Research*, 31-38. Waterloo.

Q#	Hand-translated English	Corpus High-Frequency Spanish
26	Indicators of economic and business relations between Mexico and European contries.	Checoslovaquia En Ghana Polonia nacional programa Australia Bajos Egipto España Filipinas La Países Portugal Igualdad Italia Paz recursos Austria Finlandia Acción Pide Venezuela Naciones gubernamentales Unidas como período una Comisión Desarrollo regionales sesiones Mujer Mundial información nacionales informe México resolución no proyecto un actividades países Estados organizaciones desarrollo sus su E/CN mujer Secretario General por República al con se Conferencia sobre para del las que los el en la de
27	Indicators of economic and business relations between Mexico and African contries.	Checoslovaquia Democrática Egipto Filipinas Francia Indonesia Irlanda Los Países Secretario Uruguay aplicación más proyectos servicios Alemania Colombia La fuentes trabajo Asamblea Iraq Naciones Nigeria Pakistán Unidos documento han DE Unidas energía nuclear sus Brasil principios siguientes utilización Argentina Chile En Venezuela como desarrollo espacio ultraterrestre El General una período sesiones al países su Estados sobre un para República por con se México que las del los en el la de

Table 1 Several Spanish TREC queries and their English translations

Q#	Hand-translated English	Statistically-significant Spanish Queries
26	Indicators of economic and business relations between Mexico and European contries.	período un una Anguila CARICOM Dos ECCB En Este Oeste Europeo Guyana Jefes Magreb Occidente Parlamento Principal T al ciencias con consentimiento consulares convenciones correo cuantitativos de del diplomáticos el empresarial en experiencias externas guías la las los para por que residente se sobre su sustituir tecnológica temporal tienden tomaron tono totalidad trabajan tradicionales transacci transacción transacciones transición transparencia tratará tratase trigésimo trimestre tropiezan trueque ultimado un un Seminario una unificado university urbanas utilizarse véanse vacantes validez vecindad vecinos venían vencimientos vende versión vigentes vinculadas vinculado vinculados voluntarios y Sudáfrica y financiación y rechazó
27	Indicators of economic and business relations between Mexico and African contries.	árboles Anguila CARICOM ECCB En Este Oeste Guyana Jefes Principal al ascenso autóctonos ciencias con consentimiento consulares convenciones correo cuantitativos de del diplomáticos el empresarial en experiencias externas guías la las litorales los mar nato occidental para por que se semillas sobre su títulos tecnológica temporal terremoto tienden tierras titular tomaron tono totalidad trabajan tradicional tradicionales transacción transacciones transición transparencia tratará tratase trimestre tropicales tropiezan trueque un un Seminario una unas unificado urbanas utilizan véanse víctima vecindad vecinos venían vencimientos vende verán versión vigentes vinculadas vinculado vinculados voluntarios vulnerables y Sudáfrica y financiación y rechazó

Table 2 Examples of Statistically-significant Spanish Queries

Q#	Hand-translated English	Evolutionary Optimized Spanish Queries
26	Indicators of economic and business relations between Mexico and European contries.	Checoslovaquia En nacional Egipto Filipinas Portugal Finlandia gubernamentales Unidas una sesiones Mundial México resolución no un países organizaciones sus su República al sobre que en la Egipto nacional Filipinas Conferencia países México Checoslovaquia México México Egipto México México una Finlandia mujer México Egipto las se Finlandia Egipto como Comisión información E/CN sobre un Unidas General Unidas desarrollo países Finlandia Filipinas México actividades un nacional no Conferencia Filipinas Checoslovaquia Portugal nacionales Conferencia México República Egipto México al nacional proyecto México Secretario mujer que proyecto Filipinas que México Filipinas Finlandia la México En Checoslovaquia mexicana mexicano México
27	Indicators of economic and business relations between Mexico and African contries.	Egipto Los servicios Colombia Asamblea Naciones Unidos documento sus Argentina En General una al países Estados sobre un República con México del en una Colombia México servicios una México que Estados Egipto México en México siguientes Argentina trabajo Egipto México Asamblea documento Egipto Argentina República con de Secretario trabajo México principios la aplicación Colombia Argentina DE Egipto Colombia han las aplicación General Colombia Argentina servicios Colombia un documento han México los una en las México México con mexicana mexicano México

Table 3 Evolutionary-Optimized Spanish Queries

Q#	Hand-translated English	Corpus High-Frequency Spanish
26	Indicators of economic and business relations between Mexico and European contries.	Exteriores Relaciones Guillermo Bedregal Culto Ioan Bolivia Ministro documento párrafos México con parte reproducido oficiosas ex Simone decisión ° período Voicu Rumania externas Ayuda titulado si Gutiérrez asimismo decían mexicana mexicano México
27	Indicators of economic and business relations between Mexico and African contries.	costeras Los constituir INTERES MUNDIAL principales probablemente cambios bien curso profundamente posibles DE pobladas PROBLEMAS sí comprender particular contiguas Ministro próximo Las verán Culto donde pronosticado camino climáticos Zelandia causados mexicana mexicano México
30	Are there sports programs and exchanges between Mexico and the United States?	Exteriores Relaciones Guillermo Bedregal Culto Finlandia Bolivia Ministro relacionados programas Rumania sí serie conjunto distingue denominan Unión Soviéticas determinarse motivos México Voicu asociación convenios integrado Nam Gutiérrez del SIDA entre mexicana mexicano México

Table 4 SVD generated queries

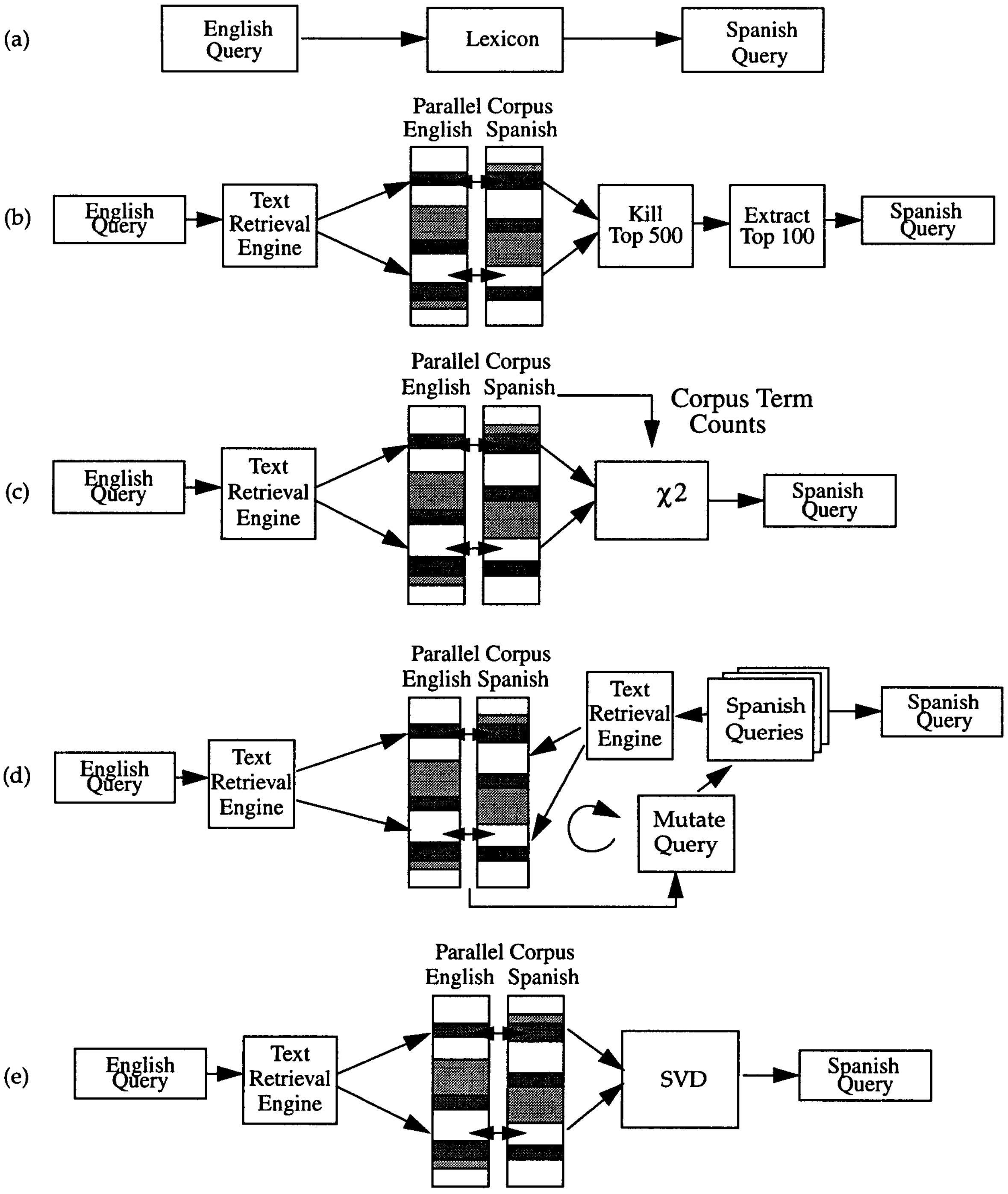


Figure 1 Diagrams of approaches to query translation: (a) lexical substitution, (b) corpus high-frequency, (c) corpus statistically-significant, (d) evolutionary optimization and (e) SVD approaches.

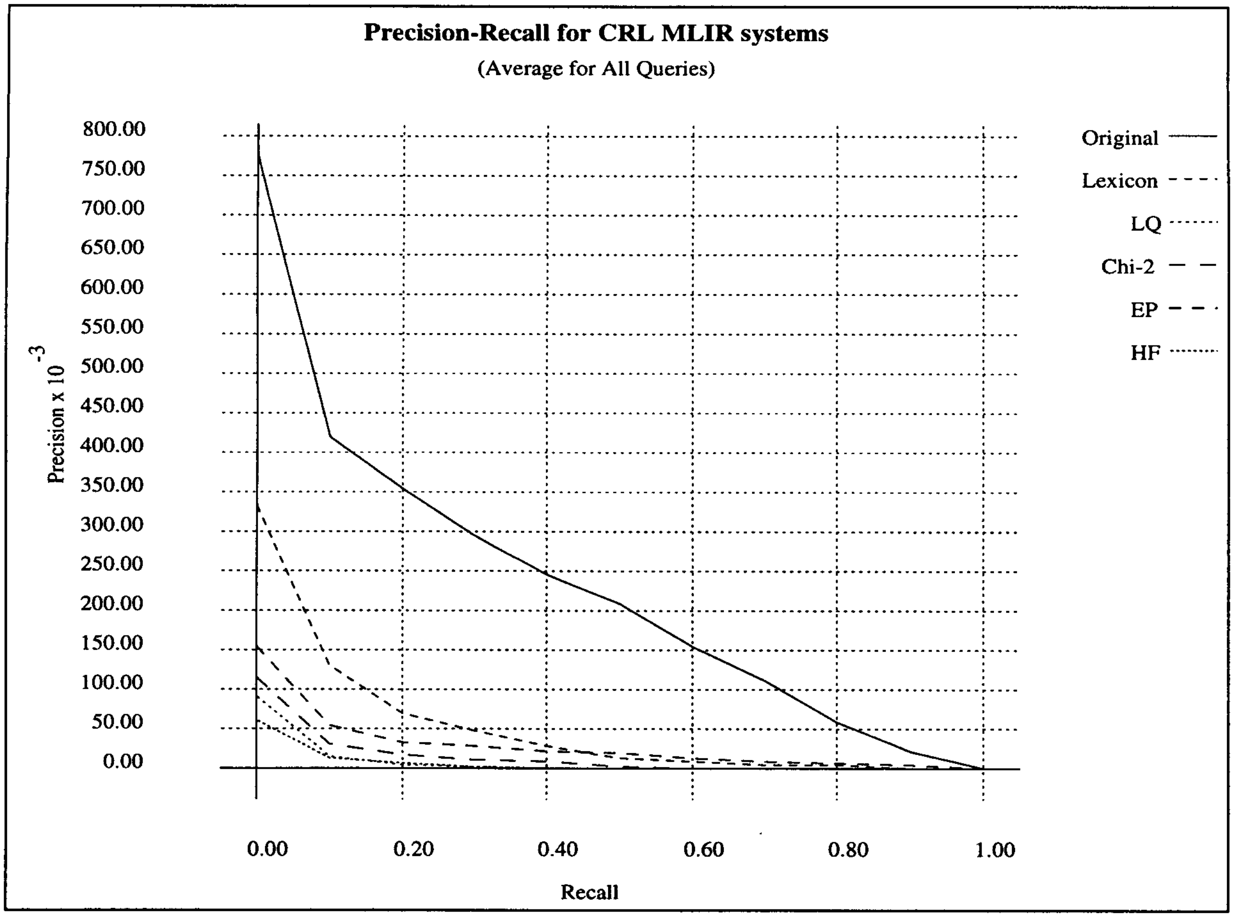
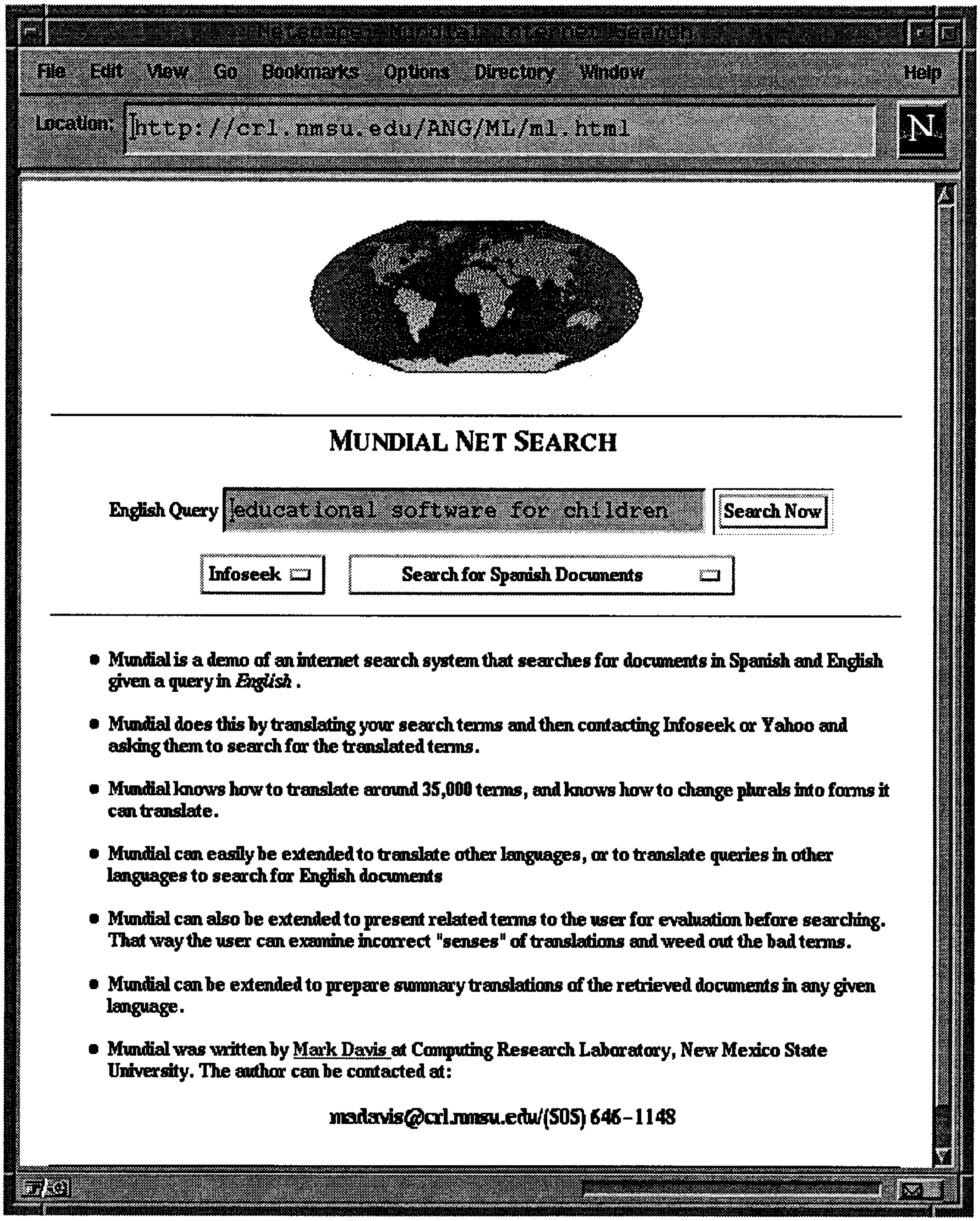


Figure 2 Average precision-recall curves for MLTR methods over 25 Spanish queries.



Location: http://crl.nmsu.edu/ANG/ML/ml.html

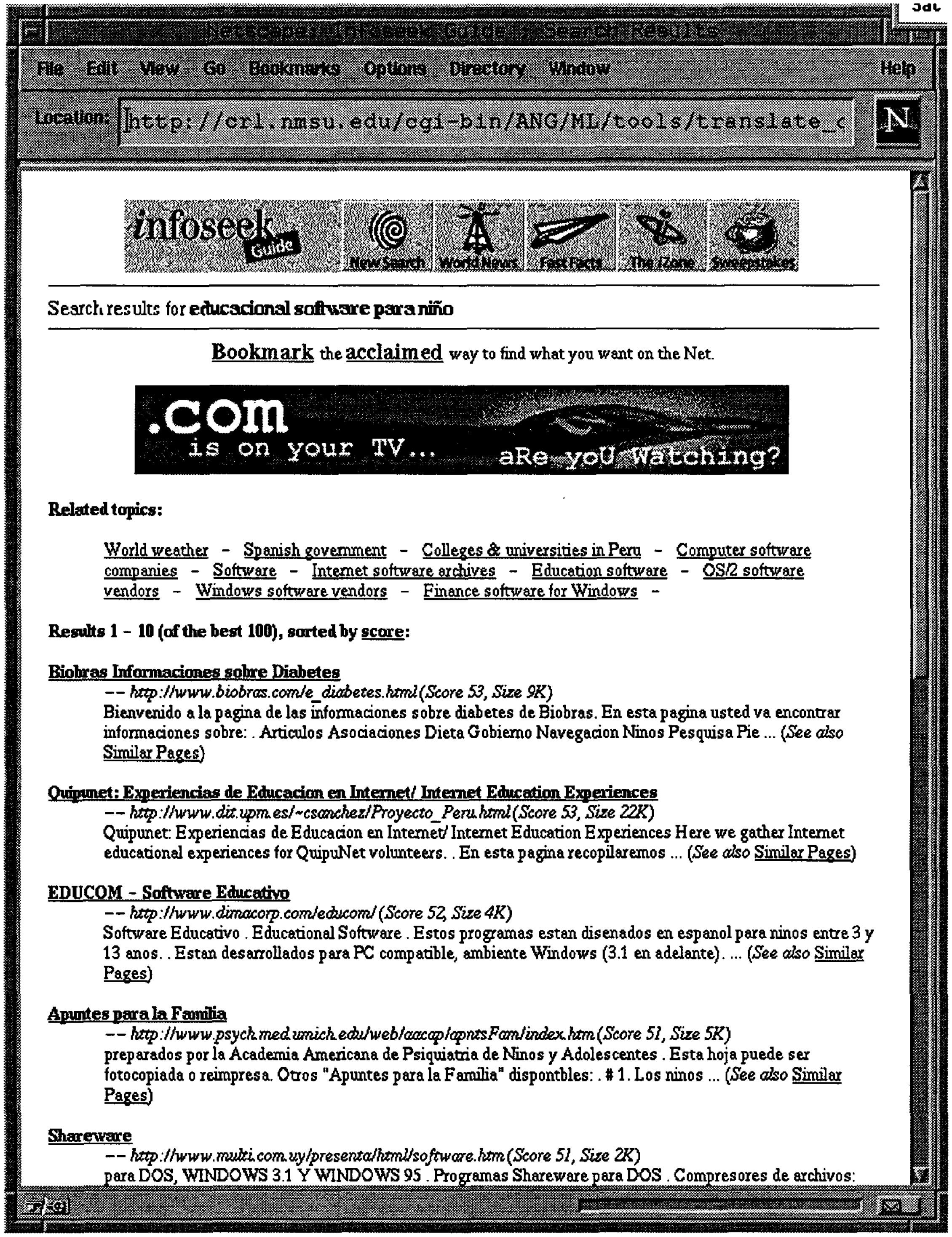
MUNDIAL NET SEARCH

English Query: educational software for children

- Mundial is a demo of an internet search system that searches for documents in Spanish and English given a query in *English*.
- Mundial does this by translating your search terms and then contacting Infoseek or Yahoo and asking them to search for the translated terms.
- Mundial knows how to translate around 35,000 terms, and knows how to change plurals into forms it can translate.
- Mundial can easily be extended to translate other languages, or to translate queries in other languages to search for English documents
- Mundial can also be extended to present related terms to the user for evaluation before searching. That way the user can examine incorrect "senses" of translations and weed out the bad terms.
- Mundial can be extended to prepare summary translations of the retrieved documents in any given language.
- Mundial was written by Mark Davis at Computing Research Laboratory, New Mexico State University. The author can be contacted at:

madavis@crl.nmsu.edu/(505) 646-1148

Figure 3 *Mundial*, a World Wide Web MLTR interface. English queries can be submitted to Infoseek or Yahoo after translating into Spanish or a mixture of English and Spanish for searches over both English and Spanish documents.


Netscape: Infoseek Guide: Search Results
File
Edit
View
Go
Bookmarks
Options
Directory
Window
Help
Location:
http://crl.nmsu.edu/cgi-bin/ANG/ML/tools/translate_c
N
infoseek Guide
New Search
World News
Fast Facts
The iZone
Sweepstakes
.com is on your TV... aRe yoU Watching?

Search results for **educacional software para niño**

Bookmark the acclaimed way to find what you want on the Net.

Related topics:

World weather - Spanish government - Colleges & universities in Peru - Computer software companies - Software - Internet software archives - Education software - OS/2 software vendors - Windows software vendors - Finance software for Windows -

Results 1 - 10 (of the best 100), sorted by score:

Biobras Informaciones sobre Diabetes
-- *http://www.biobras.com/e_diabetes.html (Score 53, Size 9K)*
Bienvenido a la pagina de las informaciones sobre diabetes de Biobras. En esta pagina usted va encontrar informaciones sobre: . Articulos Asociaciones Dieta Gobierno Navegacion Ninos Pesquisa Pie ... (*See also* Similar Pages)

Quipunet: Experiencias de Educacion en Internet/ Internet Education Experiences
-- *http://www.dit.upm.es/~csanchez/Proyecto_Peru.html (Score 53, Size 22K)*
Quipunet: Experiencias de Educacion en Internet/ Internet Education Experiences Here we gather Internet educational experiences for QuipuNet volunteers. . En esta pagina recopilaremos ... (*See also* Similar Pages)

EDUCOM - Software Educativo
-- *http://www.dimacorp.com/educom/ (Score 52, Size 4K)*
Software Educativo . Educational Software . Estos programas estan disenados en espanol para ninos entre 3 y 13 anos. . Estan desarrollados para PC compatible, ambiente Windows (3.1 en adelante). ... (*See also* Similar Pages)

Apuntes para la Familia
-- *http://www.psych.med.umich.edu/web/aacap/apntsFam/index.htm (Score 51, Size 5K)*
preparados por la Academia Americana de Psiquiatria de Ninos y Adolescentes . Esta hoja puede ser fotocopiada o reimpresa. Otros "Apuntes para la Familia" dispontbles: . # 1. Los ninos ... (*See also* Similar Pages)

Shareware
-- *http://www.multi.com.uy/presenta/html/software.htm (Score 51, Size 2K)*
para DOS, WINDOWS 3.1 Y WINDOWS 95 . Programas Shareware para DOS . Compresores de archivos:

Figure 4 Results of a *Mundial* search on Infoseek.