A TREC EVALUATION OF QUERY TRANSLATION METHODS FOR MULTI-LINGUAL TEXT RETRIEVAL

Mark Davis and Ted Dunning
{madavis, ted}@crl.nmsu.edu
Computing Research Lab
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003

ABSTRACT

In a Multi-lingual Text Retrieval (MLTR) system, queries in one language are used to retrieve documents in several languages. Although all of the collection documents could be translated to a single language, a more efficient approach is to simply translate the queries into each of the document languages. We have investigated five methods for query translation that rely on lexical-transfer and corpus-based methods for creating multi-lingual queries. The resulting queries produced by these systems were then used in a competitive information-retrieval environment and the results evaluated by the TREC evaluation group.

INTRODUCTION

The goal of text retrieval is to retrieve documents that are closely related to a user's needs. In general, the problem of translating the user's needs into queries for text retrieval systems is central to the field of text retrieval (Salton, 1970). The query may contains many more words than appear in the original user-input. In addition, users are generally rather poor at assigning weights to the terms in the query. Choosing good terms and weighting them is hard enough in one language, but when the problem is extended to involve documents in multiple languages it becomes considerably more difficult. In one possible scenario, the user generates input for the system in only a single language, but expects to retrieve documents in multiple languages (multi-lingual text retrieval or MLTR). In the context of conventional vector-based retrieval systems, this could be accomplished by automatically translating all of the documents into a single language when indices of the documents are created, or by translating the user-input into a multi-lingual query.

This paper describes the results of work on several approaches to MLTR that build multi-lingual queries using a lexicon and a corpus of previously translated documents. The comparative performance of the different methods is evaluated using original Spanish queries run with a competitive TREC text-retrieval engine as a baseline. The query translation process then operates on hand-translated versions of the original queries and the resulting translations are run with the same system.

A BRIEF HISTORY OF MULTI-LINGUAL TEXT RETRIEVAL

Salton (1970) first demonstrated that text retrieval could be used in a multi-lingual setting. His system used a thesaurus for generating query translations by taking the terms in the thesaurus for each query term and forming a new translated query. The thesaurus was created by hand for the retrieval corpus and the entries were therefore inherently disambiguated with respect to the corpus domain prior to query generation. Nevertheless, Salton's results demonstrate that IR systems can perform well in a multi-lingual setting using simple translation resources. Unfortunately, domain-specific, up-to-date glossaries are generally difficult to obtain. Those that are produced are typically constructed by and for translators, who write them in the process of translation, suggesting that an approach which makes use of the translations directly in combination with other resources is needed.

Experiments with latent semantic indexing (LSI) (Landauer and Littman, 1990) showed that paragraphs which were translations of each other could be retrieved but no actual retrieval system was constructed, nor was it clear how the system would perform in practice. This use of parallel corpora eliminates many of the problems of using bilingual dictionaries, but introduces new problems. In particular, in the context of a traditional vector based retrieval system, it has not been clear how to perform multi-lingual retrieval based on the information contained in parallel translated corpora. The success of experiments with LSI does not directly provide a method to make a more traditional vector based system work. Furthermore, LSI makes the use of inverted indexes problematic, which may hinder the practicality of this system.

Dunning and Davis (1993a,b) developed a system for multi-lingual text retrieval based on a novel method for solving very large systems of linear equations. In this system, query translation was viewed as a linear transformation of a query feature vector. For long strings, the translation of the concatenation of the strings is approximately the translation of the strings independently. This is true because the translation of two strings is nearly the concatenation of their translations. While this linearity breaks down dramatically at the word level, at the sentence level and above, it works fairly well. Despite the simplification afforded by linearity in the transformation, the actual translation matrix was derived through a computationally-taxing error minimization strategy that used a parallel aligned corpus of 50,000 words as exemplars to iteratively update the transformation matrix. At that time, machine resources were very limited and the algorithm had poor convergence properties.

Davis and Dunning (1995) applied an evolutionary approach to parallel collections by optimizing for translated query performance over a collection of parallel texts. The initial queries were created from term-by-term lookup in a bilingual machine readable dictionary. Although promising, the work left many unanswered questions about the usefulness of general lexicons for highly specific text domains and the value of corpus-based term disambiguation in an IR framework. Moreover, the results appeared to be affected by the lack of high-quality parallel text for training and evaluation.

NEW APPROACHES TO MLTR

Starting with TREC- 3, a Spanish corpus and query sets have been available for evaluating text retrieval engines. The queries and corpus are monolingual, however, so testing a multi-lingual 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. This assumes, of course, that the English hand translation preserves the retrieval characteristics and remains an unknown in our evaluation efforts. The Spanish TREC queries and their hand-translated versions are shown in Table 1, below.

Table 1: Spanish TREC queries and English Translations

Q#	Original Spanish Queries	Hand-translated English
26	Indicaciones de las relaciones económicas y comerciales de México con los paises europeos.	Indicators of economic and business relations between Mexico and European contries.
27	Indicaciones de las relaciones económicas y comerciales de México con los paises africanos.	Indicators of economic and business relations between Mexico and African contries.
28	Indicaciones de las relaciones económicas y comerciales de México con los paises asiáticos, por ejemplo Japon, China y Corea.	Indicators of economic and business relations between Mexico and Asian countries, such as Japan, China and Korea.
29	Indicaciones de las relaciones económicas y comerciales entre México y Canadá.	Indicators of economic and business relations between Mexico and Canada.
30	¿Hay programas y intercambios deportivos entre México y los Estados Unidos?	Are there sports programs and exchanges between Mexico and the United States?
31	¿Cuáles son las medidas tomadas por el gobierno mexicano para resolver la disputa con los rebeldes zapatistas en el estado de Chiapas?;	What measures has the Mexican government taken to resolve the quarrel with the rebel Zapatitas in the state of Chiapas?
32	¿Cuál es la importancia de las Naciones Unidas (NU) para México?	What importance does the United Nations (UN) have for Mexico?
33	¿Cómo van las relaciónes entre México y la Organización de los Estados Americanos (OEA)?	How are relations between Mexico and the Organization of American States (OAS)?
34	Indicaciones de los potenciales y debilidades del ejército mexicano.	Indicators of the Mexican Army's strengths and weaknesses.
35	Indicaciones de los potenciales y debilidades de las fuerzas aéreas militares de México.	Indicators of the Mexican Air Force's strengths and weaknesses.
36	Indicaciones de los potenciales y debilidades de la marina de guerra (fuerzas navales, armada) de México.	Indicators of the Mexican Navy's (naval and marine forces) strengths and weaknesses.
37	Evidencia de la herencia y cultura azteca en México.	Evidence of Aztec heritage and culture in Mexico.
38	¿Hay programas para la renovación urbana en México?	Are there urban renewal programs in Mexico?
39	¿Cuáles son las medidas modernas para mejorar la agricultura en México?	What modern measures for agricultural impovement are there in Mexico?
40	Información sobre el ballet folklórico en México.	Information about Mexico's traditional dance (ballet folklorico).
41	Medidas para controlar o evitar inundaciones en México.	Flood prevention and control measures in Mexico.
42	¿Tendrá éxito el NAFTA (TLC) en México?	Will NAFTA (TLC) be successful in Mexico?
43	Hay programes para reprimir o limitar epidemias en México?	Are there epidemic control programs in Mexico?
44	Información sobre la industria de computadoras mexicana.	Information about Mexico's computer industry.
45	Actitudes en México sobre la censura de la prensa.	Attitudes in Mexico regarding censorship of the press?
46	Informes sobre visitas oficiales y privadas a México por jefes de estado y de gobierno.	Reports regarding official and private visits to Mexico by chiefs of state and heads of government?
47	Hay programs en México para investigar la causa de cáncer?	Does Mexico have research programs for the cause of cancer?
48	Hay programas internacionales para el intercambio de estudiantes en México?	Are there international student exchange programs in Mexico?
49	El turismo como fuente de divisas para México.	Tourism as a source of Mexico's income.
50	La fabricación en México de joyas de plata y oro.	Silver and gold jewelery manufacturing in Mexico.

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 (Davis, Dunning and Ogden, 1995) 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 coun-
tries, such as Japan, China and Korea.
```

to

```
indicador indicador ayuda expansión previsiones crecimiento comercio comer-
cio 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 mexicano 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”. The complete set of queries is listed in Table 1.

Table 2: Lexicon-generated Spanish Queries

Q#	Hand-translated English	Lexicon-Generated Spanish
26	Indicators of economic and business relations between Mexico and European contries.	indicador indicador ayuda expansión previsiones crecimiento comercio comercio narración relación parentesco Méjico Ciudad Comunidad Comisión ribunal
27	Indicators of economic and business relations between Mexico and African contries.	indicador indicador ayuda expansión previsiones crecimiento comercio comercio narración relación parentesco Méjico Ciudad
28	Indicators of economic and business relations between Mexico and Asian countries, such as Japan, China and Korea.	indicador indicador ayuda expansión previsiones crecimiento comercio comercio narración relación parentesco Méjico Ciudad gripe patria campo región amor semejante parecido tanto el laca China Mar té porcelana vitrina coalín Corea Corea Corea

Q#	Hand-translated English	Lexicon-Generated Spanish
29	Indicators of economic and business relations between Mexico and Canada.	indicador indicador ayuda expansión previsiones crecimiento comercio comercio narración relación parentesco Méjico Ciudad
30	Are there sports programs and exchanges between Mexico and the United States?	allá deporte caza deporte juego cambio canje intercambio intercambio Méjico Ciudad estado estado
31	What measures has the Mexican government taken to resolve the quarrel with the rebel Zapatitas in the state of Chiapas?	medida medida excesivamente mejicano gobierno administración Estado régimen resolución resolución reyerta cabecilla estado estado
32	What importance does the United Nations (UN) have for Mexico?	importancia nación des no nada poco aciones poseer Méjico Ciudad
33	How are relations between Mexico and the Organization of American States (OAS)?	narración relación parentesco Méjico Ciudad especialista hule inglés fútbol estado estado
34	Indicators of the Mexican Army's strengths and weaknesses.	indicador indicador mejicano resistencia fuerzas intensidad fuerza flojedad tenuidad flaco desventaja
35	Indicators of the Mexican Air Force's strengths and weaknesses.	indicador indicador mejicano resistencia fuerzas intensidad fuerza flojedad tenuidad flaco desventaja
36	Indicators of the Mexican Navy's (naval and marine forces) strengths and weaknesses.	indicador indicador mejicano escuela arquitectura marítimo ingeniero ingeniería seguro fuerza fuerza resistencia fuerzas intensidad fuerza flojedad tenuidad flaco desventaja
37	Evidence of Aztec heritage and culture in Mexico.	evidencia testimonio hechos herencia patrimonio cultura choque choque Méjico Ciudad
38	Are there urban renewal programs in Mexico?	allá zona éxodo guerrillero renovación renovación reanudación extensión prorrogación Méjico Ciudad
39	What modern measures for agricultural impovement are there in Mexico?	medida medida excesivamente agropecuario escuela-granja perito feria allá Méjico Ciudad
40	Information about Mexico's traditional dance (ballet folklorico).	información noticias aproximadamente baile danza ballet
41	Flood prevention and control measures in Mexico.	inundación avenida pleamar torrente prevención medida medida excesivamente Méjico Ciudad
42	Will NAFTA (TLC) be successful in Mexico?	Méjico Ciudad
43	Are there epidemic control programs in Mexico?	allá control control dominio Méjico Ciudad
44	Information about Mexico's computer industry.	información noticias aproximadamente ordenador informático industria
45	Attitudes in Mexico regarding censorship of the press?	actitud ademán postura disposición Méjico Ciudad censura presión apretón presa
46	Reports regarding official and private visits to Mexico by chiefs of state and heads of government?	relato parte informe noticia autorizado huelga particular detective visita Méjico Ciudad jefe jerarca jefe estado estado cabeza cabellera gobierno administración Estado régimen
47	Does Mexico have research programs for the cause of cancer?	Méjico Ciudad poseer investigación investigación investigaciones causa causa cáncer canceroso investigación
48	Are there international student exchange programs in Mexico?	allá Cámara Corte línea derecho estudiante investigador asociación alumnado cambio canje intercambio intercambio Méjico Ciudad
49	Tourism as a source of Mexico's income?	turismo el fuente procedencia fuente foco ingresos rédito
50	Silver and gold jewelery manufacturing in Mexico?	abedul abeto hoja oro barra galón oro capacidad costos industrias Méjico Ciudad

The lexical-transfer approach produced Spanish queries rapidly, requiring only a simple database lookup procedure.

High-frequency Terms from a Parallel Corpus

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, below:

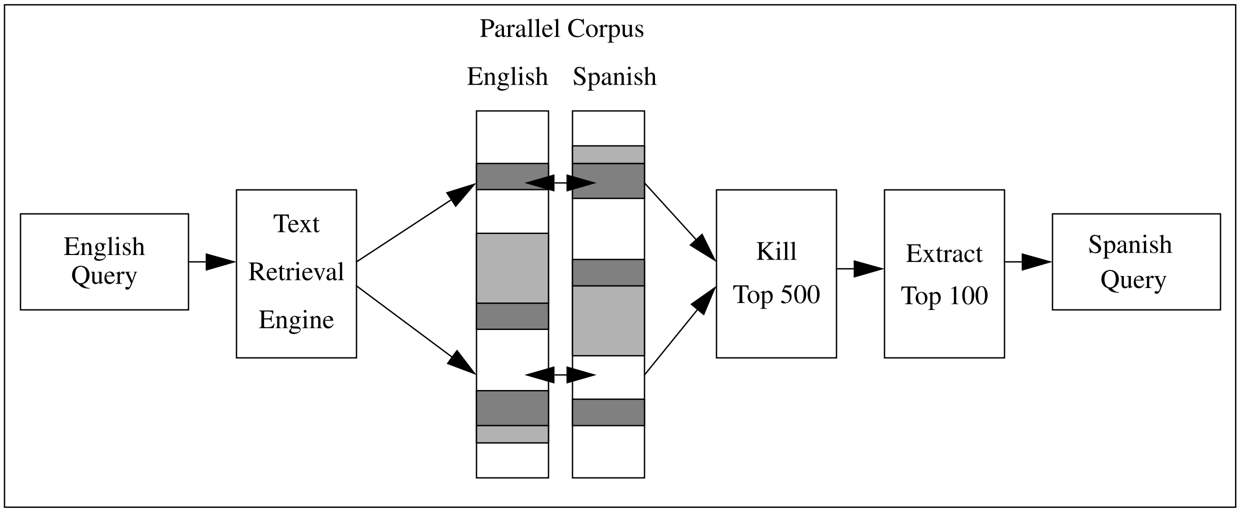

Figure 1. Building Spanish queries by extracting high-frequency terms from a parallel

Several of the resulting queries are given in Table 3. 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 four queries are shown in Table 3.

Table 3: High-frequency Queries from a Parallel Corpus

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
28	Indicators of economic and business relations between Mexico and Asian countries, such as Japan, China and Korea.	Italia Repúblicas Chile Ecuador General Gran India Pakistán Suecia Unión Venezuela Votos sus Bretaña Filipinas Norte Nueva Socialistas Soviéticas siguiente Alemania Bajos Colombia España Grecia Indonesia Reino su Argentina Francia Japón Unido especializados Brasil Países sistema Arabe organismos países Nicaragua Sudáfrica China organizaciones contra El México América sobre se Representante ante Irlanda con Permanente Unidos Seguridad para Estados Carta fecha dirigida Presidente Consejo Naciones Unidas al que República por en las los del el la de

Q#	Hand-translated English	Corpus High-Frequency Spanish
29	Indicators of economic and business relations between Mexico and Canada.	Nicaragua Nigeria Uganda Yugoslavia Zelandia con favor Dinamarca Indonesia Norte Ucrania Venezuela Yemen para Australia Botswana Finlandia Italia Soviéticas Suecia Austria Bangladesh Bulgaria Chile Hungría Irán Islámica Noruega Pakistán Repúblicas Socialistas Bajos Brasil Bretaña España Gran Países Reino Unión Unido Egipto Francia Perú Polonia América Argentina India Japón Kenya Canadá Checoslovaquia China Alemania Guinea Nueva Colombia Democrática Votos El Irlanda Arabe Estados México Unidos que las los el del en la República de

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 (Dunning, 1993).

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 2 diagrams the process. The resulting queries are in Table 4, below.

Table 4: Statistically-significant Spanish Queries

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

Q#	Hand-translated English	Statistically-significant Spanish Queries
28	Indicators of economic and business relations between Mexico and Asian countries, such as Japan, China and Korea.	Aprobación Bin Development En Estudio Kumar Nehru Omán Panjaponés Phase Productos Raczkowski Reimnitz Resources Sindicato Trabajadores Water al austríacos centavos chelines computadoras con de del diplomáticas el el en francos la las los mantienen mm navales nombró págs para poblaciones por que revestía sazón se secundario semana semanas siglos sindicatos sobre solo su sucesión sudoriental suizos suizos sujeción sumamente suspendió sustantivas sustituya títulos termina tradicionales transacciones tribales trigésimo un una unificación vacante vecindad venía vicepresidentes vinculadas vio visitaba visitas voz vuelos y Venezuela y el Reino y estudios y financiación

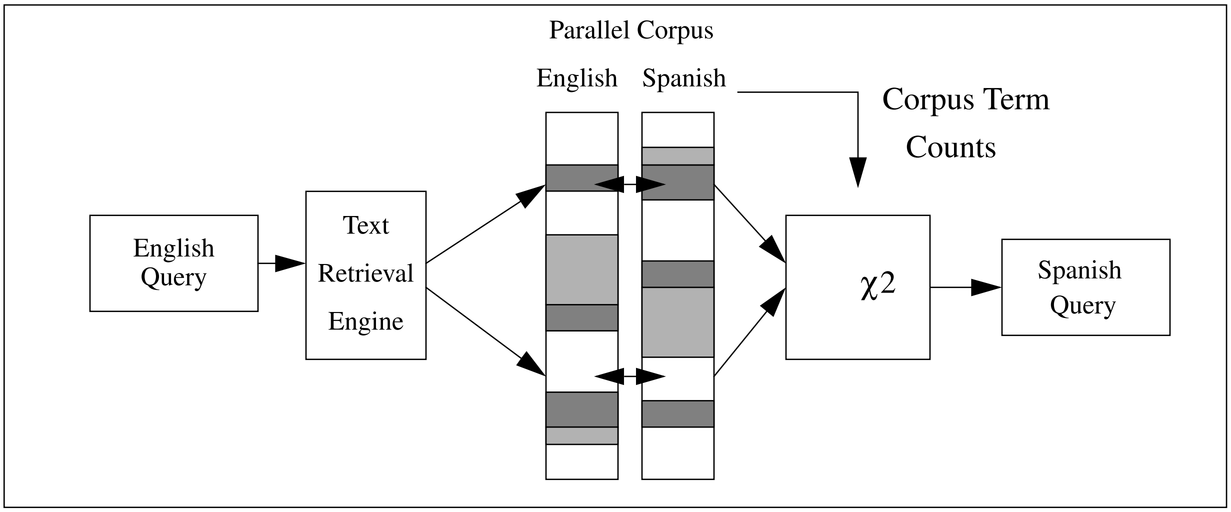

Figure 2. Extracting statistically-significant terms from parallel text look-up.

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) (Fogel, 1992) 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.

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 (Davis and Dunning, 1995). 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 fitness changes of the queries over 50 generations are shown in Figure 2.

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. The entire 680,000 aligned sentences were not used because we were uncertain that these results would be evaluated. We therefore intended an in-house evaluation over the remainder of the corpus should TREC evaluation be unavailable. Also, the retrieval engine for training was a traditional vector-based engine and not the same engine that was used to evaluate the TREC retrieval results (which was unavailable at the time of query preparation). Table 5, below, shows several of the resulting queries from the EP method.

Table 5: Evolutionary-Optimized 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
36	Indicators of the Mexican Navy's (naval and marine forces) strengths and weaknesses.	nivel respecto internacional internacionales Los lo marina sistema todos DE programas Asamblea General entre Consejo Mundial actividades países no climáticos Unidas como una por del para los en el la de climáticos marina marina R en sobre climáticos las marina Consejo los fracciones aplicación Consejo marina al período marina respecto marina entre zonas Mundial respecto marina zonas marina un sobre Gobierno climáticos todas sin sus mundial marina sus climáticos marina marina sus marina marina marina fracciones Estados del General marina por Los las En mexicana mexicano México
44	Information about Mexico's computer industry.	forma nuevas particular resultados contra económico fin lo medios Comité período sesiones Comisión internacionales Estados computadoras cuestiones mediante medio programa sistema computadora actividades El así servicios al sus países como su desarrollo una Naciones la La computadora organizaciones para actividades para computadora internacional así forma información computadora como período particular del computadoras el con Se computadoras General computadoras actividades programa computadora computadoras período del como computadora lo forma computadora computadora desarrollo ese servicios En computadora se sobre lo computadoras país que computadoras computadoras económico En computadora computadora computadoras mexicana mexicano México

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 Dunning and Davis (1993) and Landaurer and Littman (1990). This method is at heart a numerical approach to derive a translation matrix from parallel texts.

Let us suppose that there exists a translation operator that translates one query into another and, furthermore, that the queries can be represented as vectors of real-valued weights. This latter assumption is reasonable for virtually all text retrieval systems that treat queries as unordered "bags of words." In practice, a linearity assumption is reasonable in that translations of separate

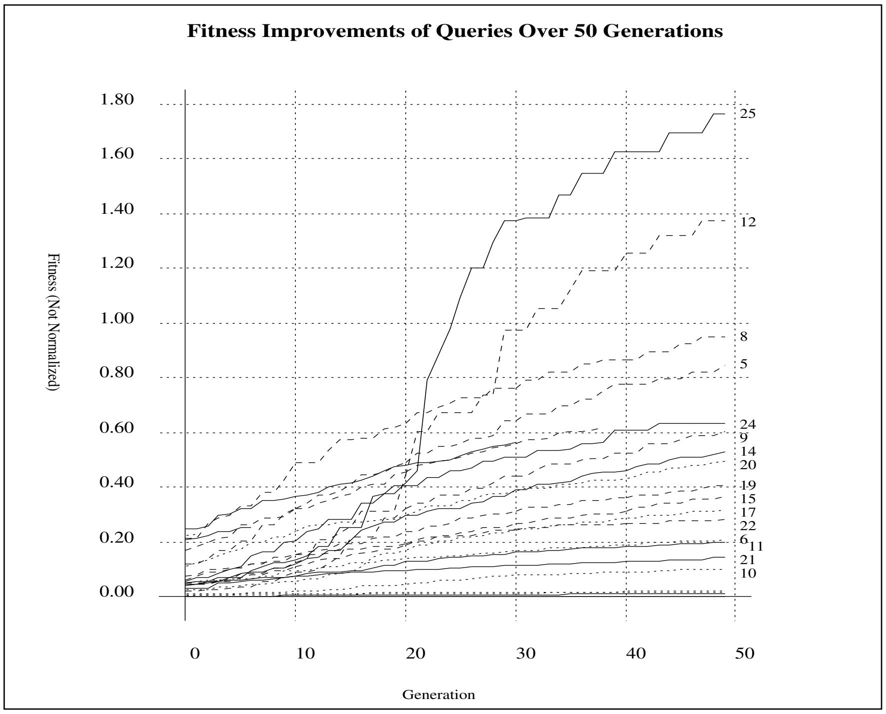

Figure 3. Fitness improvements of queries 1-25 over 50 generations during training. For TREC purposes, query 1 is equivalent to query 26, and so forth.

sections of texts are generally equivalent to translations of the text as a whole. This assumption breaks down below the level of the sentence, but is a useful approximation at the sentence level and above.

More specifically, given that we have a large number of short pieces of text which are translations of each other, we can use these short texts to form a set of linear equations which, when solved, yield a translation matrix T. Since T will be approximately a d by d matrix, it would seem that we would need millions of these translated bits of text. In fact, we can solve for an underdetermined least squares solution for T using singular value decomposition. This solution method avoids the numerical problem of not having enough samples, as well as handling the inevitable situation where the derived equations will not be quite consistent. The effect of using an underdetermined solution should be to preserve any ambiguity present in the query relative to the original parallel text.

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. The generated queries are given in Table 6. It should be noted that these numerical methods are still very preliminary and relatively untested. One early result that we discovered was that English terms occurring in the Spanish text and vice-versa are highly singular features for the translation matrixes. These early trials must therefore be treated with caution.

Table 6: SVD generated 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
28	Indicators of economic and business relations between Mexico and Asian countries, such as Japan, China and Korea.	informe aprobó octubre Junta sobre Voicu Mesa sesión Conferencias Finlandia período Guillermo básicos ˚ Consultivo UNCTAD/GATT Sucedáneos problems health Tungsteno relaciones México Ioan Arabes Exteriores grupos aportaciones credenciales Tal reuniones mexicana mexicano México
29	Indicators of economic and business relations between Mexico and Canada.	Exteriores Relaciones Guillermo Bedregal Culto México Bolivia Ministro Ley ideologías Rumania ejemplo cuerpo Otra resguarda provocarlos afectan étnico Voicu racistas exige aludida Tadanori Gutiérrez contribuirá cinematográficas 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
31	What measures has the Mexican government taken to resolve the quarrel with the rebel Zapatitas in the state of Chiapas?	mexicanas desequilibrios ecológicos Realiza arreglar costas presidente Pacífico evitar cuantía Mexicano recibido NASA información Comité siguiente El diga métodos futuro Nigeria junto Cirugía - Consejo propuestas archivado embargo comunidad actividades federal limitada mexicana mexicano México

RESULTS AND EVALUATION

The TREC evaluation procedure was an ideal approach to evaluating the effectiveness of the lexicon and corpus-based translation methodologies presented in this paper. Some caveats have to be considered with regard to the overall value of TREC results, however, including:

1. The different domains of text covered by the UN corpus and the target corpus for TREC-4 evaluation.
2. The English queries were hand-translated versions of the Spanish TREC queries that may not fully articulate the retrieval profile of the original Spanish query over the TREC corpus (a good test of this would be to translate back the English queries by hand with another translator and see whether the retrieval results differ.)
3. The unavailability of the *same* text retrieval engine for both training phases and for generating the final retrieval results.

Of all of these considerations, (3) is perhaps the most important. The Inquery Spanish retrieval engine was not fully functional until a few days prior to the submission date and therefore was unavailable for optimization in the case of EP evaluation iterations or for finding parallel text segments for query term extraction in all of the other cases. A straightforward idf-based retrieval engine was therefore substituted. What effect this has had on the ultimate results remains the subject for future research.

Despite these caveats, the full set queries were successfully run for each of the methods and results were evaluated at NIST. All of the methods resulted in substantial reductions in the precision-recall averages for the queries, although notable exceptions existed in which individual queries outperformed, or performed approximately as well as, the original Spanish queries. The former is a matter for some elaboration. It is always conceivable that a query modification process may result in better performance with respect to the original query but it is also very unusual for a process that set out to perform a radical transformation with the intent of "recreating" that query to actually outperform the original. This last result is a matter for further analysis.

Precision-Recall Averages

The average precision-recall curve for all 25 queries is shown in Figure 3, below.

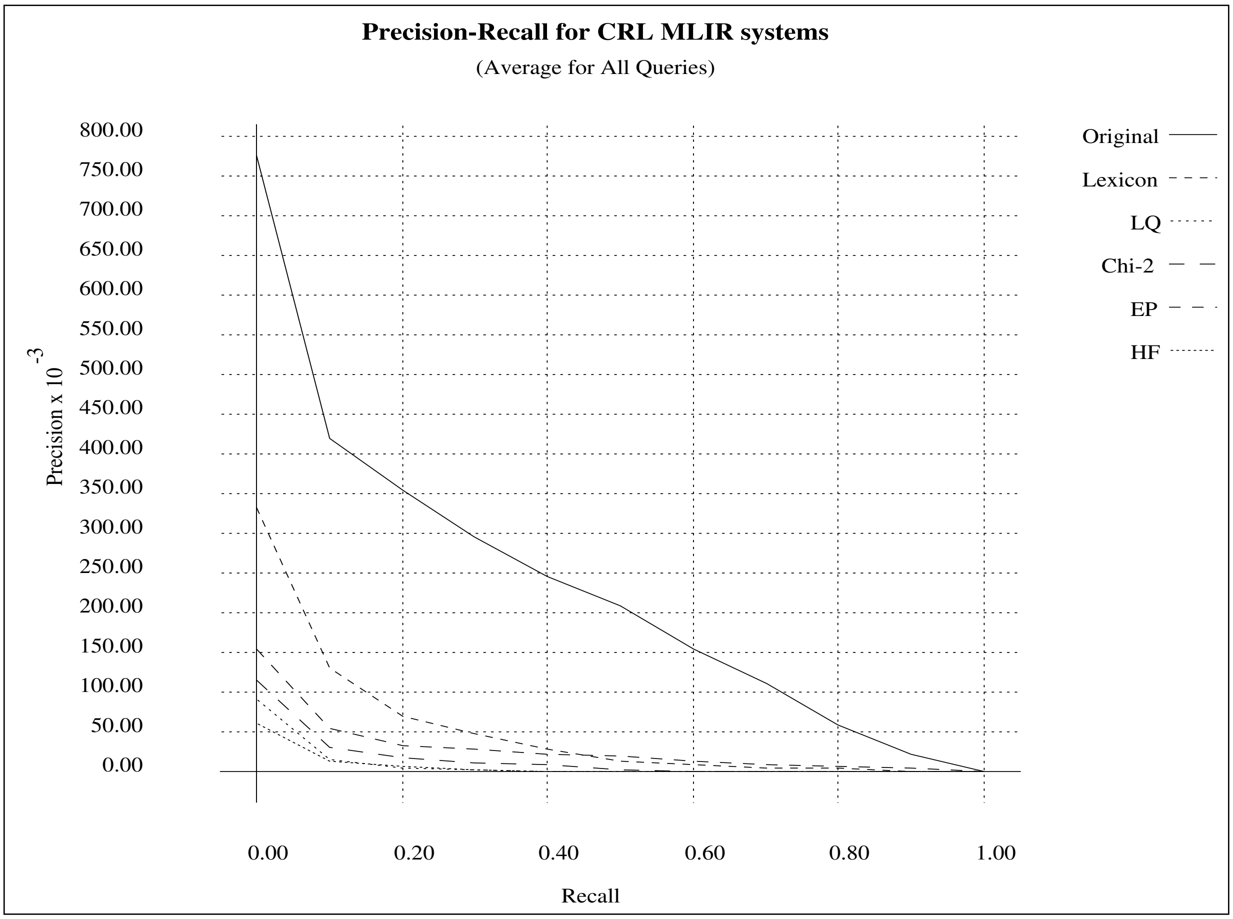

Figure 4. Average precision-recall curves for 25 Spanish queries.

The top line, representing the performance of the original Spanish query shows a competitive TREC text retrieval system. Below that are the lexical transfer curves, followed by the EP method, the Chi-squared approach and the SVD and high-frequency methods, respectively. Of special interest is the bow in the performance of the EP method in the high-end of the recall curve. In several cases (below) the EP method appeared to promote higher precision at higher recall rates than the other methods.

Some Interesting Anomalies

Although the overall results are somewhat discouraging, individual retrieval results show some startling anomalies. Figure 5 shows the precision-recall curve for queries 30, 36 and 44.

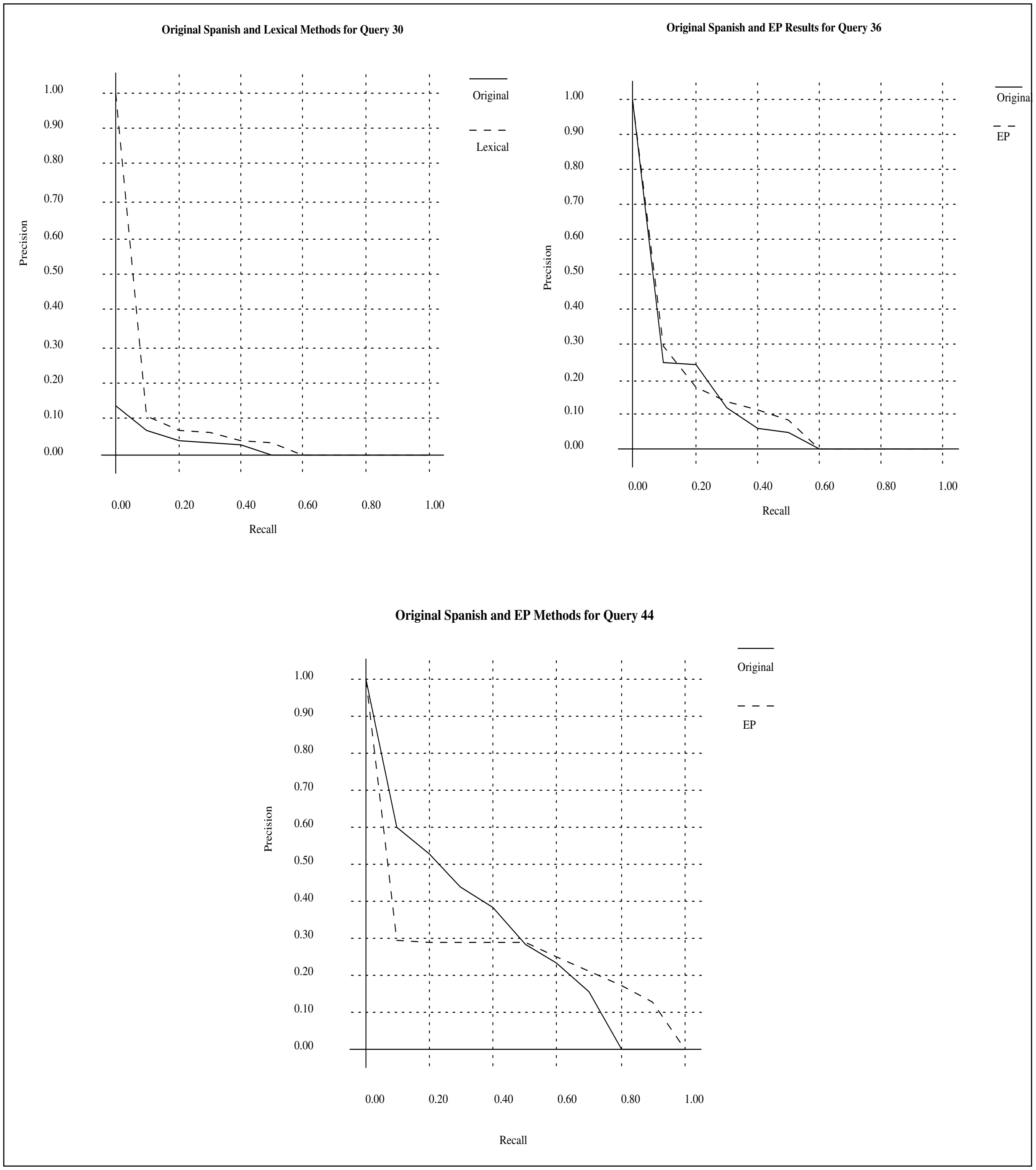

Figure 5. Anomalous results for Queries 30, 36 and 44.

The precision-recall curve for query 30 shows a situation where the lexical methods actually outperformed the original query substantially. The Inquery performance on the original query is probably suspect here; it may very well be that the Spanish morphology in Inquery failed to properly stem/expand the key term “deportivos”, while the lexical method generated “deporte.” Query

36 and 44 show the interesting property of the EP method to produce relatively high precision at higher rates of recall. Why this happens is unclear at the moment.

CONCLUSIONS AND FUTURE WORK

The five query translation methods proposed in this paper all utilize slightly different approaches to the problem of automatic query translation. The only approach not covered is a full machine-translation system. Although the overall results are not encouraging, the several anomalous results reported here do suggest that under certain, as yet to be determined, circumstances lexical and EP-based methods may outperform or perform as well as original queries. Future efforts will therefore focus on the following areas:

- Establishing an on-domain corpus of parallel text. In a setting where translators are constantly creating new translations, the availability of on-topic parallel texts is common. In our pursuit of a MLTR system for real-world applications, we therefore believe that the availability of on-domain corpora is a realistic assumption.
- Using the Inquery text retrieval engine for both query translation and results generation.
- Combining lexical and EP methods to gain the best properties of both systems. Early trials (Davis and Dunning, 1995) demonstrated that such a system appeared to work well when evaluation was only over a novel section of a corpus that was also used for training. Further evaluations are needed, however, at the level of TREC.
- Continued research on the SVD methods.

Finally, further examination of the TREC-4 results is needed. The results presented herein were only available five days prior to the deadline for this paper, limiting the depth of analysis that we could perform on the results. Closer examination of the results will likely suggest additional ways that the query translation methods can be improved.

ACKNOWLEDGEMENTS

The Inquery system runs were graciously donated by the University of Massachusetts at Amherst information retrieval research group. CRL is indebted to UMASS for taking the time to run and submit the numerous query result collections to NIST. We also wish to thank NIST for the Herculean effort to evaluate the entire collection of five runs of queries.

This research was funded under contract number MDA904-91-C-6153 from the United States Department of Defense.

REFERENCES

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.

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.

Dunning, T. E., and M. W. Davis (1993a), A Single Language Evaluation of a Multi-Lingual Text Retrieval System. In NIST Special Publication 500-207: *The First Text Retrieval Conference (TREC-1)*, ed. D.K. Harman, Computer Systems Laboratory, NIST.

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.

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

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.

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.

Salton, G. (1970). Automatic Processing of Foreign Language Documents. *Journal of the American Society for Information Sciences*, 21: 187-194.

Salton, G. and M. J. McGill (1983), *Introduction to Modern Information Retrieval*. New York: McGraw Hill.