

Harman, D. K. (1993). Overview of the First Text Retrieval Conference. In NIST Special Publication 500-207: *The First Text Retrieval Conference (TREC-1)*, ed. D.K. Harman, Computer Systems Laboratory, NIST.

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.

Leacock, C., G. Towell, and E. Voorhees (1993). Corpus-Based Statistical Sense Resolution. In *Proc. of the ARPA Workshop on Human Language Technology*, Section 8. Princeton, New Jersey.

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.

Yarowski, D. (1992). Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. 1992 Conference on Computational Linguistics (*COLING-92*).

## 6 CONCLUSIONS

Disambiguating queries derived from MRDs appears to work well when the original query is closely related to the document collection. For queries that are not closely related to the documents, the results are unclear. These tests were hampered by several factors, all of which would seem to bias the results against the disambiguation approach. In particular, the aligned training corpus was minute and standard alignment programs had severe difficulties with the noisy nature of the parallel corpus. Nevertheless, the fact that the disambiguation of query translations results in improvements over the use of raw MRD entries in the cases where the queries correspond to the domain of the corpus supports the value of corpus-based disambiguation in MLIR systems.

Evolutionary optimization for discovering optimal queries using a parallel training corpus does present difficulties for "on-line" IR systems. On a Sun 4, query disambiguation took an average of five minutes, making this approach perhaps most attractive for batch processing of query requests or in automatic document routing applications where standing queries are used repetitively.

## 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. To appear in *Proceedings of the Conference of the European Chapter of the Association of Computational Linguistics*. March 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*, MCCA-93-252, Computing Research Laboratory, New Mexico State University.

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.

Gale, W. and K. Church (1991), A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the Association of Computational Linguistics*,

Gale, W., K. Church, and D. Yarowski (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Statistical Research Report 104*, AT&T Bell Laboratories.

Best and Worst Fitness Scores for Query 1 Over 20 Generations

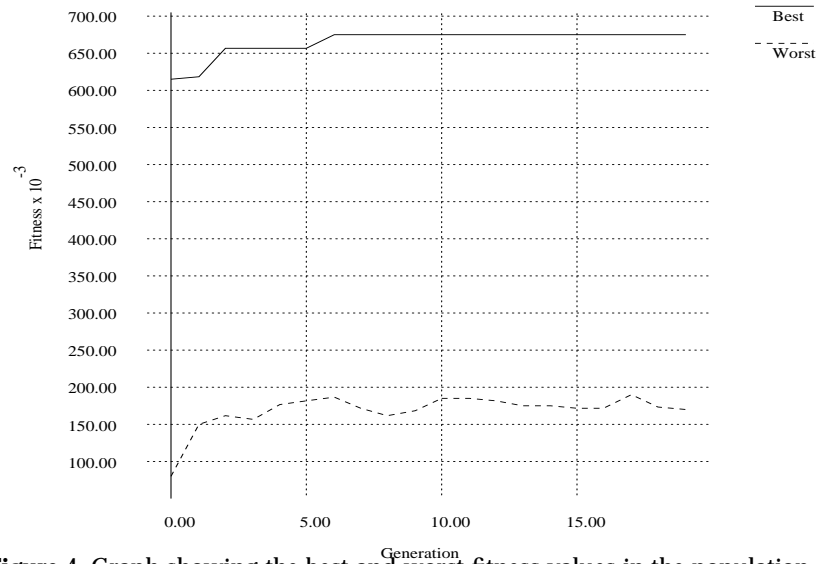


Figure 4. Graph showing the best and worst fitness values in the population over 20 generations of evolution. Fitness gains are extremely modest, except in the early generations. No further improvements were found beyond generation 20.

## 5 RELATED APPROACHES

Although the availability of machine readable dictionaries presents an easy path for deriving hypothesis translations of queries, MRDs are by their nature very broad and lack much of the domain-specific and technical terminology that is found in a body of documents like the PAHO corpus. The direct application of the parallel corpus to filling out the hypothesis queries is not subject to these limitations. An approach that substitutes the terms in the English query parallel texts directly into the population of Spanish queries, and then chooses among the Spanish hypotheses using the evolutionary programming approach described above, would eliminate any reliance on machine readable dictionaries.

A generalization of this approach is to regard the disambiguation process as a score-matching procedure on retrievals over the parallel training corpus. Given a collection of 40,000 unique terms in a document collection, a transformation matrix consisting of 40,000 equations in 40,000 variables can be devised (assuming linearity in the translation process). In the evolutionary approach presented in this paper, the fill-in of the matrix was effectively restricted to terms that also occurred in the bilingual dictionary, substantially increasing the sparseness of the matrix. Singular-value decomposition techniques may also be an alternative for the manageable calculation of the translation matrix.

**Table 1.** Order of documents in the Spanish query results. Numbers in the table give the order in the query results of the parallel documents when compared to the English query results. The “All” column for each query shows the results of applying all senses of each English term, without disambiguation. The “Dis” column shows the results when corpus-based disambiguation has been performed on a population of Spanish queries.

Eng. Doc	Q1		Q2		Q3		Q4	
	All	Dis	All	Dis	All	Dis	All	Dis
1	15	5	4		1	1	1	10
2	2	1	1	1	7	3		
3	6	2			14	13		
4	8	3			4	4	7	
5		7	20		9	11		14
6	7	15						
7					6	5		
8	3	4	13			16		
9	5	11		6	3	2		
10		13		4	4	10		3
11		16				7	5	1
12	1	12			11	8		
13	20				2			
14				8		9		
15							17	
16	14	17				20	15	
17			9					
18			11		20	6	3	4
19								
20								
Total	10	12	6	4	11	14	6	5
Rank Score	5.10	2.75	6.50	4.00	4.27	4.50	4.50	9.0

Q1: intestine parasite rate and the health official response from Caribbean country

Q2: example of entertainment for educate on health issue that use sing and dance

Q3: uterine cancer and the growth of woman support structure in Honduras

Q4: meeting and conference on workplace and factory safety

**Figure 3.** Four queries, written in their stemmed forms, that were used as test cases for evolutionary disambiguation. The stemming to morphological base forms was performed by hand. The queries were formulated by hand based on examination of the subject matter of the PAHO corpus.

derived Spanish query should retrieve exactly the same set of documents as the English query, assuming that the parallel texts are translations. A second measure of the success of a translation is the closeness of the relative ordering in the Spanish results to the English document order. In Table 1, disambiguated queries perform 10% and 15% better on Q1 and Q3, respectively, but perform 10% and 5% worse on Q2 and Q4, using the total number of documents shared between the two retrievals as the measure. The poorer performance of Q2 and Q4 after disambiguation is likely related to the paucity of retrieval results for these queries. If a query is not related to the majority of documents in a corpus, the spectrum of documents retrieved will be ranked primarily according to the presence of high-frequency terms in the query. Removing terms from a query during a disambiguation phase will almost always reduce the performance of the subsequent retrieval in this case.

The similarity of the rankings was measured using the sum of the distances of the entries in the table from the correct rankings. The results are shown in the bottom-most row of the table. The sum of the distances has been normalized by the total number of correct entries. Q1 and Q2 have significantly lower rank-sum distances for the disambiguated queries. Q3 shows comparable results for both queries, although the disambiguated query did slightly poorer despite the higher number of overall correct occurrences.

Figure 4 shows the best and worst fitness values of the population for query, Q1, over twenty generations. The fitness gains were rapid and modest, which closely parallels the relatively modest gains for Q1 due to disambiguation in Table 1.

tigations into improved methods of aligning parallel texts (Davis, Dunning and Ogden 1995) have the potential to dramatically improve the quality of the results presented here, but for the work described here, highly accurate alignments were not yet available.

Of the 181 documents pairs, the first 30 documents became the training corpus for disambiguation and translation of the English query. The remaining 151 documents then served as a “novel” corpus over which the retrieval results for the derived Spanish query could be compared to those of the English query.

A population of 200 hypothesis Spanish queries was generated by randomly assigning a single term from a sense entry obtained by looking up each English query term in the Collins English-Spanish bilingual dictionary. It is important to note that each sense entry in Collins for a given headword contains other information besides just potential single-word translations. Phrasal translations are very common and examples of usage patterns are often provided in the Collins dictionary. No distinction was made between usage examples, phrasal entries or single headwords for this experiment, although part-of-speech tags and other lexical information were removed.

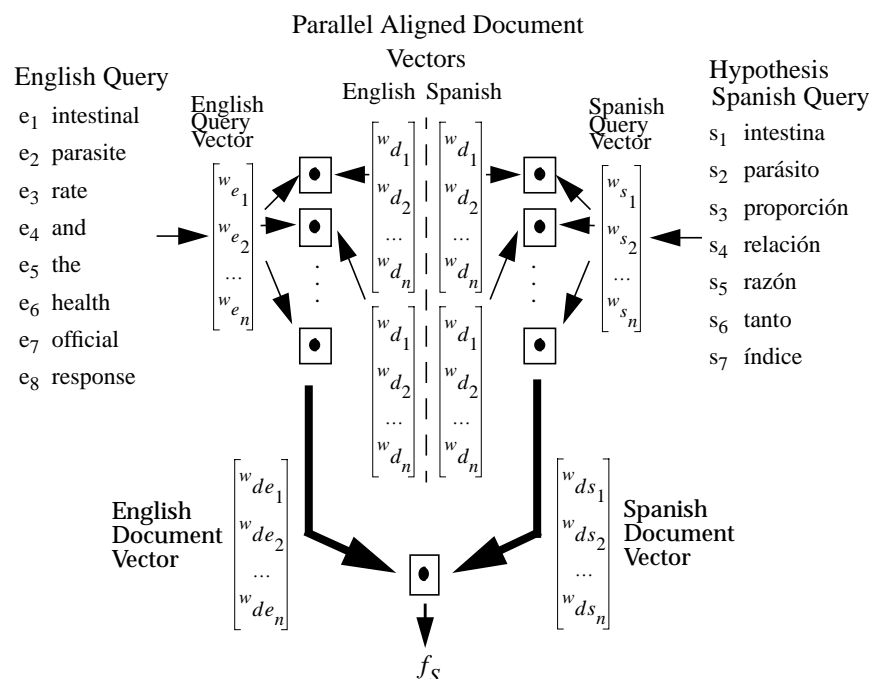
Term frequency-Inverse document frequency (IDF) weighting was used (Salton 1983) for the vector weights. IDF weights are calculated based on the frequencies of a term in the corpus and the number of documents in which the term appears. The term frequency-IDF weight,  $w_{TF-IDF}$ , for a document is

$$w_{TF-IDF} = f_{ik} \log_{10} \left( \frac{N}{n_k} \right)$$

where  $f_{ik}$  is the frequency of the  $k$ th term in the  $i$ th document of the corpus,  $N$  is the total number of documents in the collection and  $n_k$  is the number of documents containing the  $k$ th term. IDF weighting schemes like this penalize terms that occur throughout the document collection ( $n_k \approx N$ ) and exaggerate the scores of terms that are comparatively rare.

Figure 3 shows the four queries that were evaluated. Terms in each query were submitted as morphological base forms (stemmed) to the translation engine. This facilitated easy comparison with the headwords in the bilingual dictionary. Automatic stemming technology could equally well be used so that grammatically correct English queries can be accommodated without loss of generality.

The results of the queries are shown in Table 1. The left-most column shows the top twenty documents retrieved by the English query over a novel data set. The higher the document is in the list, the more relevant it was judged by the vector IR system. For each query, two columns are shown giving the corresponding positions of the English documents in the Spanish retrieval results. If a table cell is empty, that document was not found in the top 20 documents retrieved by the Spanish query. The column labeled “All” shows the retrieval results with no disambiguation. The second column shows the ranking of the documents by the disambiguated query. The total number of retrievals that a Spanish query shares with the English query is one indication of the success of the translation. Ideally, a



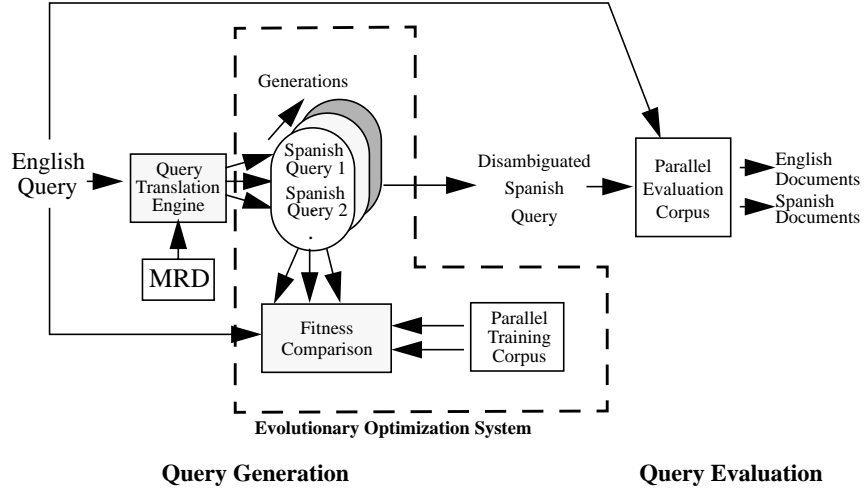
**Figure 2.** Schematic showing the fitness calculation of a hypothesis Spanish query and an original English query. The English query (at left) is transformed into a query vector by a weighting scheme. The cosine dot product of the query vector with each document in the training corpus results in an English Document Vector. Similar operations are performed on the hypothetical Spanish query (at right). The cosine dot product of the resulting document vectors is the fitness (at bottom)

Where  $T$  is the translation operator. For a query containing 10 words, each of which has an average of three senses with five terms each, this number is  $15^{10} \approx 5.8 \times 10^{11}$ . Given such a large combinatoric space, evolutionary programming methods are a potentially useful tool for optimizing queries.

An atomic mutation step in this system consists of picking an English term at random, and then adding or deleting a single Spanish word which appears in any translation or example for the chosen English term. The severity of the mutation operation is therefore based on the number of single mutations applied to a given Spanish query.

#### 4 RESULTS

For the current effort, the system used the Pan American Health Organization (PAHO) corpus for both the training and the novel document set. The PAHO corpus consists of 181 English and Spanish document pairs, with approximately 600,000 English and 600,000 Spanish words. The documents all concern Latin American health issues, conferences, and meetings. The documents were automatically aligned at the sentence level by a variant of the algorithm by Church and Gale (1991). The quality of alignments produced by these methods was often very poor. Subsequent inves-



**Figure 1.** Diagram of the query translation and evaluation process. Note that the training corpus is different from the evaluation corpus. The latter is only required for determining the accuracy of the derived Spanish query on a novel document collection. In a real system, the Spanish query would be applied to monolingual Spanish collections.

$$w_{mj} = Q_j \bullet D_m = \frac{\sum_i q_{ji} d_{mi}}{|Q_j| |D_m|}$$

High scores are assumed to indicate correspondingly high degrees of “conceptual matching” between the query and document. The measure of the fitness of a given  $S_j$  is:

$$f_j = \Delta_E \bullet \Delta_{S_j}$$

Where  $\Delta_E$  is the document score vector for English query  $E$  over the English half of the parallel training corpus. Figure 2 shows schematically the process of calculating the fitness of a Spanish query using a training set of parallel documents. Beginning with the English query and a hypothesis Spanish query derived by dictionary lookup, a query vector is generated by frequency-weighting the query terms. The dot product between the query vector and each of the document vectors in the training corpus then produces a document score vector. The product of these document vectors is the fitness.

The number of possible unique sense combinations, and therefore hypothesis Spanish queries, is

$$\prod_{e_k \in E} |T(e_k)|$$

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 information 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.

### 3 AN EVOLUTIONARY APPROACH

Bilingual machine-readable dictionaries (MRDs) provide a groundwork for constructing translated queries. The limitations of MRDs are that most words have multiple senses (polysemy) in the MRD and that the domains spanned by most MRDs are limited. Polysemous terms can be disambiguated to a degree through the use of information retrieval over a sense-tagged training corpus. Leacock, *et. al.* (1993), expanding on work by Yarowsky (1992) and Church, *et. al.* (1992), demonstrated that IR systems can achieve disambiguation results comparable to Bayesian methods and supervised neural networks, although they also demonstrated that the IR methods produced failures of a different type than the other two approaches.

The method proposed here expands on the IR approach to term disambiguation by choosing a collection of Spanish senses from the MRD translations of an English query based on the similarity of the Spanish and English query results. This approach to disambiguation uses evolutionary programming (Fogel 1991) to iteratively optimize the Spanish query until it matches the results of the original English query on a parallel training corpus. Figure 1 diagrams the overall process.

An English query,  $E$ , composed of terms  $e_k$ , can be translated using an MRD into a collection of possible unique Spanish queries,  $S_j$ , by assigning one term from a sense from the MRD for each  $e_k$ . The result of a retrieval, English or Spanish, is a vector of documents and scores. Given document vectors,  $D_m$ , in a training corpus, a document score vector,  $\Delta$ , contains a score,  $w_{mj} = Q_j \bullet D_m$ , for each of the documents with queries  $Q_j$ , which may be either English or Spanish. The dot-product operation is the cosine dot product in this case:

use short phrases in addition to single words are as effective as any others. (Harman, 1993).

In general, the problem of translating the user's needs into queries for such systems is central to the field of information retrieval [Salton, 1970]. The query may contain 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 information retrieval or MLIR). 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 the second approach. It combines a bilingual dictionary with a set of training documents (corpus) which have been previously translated. This corpus or example-based approach has advantages in that it is inherently sensitive to language as it is actually used. Furthermore, if the parallel training corpus is similar to the overall collection of documents then specialized terminology is likely to be handled well. The system described here uses an objective evaluation of the results of trial retrievals in an evolutionary framework to iteratively improve the translated queries.

## 2 APPROACHES TO QUERY TRANSLATION

Salton (1970) first demonstrated that information 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 again 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

# Query Translation Using Evolutionary Programming for Multi-lingual Information Retrieval

Mark W. Davis and Ted E. Dunning

## Abstract

Multi-lingual information retrieval (IR) systems apply queries in one language to a document collection in several different languages with the goal of retrieving only those documents relevant to the query. At first glance, deep linguistic analysis and translation of the query appears necessary before retrievals can be performed. IR systems are unique in natural language processing, however, because a pattern of term occurrences in a document generally suffices to determine the subject matter; word order is largely irrelevant. Translated queries are therefore primarily derived by a mapping from a word set in the query language to a word set in the language of the derived query. Large parallel text collections with sentence-level alignments can provide a baseline for evaluating the correctness of a query translation, but the determination of members of the query translation remains problematic. Constructing a query from machine-readable, bilingual dictionaries and assigning term weights by the evolutionary optimization of a population of potential weighting schemes presents a solution to the difficulties of generating translated queries. In this approach, differences in the rank statistics on the comparative recall results for a query against its native language and its translation against its native language determine the fitness of a tentative query translation.

## 1 INTRODUCTION

The goal of information retrieval (IR) is to retrieve documents that are closely related to a user's needs. For performance and simplicity, most systems avoid sophisticated linguistic analysis of the documents.

Some systems go so far as to consider documents to be unordered bags of words. These documents are related to a query which is itself composed of unordered words to which numeric weights are attached to indicate importance. Systems which primarily combine these weights from the query with word counts from the documents in a linear fashion are called vector-based retrieval systems. The relevancy of a document to a query is determined by some sort of inner product on the document and query vectors. Until recently, such systems represented the state of the art in retrieval for moderate to large sets of documents. Even now, systems that