NEW EXPERIMENTS IN CROSS-LANGUAGE TEXT RETRIEVAL AT NMSU'S COMPUTING RESEARCH LAB

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

ABSTRACT

In Cross-Language Text Retrieval, queries in one language retrieve documents in other languages. Query translation is the least expensive approach to the retrieval task when compared to full document translation. The simple combinatorial properties of vector-based text retrieval systems simplify the translation task enormously, reducing most translation to the correct substitution of equivalents from a bilingual lexicon or corpus. New experiments are presented on methods for selecting among potential equivalents from a bilingual lexicon, including one fully-automatic method that achieves 73.5% of the performance of a monolingual system operating on the same retrieval task.

Introduction

In Cross-Language Text Retrieval[1] (CLTR), queries in one language retrieve documents in other languages that are related to that query. Between translating all of a document corpus into one single language prior to indexing or, alternatively, simply translating the user query at query time, translating just the query is generally considered the least expensive in terms of resources, and potentially more accurate given the current state of machine translation technology.

Early experiments by Salton (1971) demonstrated that CLTR could do as well as monolingual approaches given certain experimental constraints. Primarily, this meant preparing a transfer dictionary in advance that contained precise translations of terms in the query language. Homographs and polysemous terms were not a significant obstacle because the terminology in the dictionary was disambiguated by a human in advance of the retrieval experiment. The added problems introduced by translation pragmatics similarly dissolved.

In recent years, however, it has become apparent that the issues in practical, fully-automatic CLTR systems are substantially more complex than originally conceived. Foremost among these issues is the question of whether the linearity of vector-based retrieval systems leads directly to the application of term-for-term translations. This issue is already being answered by the realization that phrases are not always reducible in machine translation or CLTR systems (Hull and Grefenstette, 1996). A related issue is whether the information retrieval model makes corpus-based term disambiguation practical. Thus far, only mixed results have been achieved for large-scale evaluations of CLTR systems, although TREC multilingual corpora have made further studies much easier (Davis and Dunning, 1995; Hull and Grefenstette, 1996). In recent years, NLP tools like parts-of-speech taggers have improved to beyond the 90% performance level. For CLTR systems, this means that these tools are no longer an unknown quantity, and any performance gains due to using them should become less ambiguous.

Parallel corpora have been shown to be useful for disambiguating monolingual term senses in limited tests (Leacock, Towell and Voorhees, 1993). Parallel corpora have also been used for training statistical text models for translation (Church and Mercer, 1993), and parallel corpora have been implicitly applied to the CLTR disambiguation problem by Landauer and Littman (1991) who generated query translation matrixes using Latent Semantic Indexing. Davis and Dunning (1994, 1995)

[1] At SIGIR 96, participants in the Cross-Linguistic Information Retrieval Workshop voted to refer to text retrieval with differing query and document languages as Cross-Language Text Retrieval to reduce some of the confusion that other names have caused in the past. I follow this convention in this paper.

applied evolutionary programming methods to attempt to refine Spanish translation of English queries by iteratively comparing the retrieval profiles of English and Spanish queries over a parallel corpus. In Davis and Dunning (1994), a transfer dictionary was used to create the Spanish queries, but no large-scale retrievals were performed, and the later work (Davis and Dunning, 1995) used initial Spanish equivalents derived directly from a parallel corpus. Results from the latter were shown to be comparatively poorer than even the full transfer dictionary methods. In both cases, the evolutionary optimization methods were computationally expensive, requiring around 50,000 retrievals per query to achieve acceptable levels of optimization.

An alternative model of CLTR using parallel corpora is to attempt to disambiguate the Spanish equivalents by comparing their retrieval results one at a time against the English query retrieval results as a whole. Vector-based retrieval models use linear combinations of term occurrence features. As a result, the subspace of the Spanish term-document space projected along an axis of a given equivalent may be adequate for determining the correct equivalent for an English term.

A yet further alternative is to make use of NLP tools like parts-of-speech (POS) taggers. Wilks (1996) has suggested that POS tagging may be combined with full lexicons to disambiguate up to 95% of English. The role of a tagger in a CLTR system then becomes pre-selecting equivalent sets for each query term based on the POS of the tagged query terms. Corpus-based methods can then be applied to the remaining equivalent set where ambiguities still exist to further refine the translation.

In this paper, results are presented for the disambiguation of transfer dictionary equivalents using parallel corpora, equivalent sets selected by matching parts-of-speech (POS) generated by an automatic POS tagger and a bilingual lexicon, and combined results from the two methods. Disambiguation appears to be effective even without the added complexity of considering the entire range of possible equivalent combinations as was done in the evolutionary programming models. The effectiveness of corpus-based disambiguation alone appears mixed in these experiments. The best performance is to be found in disambiguating POS-tagged queries over a parallel corpus, achieving 73.5% of the performance of the original monolingual queries on the same retrieval task. The addition of corpus-based disambiguation represents 6% of this figure, despite the fact that the parallel corpus was drawn from distinct domains of documents.

The experiments reported in this paper are all based on making use of English, human-produced translations of TREC Spanish topic descriptions. The CLTR system translates these query descriptions into Spanish in a fully-automatic manner and the new Spanish queries are run against the TREC Spanish document corpus. The best we could reasonably hope for would be that the automatic query translations performed at least as well as the original Spanish queries over the same document set. The performance of the monolingual Spanish topics then serve as a comparative baseline for automatic translation methods. In an operational setting, a CLTR system user would be creating, for example, English queries to retrieve documents in multiple languages. The translation process would convert the query into the range of languages that are represented by the document copora of interest to the user, and the retrieved documents could then be submitted to a translation staff or to a machine translation system for a quick gloss of the document contents.

Recuerdo: A Spanish Retrieval Engine

In order to perform our CLTR experiments, we needed a retrieval engine with competitive performance characteristics. The system also needed to be able to operate over parallel corpora for disambiguation in addition to working on a monolingual document collection. Current Spanish monolingual retrieval systems are primarily vector-based (using variants of tf-idf document and term weighting), inference-net based, and derived from logistic regression of a retrieved document set. The flexibility of the vector-based tf-idf approach suggested that it was a reasonable approach. Further, a vector model is an inherently linear combination of term weightings, making the substitutions of term equivalents in a CLTR scenario straightforward, with special handling of phrasal components an added option that can be accommodated easily without significant modification of the system.

The *Recuerdo* system developed at CRL has some substantial modifications over the Smart system from Cornell. Among these was the development of new Spanish stemmer based on the Porter stemmer model that contains 145 rules for stemming Spanish terminology. The complexity of irregular Spanish verbs was partially handled within this framework, although it was decided to do without specifying irregular verb paradigms precisely to maintain the speed of the stemming algorithm. The effectiveness of this approach has only been tested within the framework of the retrieval experiments presented in this paper.

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

For CLTR applications, the system can read multiple indexes for parallel texts, and can perform comparisons between retrieval results for queries across parallel corpora using either a transfer dictionary or the direct extraction of equivalents from the parallel corpus. The system can also perform term expansions by finding the subset of terms it has encountered at index time that have up to a certain number of character differences with the source term. This fuzzy matching capability is used for finding cognates of query terms in CLTR where no dictionary or parallel corpus term is available.

For the system to operate in a fully CLTR mode, it is necessary to supply kill lists in both query source and target languages, transfer dictionaries, an indexed target collection and indexed parallel text collections. For these English-Spanish CLTR experiments, the Collins bilingual dictionary was used as a transfer dictionary and one year of the UN parallel corpus was used as a parallel text collection. POS-tagged queries were handled by a second version of Collins that included POS groupings of lexical items.

Collins Bilingual English-Spanish Dictionary

Collins is a comprehensive bilingual dictionary containing around 50,000 English headwords. For this experiment, English headwords and a subset of the collected *equivalents* and sense discriminating terminology were extracted. Equivalents from homographs and discriminating terms were conflated after case normalization and Porter English headword stemming. Duplicate equivalents were not removed from the conflated term set. The Spanish equivalents were case normalized and stemmed using a Spanish variant of the Porter stemming algorithm developed at CRL. For this experiment, phrasal headword entries were also discarded.

After this preprocessing, 23,932 English headwords remained with an average of 1.394 equivalents per headword (variance of 0.648), with the largest headword having 16 equivalents. This set was checked by a Spanish-fluent graduate student against the original Collins entries, who added missed equivalents to English headwords that also appeared in the queries. The student was provided only the pooled terms from the 25 TREC Infosel queries and was instructed to make certain that the equivalent sets were complete.

A second transfer dictionary was prepared for use in the POS-tagging experiments. The Collins markup for nouns, transitive and intransitive verbs and adjectives were mapped onto the Penn Treebank POS tagset by conflating all Collins nouns to NN tags, all Collins verbs to VB tags and all Collins adjectives to JJ tags. Headwords with multiple parts-of-speech became separate lexical entries in the resulting dictionary, with the headword prefixed by the tag and an underscore.

The UN Parallel Corpus

The 1991 UN parallel documents were automatically aligned (Davis, Dunning and Ogden, 1995) resulting in 97,594 alignment pairs at the sentence or double-sentence level. The English documents contained 91,915 unique terms out of a total of 4,483,677. On the Spanish side, there were 122,827 unique terms in a total of 5,259,124. The alignment process has previously been estimated to be 83% correct, although a comprehensive evaluation of the UN alignments was not performed.

The 1991 UN document set was chosen because it was suspected that current issues might be better represented by the most current document set from the UN collection which includes years 1988 through 1991.

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

The MITRE Parts-of-Speech Tagger

For the POS-based experiments, the MITRE English POS tagger was applied to TREC-5 queries. The English description field of the SGML markup was modified slightly by adding <s> and </s> tags at the beginning and end of the field. For example, Trec-5 Spanish query 56 is:

```
<top>
<num> 56  </num>
<title> Swine Fever
<E-desc>
How has the threat of swine fever affected
international trade?
</E-desc>
<E-narr> Narrative:
A relevant document will contain information
detailing some effect caused by
fears of swine fever.
<S-desc>
¿Qué efecto ha tenido en el comercio inter-
nacional la enfermedad "fiebre
porcino?"
```

```
</S-desc>
<S-narr> Narrative:
Un documento relevante tendrá información
sobre el efecto que ha tenido el
temor a la fiebre porcino en el comercio
internacional.
</top>
```

The tagged version of the query is (just the <E-desc> field is shown for brevity):

```
<top>
<num> 56  </num>
<title> Swine Fever
<E-desc><s>
<lex pos=WRB>How</lex> <lex pos=VBZ>has
</lex> <lex pos=DT>the</lex> <lex
pos=NN>threat</lex> <lex pos=IN>of</lex>
<lex pos=NNS>swine</lex> <lex pos=NN>fever
</lex> <lex pos=VBD>affected</lex> <lex
pos=JJ>international</lex> <lex
pos=NN>trade?</lex>
</s></E-desc>
...
</top>
```

A further filtering step was performed by collapsing the spectrum of noun and verb tags generated by the MITRE POS tagger to JJ, NNP, NN,VB, CD and FW, and then prefixing each query term with the POS tag. Other tags were eliminated. For the query above, this resulted in:

```
<top>
<num> 56  </num>
<title> Swine Fever
<E-desc><s>
VB_has NN_threat NN_swine NN_fever
VB_affected JJ_international NN_trade
</s></E-desc>
...
</top>
```

Overall, the performance of the MITRE tagger on the English description fields was very good. Among the tagset that remained, there were 8 errors by the MITRE tagger over a total of 222 labelled terms in 25 queries, resulting in a 3.6% error rate on query tagging. Notable errors included: "in" was identified as a Foreign Word (FW), "steps" was incorrectly identified as a Verb (VB) twice, and "extinction" was identified as a Verb (VB) once.

Disambiguation and Retrieval

Figure 1 diagrams the method used to retrieve the baseline retrieval results. The Spanish monolingual query retrieves Spanish TREC documents in this approach.

Figure 2 shows the simple extension to this procedure that involves simply replacing each English query term with all of its Spanish equivalent terms from the Collins bilingual dictionary.

Disambiguation of term equivalents was performed by selecting the best Spanish equivalent or equivalent set for each English query term. In the corpus-based disambiguation experiments, the criterion for equivalent selection was based on examining the distribution of English query terms and candidate Spanish equivalents across aligned parallel texts. For POS-based disambiguation, the approach was to select sets of equivalents from a bilingual lexicon that match the POS of the English query term.

The corpus-based system scored the inner products of weighted document vectors for the English and Spanish retrievals over parallel documents, selecting the term with the highest score. This process thus favored Spanish equivalents that had the most in common with the English query results. This process is diagrammed in Figure 3.

If the English term had no dictionary entry, a fuzzy match was done between the English term and the target retrieval term database to discover potential cognates in the target index. The fuzzy matching process first used an edit distance of zero and, if the term was not found, used an edit distance of two characters.

Adding the fuzzy matching process addressed two fundamental problems associated with general transfer dictionaries: limited coverage and dated material. Specialized terminology is often of neo-latin origin or loan words. In many cases, proper nouns do not have a translation but become loan words in the translation process. The fuzzy matching process makes matching these terms function automatically. If the term is not in the dictionary, an equivalent can often be directly resolved from the target document collection.

The POS disambiguation approach used the MITRE tagger markup of the English query to select among candidate equivalents in the Collins dictionary. This process is diagrammed in Figure 4. NNP tags were handled as special cases of nouns. If equivalents were found in the dictionary under the NN tag category, corresponding to all Collins nouns, then the substitution was performed. Hence, NATO was correctly translated in Collins as OTAN. If no equivalent existed, however, as was the case for proper names, NNPs became their own equivalents. Hence, MERCOSUR translated as itself.

The combined approach is diagrammed in Figure 5. The resulting equivalent sets for each English query term after POS disambiguation were submitted to the corpus-based disambiguation engine in this approach. For the AFP TREC-5 query set, there were 166 unique query terms and 428 equivalents, for an average of 2.58 equivalents per English term. The corpus-based method

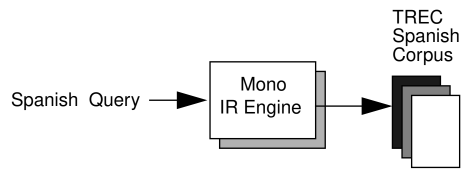

Figure 1: Monolingual Spanish Retrieval for comparison baseline.

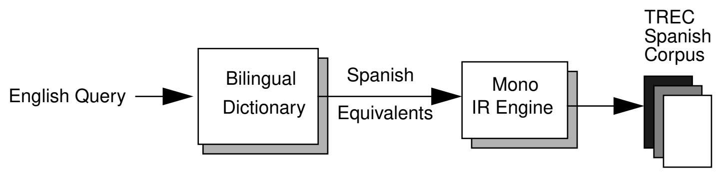

Figure 2: Replacing each English query term with all of its equivalents from the bilingual dictionary to form a new, ambiguous query.

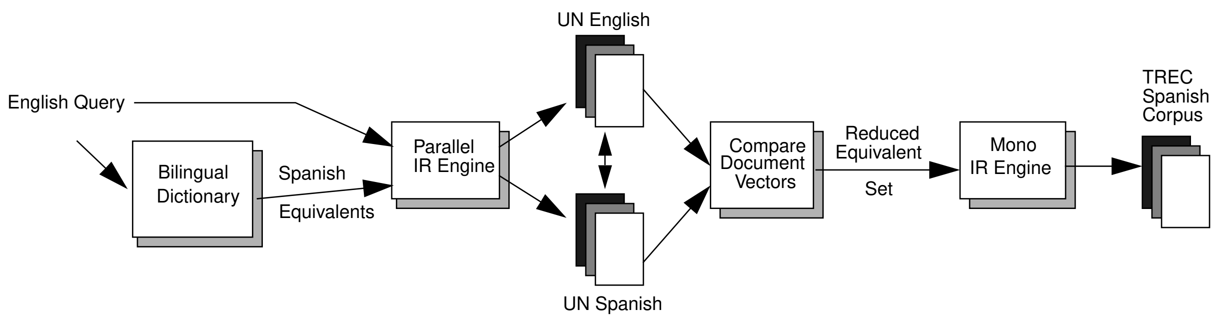

Figure 3: The corpus-based disambiguation method chooses among candidate equivalents for each term of the English query by measuring the similarity of the retrieval results for each equivalent to the English query on a parallel text retrieval task. The derived query is then submitted to the monolingual retrieval system.

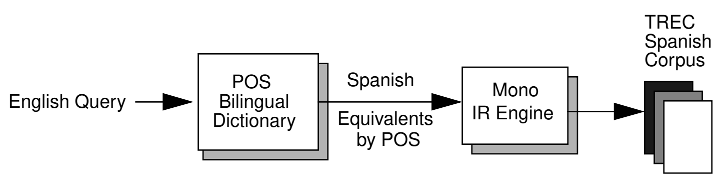

Figure 4: The POS disambiguation method chooses Spanish equivalents for each English query term by matching parts-of-speech in the bilingual corpus.

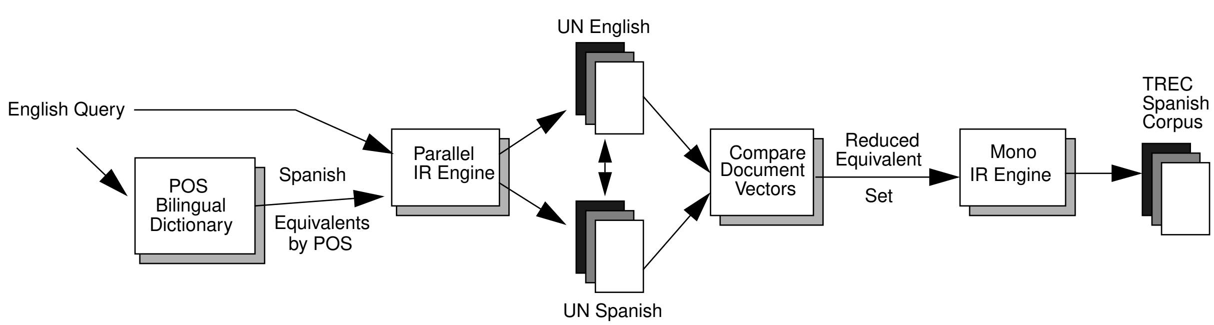

Figure 5: The combined approach uses the POS-diambiguated terms as input into the corpus-based disambiguation engine.

then chose only one of the candidate equivalents for each English query term to form the final query.

Monolingual and Cross-Language Retrieval Results

In order to evaluate the comparative performance of the monolingual system versus the disambiguated queries, TREC-4 and TREC-5 Spanish queries were used. CRL had previously provided English translations of the TREC-4 queries, so the English versions were already available and could be used alongside the Spanish versions. For the TREC-5 queries, NIST provided both English and Spanish versions of the queries.

In the discussion that follows, the monolingual Spanish results will be referred to as MONO4 and MONO5, the all-equivalent substitution approach as ALL4 and ALL5, the corpus-based disambiguation sets as CORP4 and CORP5, the POS approach as POS5, and the combined approach as BOTH5. The POS and BOTH approaches have not yet been tested on the TREC-4 query and document collections.

The pooled query-relevance judgements (qrels) from NIST were used to evaluate the system for several of these runs. It is possible that the stemming algorithm that was used for Spanish might conflate Spanish terms in a manner not represented in the other systems, so the pooled qrels are probably not a perfect measure of the system's performance. There were no other options available prior to direct TREC evaluation, however. This applies to MONO4, CORP4, ALL4, POS5 and BOTH5, since each of these runs was not directly evaluated by NIST.

The performance of the three methods is shown in Table 1. The non-interpolated average precision values are listed by category.

Table 1 Average Precision For All Methods

Method	PR (NI)
MONO4	0.1874
MONO5	0.2895
ALL4	0.0783
ALL5	0.1422
CORP4	0.1250
CORP5	0.1153
POS5	0.1949
BOTH5	0.2127

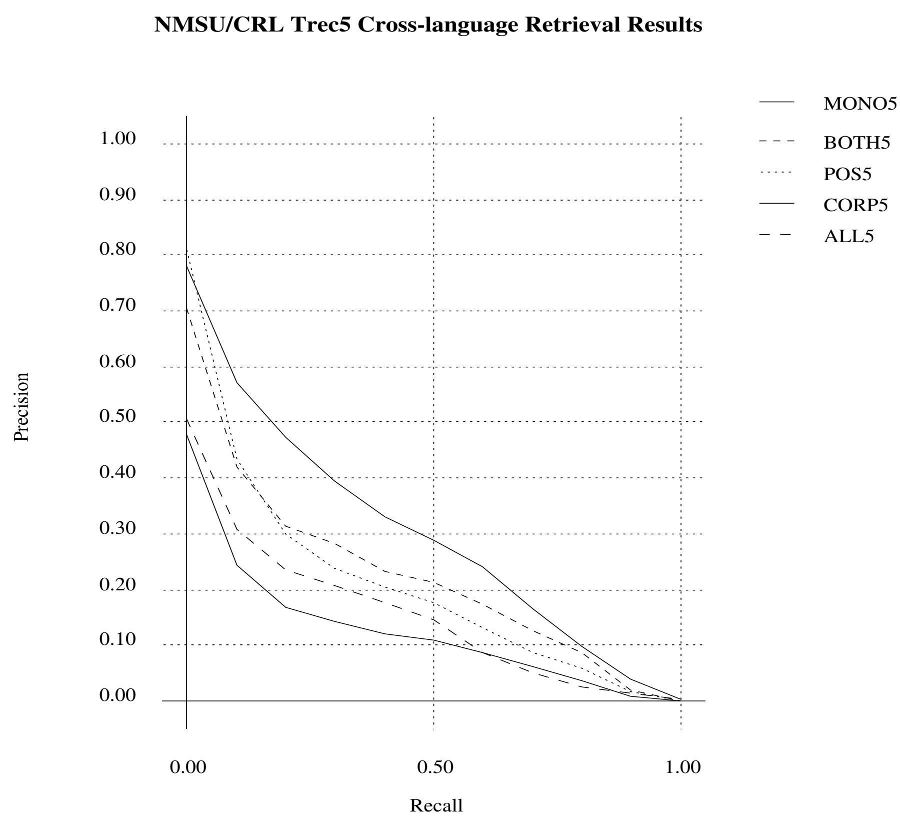

Figure 6: Precision Recall curves for Trec-5 Experiments

The complete Precision-Recall curves for MONO5, ALL5, CORP5 and BOTH5 are shown in Figure 6.

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

MONO4, ALL4 and CORPUS4 results are included because they show a slightly different pattern than the corresponding results for TREC-5. In the TREC-4 results, there was a clear performance gain that resulted from corpus-based disambiguation of the translation equivalents. For the TREC-5 results, however, corpus disambiguation decreased performance when used alone but was advantageous when combined with POS-based disambiguation. Exactly why this occurred is not altogether clear. It certainly must be due to substantial disambiguation errors being made over incorrect POS equivalents present in the TREC-5 query equivalent sets, but exactly why this query set performed so radically different from the TREC-4 set is not immediately apparent and will require further investigation.

Conclusions

Disambiguation of terms in an equivalent set supplied by a bilingual transfer dictionary can result in substantial improvements over most CLTR methods seen to date. The set of experiments presented in this paper provides a clear path to high-performance CLTR systems: combining POS-based disambiguation with corpus-based disambiguation for query translation. Further improvements are possible for the existing system, including accurate identification of phrases in the query that need specialized translations and, perhaps, interactive approaches to translating new terminology and acquiring lexicons for unfamiliar target languages.

Acknowledgements

The POS tagger used in this work was provided by MITRE Corporation. This work could not have been accomplished without it.

This research was funded under grant MDA 904-94-C-6153 of the US Department of Defense as part of the Tipster Reinvention Laboratory.

References

Church, K. W. and R.L. Mercer (1993) “Introduction to the Special Issue on Computational Linguistics Using Large Corpora,” Computational Linguistics, **19**:1. pp 1-24.

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.

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.

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.

Hull, D. and Grefenstette, G. (1996) “Experiments in Cross-linguistic Information Retrieval” in *SIGIR96*, August, Zurich, CH.

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 *Proceedings of the Human Language Technology Workshop*, 260-265, Princeton, NJ. ARPA.

Salton, G. (1971) “Automatic Processing of Foregn Language Documents,” in *The Smart Retrieval System*, ed. Salton, G., Prentice-Hall, Englewood Cliffs, NJ.

Wilks, Y. (1996) *Personal Communication*.