

The issue of model accuracy is clearly in estimating $p(\alpha_j|d_i)$. Better estimates of this quantity can be obtained through estimating the probabilities of Markov chains of increasing order using backoff methods (Chen and Goodman, 1996). Better still, hierarchical models like Prediction Suffix Trees (PSTs) can be used to estimate the probability (Pereira, Singer and Tishby, 1996).

Extending this formalism to CLTR, the entropy measure serves equally well to characterize the probability of a term in source and target languages being matched. From an implementation standpoint, being able to use a subset of documents that are chosen based on term occurrences in the source query language substantially reduces the computational complexity required for LSI or linear methods.

Our ongoing efforts in CLTR include the development of formal, empirical models of CLTR and text retrieval in general. Present research in text retrieval has almost uniformly adopted an *ad hoc* experimental methodology ("we use whatever works"). We believe that further developments in CLTR need substantial theoretical grounding in order to overcome the substantial barriers to improvement. Although the methods sketched above are statistical and empirical in design, formal linguistic models of the translation and retrieval process may perform as well, if not better. At present, however, we feel that the scale and wide coverage of most practical CLTR systems are such that further examination of empirical methods are substantially warranted.

References

Chen, S.F and Goodman, J. (1996) "An Empirical Study of Smoothing Techniques for Language Modelling," in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, June 1996.

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.

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.

Dunning, T. E. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational*

Linguistics, 19, 1: 61-74.

Dunning, T. E., and M. W. Davis (1993), "Multi-Lingual Information Retrieval," *Memoranda in Computer and Cognitive Science*, MCCS-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.

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.

Pereira, F.C, Singer, Y., Tishby, N., (1996) "Beyond Word N-Grams," original paper in *Proceedings of the Third Workshop on Very Large Corpora*, MIT, 1995

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

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.

The performance of the corpus-based methods that we have used based on this collection suggests that even this rather large amount of parallel text is insufficient to provide reasonable coverage if the Infoseek TREC collection. Further developments of corpus-based query translation approaches may very well be dependent on the availability of suitable parallel document sets.

CLTR for the World Wide Web

Recently, we developed a prototype CLTR system for the World Wide Web. *Mundial* is a query interface to Infoseek and Yahoo that takes queries in English, translates them to Spanish and submits the resulting queries to the Infoseek and Yahoo search engines. The *Mundial* prototype uses a bilingual dictionary combined with several heuristics to limit the terminological expansion of the input query. *Mundial* uses the Collins Bilingual dictionary for translating queries, although some of the definitional terminology has been removed by an automatic sorting procedure that reduces each definition to a maximum of two terms. Limiting query size is important because most search engines restrict the size of a query to around 80 characters. Overgeneration in the translation process is handled by using the longest terms (in character count) in *Mundial*. Although in some cases this may be in error, the hope is that automatic stemming of query terms at the search engine will reduce long terms to stems common to many of the keywords that might have been substituted if the entire definition was transferred. The second motivation was that long terms tend to be more precise than short terms, and content words should be as precise as possible during searches.

Mundial may be accessed at: <http://crl.nmsu.edu/ANG/ML/ml.html>.

Towards a New Model for CLTR

The grab-bag of methods that we have applied to the CLTR task share some common similarities in that they all attempt to utilize a parallel text resource to create a translated query. Creating the query requires a selection of terminology from the parallel resource according to a

matching criterion for the original query terms. Although the manner in which this is done may be deep or simple, the commonality suggests that accurate models for *monolingual* text matching are a critical first step in discovering possible bilingual terminology.

We have developed a promising new model of the text matching process for text retrieval that we are currently evaluating in a monolingual setting. In this model, our uncertainty about a collection of words being associated with a document is measured by the conditional entropy of the text and document, conditioned on the collection as a whole

$$H(\Lambda|d_i) = -\sum_{j=0}^{N_\Lambda} p(\alpha_j|d_i) \log p(\alpha_j|d_i)$$

where, as a first approximation, the probability of seeing term α_j in document d_i is the counts of the term in the document over the total counts of the term in collection:

$$p(\alpha_j|d_i) = \frac{N_{ij}}{N_i}$$

The entropy measure does not require arbitrary normalization parameters. It is highest for equal-valued discrete probabilities of term-document occurrences, and increasingly smaller as the term-document to term-collection counts diverge over the term set.

We have implemented a prototype text retrieval system using this model and results are encouraging. On the CACM collection, the following precision-recall curves:

Recall	Precision
0.00	0.7313
0.10	0.6536
0.20	0.5054
0.30	0.4157
0.40	0.3451
0.50	0.2555
0.60	0.1907
0.70	0.1080
0.80	0.0824
0.90	0.0365
1.00	0.0347
Average	0.2839

In this trial, both individual words and word bigrams were indexed after stoplisting with 356 words and stemming with the Porter stemming algorithm. This performance is comparable—if slightly worse—than tf-idf systems

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 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 drawn randomly from the UN parallel corpus.

The query

Indicators of economic and business relations between Mexico and European countries.

became

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

The mean improvement for 25 queries during optimization was approximately 1000% (Figure 1). The performance of the resulting queries when applied to the TREC collection was 60-70% worse than the reference monolingual queries, however, suggesting that the optimization was well-suited to the UN corpus but the results did not transfer to the novel document set. This disparity further suggests that corpus methods in general are highly dependent on the availability of related parallel texts.

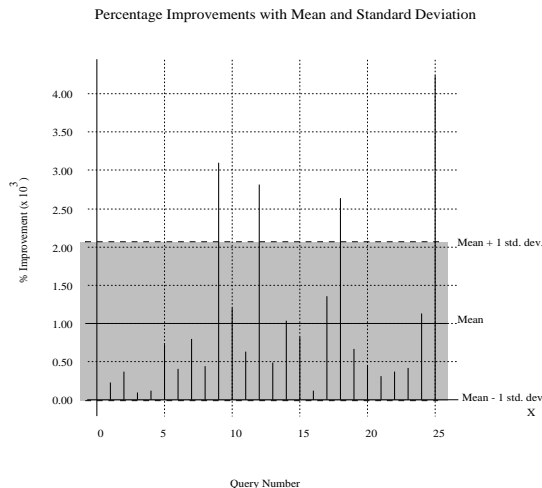


Figure 1: Percentage improvements for TREC Queries 1 through 26 during EP optimization.

Nevertheless, comparing the EP generated query terms to the corpus high-frequency terms, more European country names are prominent, confirming that the method is capturing elements of the original query that are likely valuable in the cross-linguistic domain. This was confirmed by relatively higher precision for this method at the high recall end of the spectrum, when compared to the other methods.

Singular Value Decomposition and the Translation Matrix

We have also evaluated a form of singular value decomposition methods for creating Latent Semantic Indexing methods. The final query translation method was a radical departure from the others, but is derived from earlier work by (Landaure and Littman, 1990) and (Dunning and Davis, 1992). This method is at heart a numerical approach to derive a translation matrix from parallel texts.

In this effort, we applied a *QR*-decomposition technique to reduce the complexity of calculating the singular value decomposition, resulting in query translation that took only a matter of seconds on a SPARC 10.

Due to numerical instabilities in the approach, we feel that the results were substantially incorrect. In TREC-4, the method performed poorly.

Parallel Corpora

For TREC, 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

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

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.

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 avoids normality assumptions and therefore makes better predictions of term distributions in text (Dunning, 1992). Specifically, a likelihood ratio, λ , can be described by a ratio of likelihood functions for binomial distributions

$$\lambda = \frac{\max_{\omega \in \Omega_0} H(\omega; k)}{\max_{\omega \in \Omega} H(\omega; k)}$$

where

$$H(p_1, p_2; k_1, n_1, k_2, n_2) = p_1^{k_1} (1-p_1)^{n_1-k_1} \binom{n_1}{k_1} p_2^{k_2} (1-p_2)^{n_2-k_2} \binom{n_2}{k_2}$$

relates counts of terms (k_1 and k_2) to distribution probabilities (p_1 and p_2). Applying the maximization gives

$$\begin{aligned} -2 \log \lambda &= 2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) \\ &\quad - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)] \end{aligned}$$

where $p = \frac{k_1 + k_2}{n_1 + n_2}$ and $L(p, k, n) = p^k (1-p)^{n-k}$.

To its disadvantage, the binomial model does assume independent word occurrences. To its credit, however, it does not underestimate the probabilities of small count events.

The method we applied extracted 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. An example query translation of

Indicators of economic and business relations between Mexico and European countries.

became

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

In TREC-4 evaluations, this method performed poorly, with an average precision of 0.0109 compared with the baseline monolingual system at 0.2153, or around 95% worse than the baseline system.

A better method for extracting terms would be to test term significance on a document-by-document basis, rather than pooling the terms together. The resulting set of terms could then be pooled. We are currently investigating this approach.

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 algorithm, 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

This approach bears some similarity to the Latent Semantic Indexing (LSI) methods of Bellcore (Landauer and Littman, 1990), although the computational complexity of the iterative methods are amortized slightly differently over multiple passes through the data set.

After developing the transformation matrix, queries can be translated by computing Q' in:

$$Q = TQ'$$

At the time, there were no opportunities for large-scale evaluation of the effectiveness of this query-translation strategy, so our evaluation methodology consisted of applying the resulting transformation matrix to documents in a held-out portion of our training data. Although the results were encouraging, the computational resources needed to iteratively derive the transformation matrix were costly and had only been evaluated over documents from the same text domain.

Lexical Transfer

Starting with the 1994 Text Retrieval Evaluation Conference (TREC- 3), Spanish document collections and query sets have been available for evaluating text retrieval engines. The queries and documents are monolingual, however, so testing a multilingual system is only possible if the query set or the corpus is translated into a different language. We chose to translate the queries since they were very short. With translated queries, a query translation system that produces Spanish queries from hand-translated English versions of original Spanish queries can then be compared against the original queries. The differences between the two results are then a reasonable measure of the effectiveness of the translation process in preserving the characteristics of the original query that contribute to retrieval.

For TREC-4, we submitted one set of results in which we used the Collins English-Spanish bilingual dictionary to build Spanish queries. Individual terms in the English query were reduced to their morphological stems 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, TREC Spanish Query 28 is transformed from

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

to

indicador indicador ayuda expansión previsiones crecimiento comercio comercio 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 substitution scheme, and that “relations” has included the familial sense “parentesco”. The lexical-transfer approach produced Spanish queries rapidly, requiring only a simple database lookup procedure.

While a dictionary 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. Corpus methods do, however, have the comparative advantage that they are current and often contain highly specialized usage patterns that may not be in available dictionaries.

In the TREC-4 evaluation, the transfer queries performed about 50% as well as the monolingual Spanish baseline. This is in sharp distinction to the results reported in (Salton, 1971), where a hand-generated bilingual thesaurus performed about as well as the monolingual baseline.

High-Frequency Terms from Parallel Text

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

As an example, the query

Indicators of economic and business relations between Mexico and European countries.

became

ADVANCES IN CROSS-LINGUISTIC TEXT RETRIEVAL AT COMPUTING RESEARCH LABORATORY

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

Introduction

Cross-linguistic text retrieval (CLTR) is a method for retrieving documents in languages different from the query language. Although it is possible to translate all of the documents into the query language, for large collections the most economical approach to CLTR is to simply translate the query at retrieval time into the document languages. This presupposes that the query can be translated in a reasonably accurate fashion and that monolingual retrieval systems are available for all of the document languages.

As with Machine Translation (MT) in general, query translation in a CLTR system can be done many different ways. An advanced MT system could, for example, perform sophisticated parsing and analysis of the query, derive an interlingual semantic representation and generate a new query from it. At the opposite end of the spectrum, a system could use shallow translation techniques to simply substitute terms from a transfer dictionary, ignoring the ambiguities of polysemous candidates. In shallow translation approaches, the monolingual text retrieval engine operating on the translated query bears the burden of weighting the query terms by virtue of their cooccurrences, hopefully reducing the effect of poorly translated terms. Between these two extremes are a range of approaches.

To overcome the limitations of general-purpose transfer dictionaries, tuned lexicons and thesauri built from controlled vocabulary have been applied to specific text retrieval problems with good success. Despite the growing availability of machine readable dictionaries, however, preparing special-purpose lexical resources remains a daunting task.

Using massive bilingual and multilingual corpora as translation resources is another approach that has the potential to overcome the limitations of the shallow methods, while still requiring less resources than the tuned lexical methods or the deep semantic MT approaches. Text corpora contain examples of usage patterns in the query language that can be matched to examples in the target language if the sentences or para-

graphs of the texts are aligned to one another. Although text corpora offer an intriguing possibility for CLTR query translation, the lack of domain-specific texts or a suitably large range of texts means that general purpose query translation systems remain elusive.

At the Computing Research Laboratory at New Mexico State University, Ted Dunning and I have been prototyping and evaluating CLTR systems since 1992. Our work has been concerned with the theory and application of existing text retrieval technology, as well as novel methods for translating queries. The methods we have applied have been primarily based on large bilingual text corpora. We use the corpora to derive query translations directly by selecting terminology from the target language portion of the corpus. We have also used bilingual corpora to try to optimize translated queries by eliminating terms incrementally from the target query according to the similarity of the query and target query's retrieval characteristics.

Translation as a Linear Transformation

Our earliest efforts in CLTR considered the translation process to be a linear transformation of document feature vectors (Dunning and Davis, 1992). The feature vectors were composed of document term counts. Because the feature vectors are sparse, the transformation matrix was massively underdetermined. For a pair of documents E_i and S_j , and a transformation matrix T , an iterative update method can be formulated to solve the constrained minimization problem,

$$E_i = (T + \Delta T)S_j \quad \text{subject to the constraint:}$$

$$\sum_{ij} \Delta T_{ij}^2$$

The solution to this minimization problem is:

$$\Delta T = \frac{(E_i - TS_i)S_i^T}{S_i^T S_i}$$