Spanish and Chinese Document Retrieval in TREC-5

Alan Smeaton
School of Computer Applications
Dublin City University

Ross Wilkinson
Department of Computer Science
Royal Melbourne Institute of Technology

April 5, 1997

1 Multilingual Document Retrieval in TREC

The TREC-5 conference was the third year in which document retrieval in a language other than English was benchmarked. In TREC-3, 4 groups participated in an ad hoc retrieval task on a collection of 208 Mbytes of Mexican newspaper text in the Spanish language. In TREC-4 there were 10 groups who participated, once again in an ad hoc document retrieval task on the same Mexican newspaper texts but with new topics. In TREC-5 there was a change of document corpus and new topics for the Spanish ad hoc retrieval task and a corpus of documents and topics to support ad hoc retrieval in the Chinese language was introduced for the first time. There were 7 groups who submitted runs for the Spanish track and 10 who submitted results for Chinese.

The corpus of texts used in the TREC-5 Spanish language task was approximately the same size as the one used in TREC-3 and TREC-4 but differed in that there was a more consistent use of accented characters and it was European Spanish as opposed to Mexican Spanish. This slightly affected the morphological processing of word forms.

In the Chinese language each character represents at least a complete syllable, rather than a letter as in other languages. Many characters are also single syllable words. The total number of characters is therefore quite large and somewhat ill defined. A literate adult would typically recognise at least 5-6,000 characters. The various modern standards define between 10-12,000 characters, although if early and ancient literature is included the number rises to approximately 100,000. Chinese is agglutinating – there is no space between consecutive characters, except perhaps, at the end of a sentence. Thus to perform retrieval in Chinese, the basis has to be characters unless the text is pre-segmented into words.

2 Retrieval Task

The retrieval task for the Spanish and Chinese tracks are exactly the same as the standard ad-hoc task in TREC. A given database of texts and a fixed set of topics are supplied. The task is to return a ranked list of 1,000 documents for each of the topics. For each topic, at least one run using only the description part of the topic is encouraged. The topics were supplied in both English and either Spanish and Chinese. Either the English could be used so that cross lingual retrieval could be explored, or the language of the document collection could be used for monolingual experiments.

For Spanish, a 173,950-document collection from the Agence France Presse 1994 newswire was used. It was 308 Megabytes as raw text. 25 topics were constructed for the Spanish collection and there were on average 100 relevant docs per topic. From the 7 groups whose results were pooled, each topic had an average of 679 documents manually assessed for relevance.

For Chinese, a 164,811 document collection included documents from both the People's Daily and the Xinhua News Agency. There was no segmentation information supplied. It was 170 Megabytes as raw text. 28 topics were constructed for Chinese. There were on average 93 relevant docs per topic for Chinese.

3 Spanish Results

The 7 groups who took part in TREC-5 Spanish used a variety of approaches for indexing and retrieval. We summarise these approaches before discussing their comparative retrieval effectiveness.

Cornell University

The SMART system used in the mainline submissions from Cornell University was also used in the Spanish track. A simple stemmer which removed common word endings was used to normalise word occurrences. As per Cornell's mainline submission in TREC-5, documents were initially ranked and then re-ranked based on Rocchio relevance feedback assuming documents initial top-ranked were in fact relevant. In addition to indexing by normalised word occurrences, the the Cornell approach also used any pair of non-stopwords occurring sufficiently frequently in the corpus as a *statistical phrase*. What is especially notable about the Cornell submission, apart from retrieval effectiveness, is that the porting effort needed to enable SMART to handle a corpus in Spanish was very small, and this included the development of the word normalisation module.

University of Massachusetts

As with Cornell University, the submission from the University of Massachusetts was much the same as their previous work, in this case a follow-on from a TREC-4 Spanish track approach and using a variation of their mainline TREC-5 work. This was based on combining global analysis and local feedback to generate query expansions based on local context analysis, essentially expansing queries by terms from top-ranked documents if they occur near query term occurrences. This group used a sophisticated word normalisation process developed in their TREC-4 Spanish work.

Rank Xerox Research Center

The Xerox group used a Spanish part-of-speech tagger to identify and index by noun pairs as an alternative to the simple adjacent non-stopword word pairs or "statistical phrases" used by others such as Cornell. The Xerox group also used a linguistically motivated lemmatizer to reduce word occurrences to their root forms, as an alternative to the comparatively crude stemming process used by most IR systems.

The Xerox group were also the first to use the English version of the queries as supplied by NIST, in some of their runs. They used an inflectional morphology stemmer to stem the text of the English version of the topics and automatically translated each resulting word root form into its Spanish equivalent for retrieval using a dictionary lookup. This represents the first attempt in TREC to do cross-language retrieval.

University of California, Berkeley

The Berkeley group had used a morphological stemmer in their TREC-4 Spanish track and this was enhanced for their TREC-5 submission with the addition of a larger list of irregular and regular verb forms. In addition, acronyms were identified and excluded from the stemming process and an attempt was made to correct the inconsistent use of accented characters in documents and in topics by comparing each word occurrence against a massive, unstemmed wordlist. The retrieval algorithm used by the Berkeley group was the same as in their TREC-4 Spanish.

Dublin City University

The group from Dublin City University took part in the Spanish track in order to evaluate the performance of a new stemming algorithm for Spanish developed by Martin Porter. Porter's Spanish stemmer is essentially built along the same principles as his 1980 stemmer for English; a word ending is taken off if the remaining stem is of a suitable "length" as determined by vowel-consonant patterns. There is also a list of word prefixes which are not used in calculating a stem's "length", e.g. "CON-". Porter's stemmer also knows about 500 of the most commonly occurring irregular verbs in Spanish, much more than in English. The retrieval algorithm used by the DCU group was plain vanilla *tf*IDF* term weighting and document ranking, with no phrase recognition or document partitioning, as used in the mainline ad hoc submission from this group.

George Mason University

Like the group from Cornell, GMU used a two-pass retrieval algorithm incorporating estimated relevance feedback into the second pass. The original topics were used as part of a *tf*IDF* term weighting to generate a first document ranking and then terms in the top-10 ranked documents were ranked by their *n*IDF* values and the top 10 of those were added to the query for the second pass retrieval. There was no stemming in the GMU approach though a 500-word stopword list was used. The implementation of this was on GMU's relational database prototype.

New Mexico State University

The final group from NMSU concentrated on cross-language retrieval technologies by using the English version of the topic description and automatically translating into Spanish for running against the Spanish documents. The automatic translation processes included an attempt at text disambiguation among multiple senses in a bilingual dictionary, incorporating all translations from the dictionary, and a baseline run of Spanish queries against Spanish documents. The retrieval engine used in all cases was, like most groups, based on *tf*IDF* term weighting, and also used a Porter-like word stemmer once the dictionary-based translation had been completed.

Conclusions on Spanish Retrieval

Retrieval on Spanish texts threw up some interesting observations for the groups who took part. Even though the corpus was of European as opposed to Mexican Spanish, and the use of accented characters was more consistent, it was still variable across documents. Some groups took a simple approach to this and ignored all accents while for others there was a serious attempt to use and even to correct the accented character use. In Spanish the accents on characters generally does not affect the meaning of the words so we don't know what effects the special treatment vs. the ignoring approaches have.

For most groups participating in the Spanish track, the track was an add-on, an afterthought, with a low priority after their mainline TREC submissions and the effort required to alter established systems such as SMART and INQUERY to handle Spanish texts was minimal. Some groups like NMSU and Xerox, however, made it a priority. The amount of effort put in by respective sites is not reflected in the relative performances in terms of precision and recall as approaches which worked well for the English mainline task carried over reasonably well into the Spanish retrieval task also. For groups which had developed sophisticated and effective retrieval techniques for the mainline task their reward for taking part in the Spanish task was the knowledge that their techniques worked well in Spanish also.

Despite the small number of participating groups, it is difficult to make cross-system performance comparisons, even when retrieval uses the same documents and topic descriptors. An expected result is that the longer forms of the queries incorporating the narrative as well as the description, yields better averaged performance but that does not tell us anything new. Many of the groups developed new approaches to word normalisation and it would have been interesting to compare directly, the normalisation approaches taken by Cornell, Xerox, Berkeley and Dublin, for example. This would have been the interesting and the language-specific issue in the Spanish track. Unfortunately, and this is a symptom of TREC in general, each group which develops a new technique for any part of the retrieval process must then wrap that technique within the IR system used at their site and other IR parameters such as stopword lists, term weighting, document length normalisation, document scoring, and so on, are all different from site to site and this prevents direct cross-site comparisons. Thus the only scientifically credible comparisons we can make are among the runs submitted from a single site, where the other retrieval parameters are not variable. The alternative to this would be an unworkable version of TREC.

The two groups which tried cross-language retrieval, Xerox and NMSU, both found that their cross-linguistic results were not as good as using the original Spanish form of the queries. The NMSU group claim to have recaptured 70using their best English-Spanish, whereas the Xerox performance drop is more severe. Another observation from the Xerox work is that adding linguistically derived noun phrases/noun pairs improved retrieval effectiveness over not adding, but only slightly, so this raises the question of whether it is it worth the effort. This is a big question in IR research anyway.

4 Chinese Results

The 10 groups who took part in TREC-5 Chinese generally explored the use of words vs. n-grams and methods of manually modifying queries. Some work was also done on retrieval methods particularly appropriate to Chinese retrieval. We summarise these approaches before discussing their comparative retrieval effectiveness.

City University

The experiments at City University used the Okapi system for their Chinese retrieval experiments. The tried both a character based retrieval and a word retrieval. Words were discovered using a greedy algorithm using a 70,000 word dictionary. With both character and word approaches, the use of phrases were explored. A number of probabilistic relationships were investigated based on the relative probability of a phrase appearing given that both constituents have appeared.

Claritech Corporation

The Claritech used the Clarit system for both n-gram character based and word based approaches. Words were discovered using a dynamic programming algorithm that segmented into minimal numbers of words using a 100,000 word dictionary and heuristics. Their automatic experiments showed that word based and bi-gram approaches were roughly comparable and that both were better than single character retrieval. Claritech also explored manual retrieval by introducing new terms and introducing Boolean constraints on documents to be used for feedback. Both of these techniques provided improvements. It was interesting to note that feedback that usually works well in most other circumstances failed to provide gains for manual retrieval.

Cornell University

Cornell approached Chinese retrieval with no Chinese expertise but a very good retrieval system – the SMART system. They achieved very good results. They approached the task by using character based retrieval augmented with character bi-grams. Expansion techniques worked well.

George Mason University

Only single characters were used in the George Mason approach. Their base run again showed that reasonable performance can be obtained using characters only. They then applied term expansion by selecting the top 10 terms from the top 10 documents. This gave good improvement. However having a person ensure that only relevant documents were used in selecting terms for expansion gave no improvement.

Information Technology Institute

The Information Technology Institute ran both automatic and manual experiments. The automatic experiment was a weighted word based ranked approach. Queries were then modified manually using fuzzy Boolean constructs. This approach gave very substantial improvements.

Queens College, CUNY

The approach taken here was to apply a combination of dictionary and statistical techniques to detect 2, 3 and occasionally 4 character words. The aim was to obtain good indexing features rather than "correct" segmentation. The also tried using both characters and words which gave a small improvement.

Swiss Federal Institute of Technology

ETH used fully automatic techniques using both single characters and all possible bi-grams. Term expansion was then applied. There were no official runs submitted.

Royal Melbourne Institute of Technology

The RMIT approach was to use several automatic runs based upon characters and words found by dictionary methods. No run was successful.

University of California, Berkeley

The Berkeley group put a lot of effort into building a good dictionary of 140,000 words to use to automatically segment the text. They then applied their standard ranking methods. Queries were then modified with several hours of manual effort with considerable improvement.

University of Massachusetts

A hidden markov model was used to segment text in the University of Massachusetts approach. The resulting queries used characters, groups of characters, and words. Experiments were then conducted using the Local Context Analysis approach to term expansion. Performance was better when expansion terms had lower weights.

General Remarks

This was the first year of this track, so it was not possible for groups to optimize their techniques on past data. Nevertheless it does appear to be the case that the best n-gram approaches, including single character approaches were comparable to the best word based approaches. Most techniques for improving retrieval performance using term expansion worked well – as they had in the comparable English experiments. It was noticeable how much improvement occured when groups manually modified the queries. The best runs were obtained as a result of manual modification.