Chinese Document Retrieval at TREC-6

Ross Wilkinson
Mathematical and Information Sciences, CSIRO
Email: Ross.Wilkinson@cmis.csiro.au

January 9, 1998

1 Multilingual Document Retrieval in TREC

The TREC-6 conference was the fourth year in which document retrieval in a language other than English was carried out. 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. In TREC-6 there was two tracks in which languages other than English were explored. In the Chinese track, a second set of topics were evaluated against the existing corpus. In the cross-lingual track experiments were conducted where queries in one language were used against a document corpus in another language. This report concentrates solely on the Chinese track.

2 Chinese Language

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. Segmentation is difficult – not even humans will always agree on correct segmentation, and there has been much research in successful segmentation of Chinese [1].

3 Retrieval Task

The retrieval task for the Chinese track is 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 was encouraged. The topics were supplied in both English 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.

A 164,811 document collection including documents from both the People's Daily and the Xinhua News Agency was used. There was no segmentation information supplied. It was 170 Megabytes as raw text. There were 26 new topics constructed. There were on average 114 relevant docs per topic for Chinese.

4 Chinese Results

The 12 groups who took part in TREC-6 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 summarize 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 bi-gram character retrieval, and then applying automatic feedback. A comparison was made between long and short queries, as well as using manual intervention. Manual intervention of about 7 minutes per query produced a 5% gain. Feedback was routinely helpful.

Cornell University: Cornell again approached Chinese retrieval with no Chinese expertise but a very good retrieval system – the SMART system. They approached the task by using character based retrieval augmented with character bi-grams. They applied standard expansion techniques, as well as SuperConcepts. The expansion on single characters no longer gave a performance gain.

Information Technology Institute: The Information Technology Institute applied a novel matching algorithm for character based retrieval using positional information. They then combined the technique with expansion terms selected from 3-grams. This gave a gain. Further filtering of terms was not helpful.

Institute of Systems Science: The Institute of Systems Science carried out only automatic runs by combining both bi-gram approaches and word based approaches. These gave gains. They then investigated the nature of n-grams concluding that 2-grams are roughly equivalent to function words, 3-grams are roughly equivalent to names, and 4-grams are roughly equivalent to concepts.

MDS, RMIT: The RMIT approach was to combine several automatic runs based upon characters, bi-grams, and words found by dictionary methods. Two dictionary methods were used, based on dictionary and mutual information, with the dictionary giving superior results. The best results were obtained by combining bi-grams and words, and combining with expansion terms as well.

Queens College, CUNY: Last year, Queens had the most successful automatic approach based on identification and indexing on short-words, words of no more than 3 characters. This approach was again applied successfully using a 43K lexicon. However given the success of bi-grams, a combination approach was taken this year by combining the results of short-words and bi-grams. Results again were very good.

Swiss Federal Institute of Technology: ETH used fully automatic techniques to index using bi-grams. They concentrated on reducing the size of the vocabulary by using a stop list to partially segment. The achieved a large reduction in the corresponding dictionary. A manual approach was used using feedback. 40-50 minutes were spent on each topic, including feedback.

University of California, Berkeley: In TREC-5, 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 noticed that this dictionary still did not contain many important indexing terms, in particular names, so they developed further techniques for identifying these names and further augmented their dictionary with 10,000 entries. In their runs they used a word based indexing approach which was then either manually or automatically augmented with additional query terms.

University of Massachusetts, Amherst: The UMass approach to TREC-6 was exactly the same as with TREC-5. 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.

University of Montreal: The Montreal effort concentrated on improved word identification algorithms by using more sophisticated morphological analysis. The results of this approach was then compared to the bi-gram approach. The word based approach gave slightly better performance, but the authors believe that more gain is possible after more effective segmentation.

University of Waterloo: The University of Waterloo used individual character indexing augmented by phrases based on adjacent characters using their Multitext approach. Topics were constructed manually by interacting with the collection. Several queries were constructed per topic and results merged manually. 1.5–2 hours were used to formulate queries per topic. The results gave the best manual run.

General Remarks

This was the second year of retrieving Chinese documents. Retrieval effectiveness was again very high. Median performance was above 0.5. With the average of the best run for any group at about 0.7, and the best single group performance at 0.62, it is clearly difficult to distinguish between approaches. What is less clear is why results are so high – it could be the data, the queries or the nature of Chinese. Until a new collection with different queries are obtained this question remains open. What is clear is that given the very high level of performance it is very hard to obtain insights into differences that individual techniques might make.

Team	Better run
City	character better (5%)
Claritech	character only
Cornell	character only
ITI	character only
ISS	character better (18%)
MDS	word better(1%), combination best
Queens	word better (4%), combination best
ETH	character only
Berkeley	words only
UMass	character better (2%)
Montreal	words better (1%)
Waterloo	user selected

Table 1: Comparing character-based and word-based approaches

Despite this and following last year's results, it is still the case that bi-gram approaches are comparable with any other individual technique and have the advantage of not requiring the difficult task of segmentation in order to employ word-based approaches (see Table 1). It has been suggested by several participants that this is due to the greater semantic content of characters compared to any sub-word element in English and other European languages. It was again the case that most groups obtained some gain from expansion and that manual intervention helped.

References

[1] R. Sproat and C.L. Shih. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351, 1990.