TREC-9 Cross-Language Information Retrieval (English - Chinese) Overview

Fredric Gey and Aitao Chen
UC DATA and SIMS
University of California, Berkeley
e-mail: gey@ucdata.berkeley.edu,aitao@sims.berkeley.edu

Abstract

Sixteen groups participated in the TREC-9 cross-language information retrieval track which focussed on retrieving Chinese language documents in response to 25 English queries. A variety of CLIR approaches were tested and a rich set of experiments performed which measured the utility of various resources such as machine translation and parallel corpora, as well as pre- and post-translation query expansion using pseudo-relevance feedback.

1 Introduction

For TREC-9 the cross-language information retrieval task was to utilize English queries against Chinese documents. This aspect of multilingual information access at TREC-9 was the seventh year in which non-English document retrieval was tested and evaluated, and the fourth year for which cross-language information retrieval has been experimented with. In TREC-3, retrieval of 25 queries against a Mexican newspaper corpus was tested by four groups. Spanish language retrieval was evaluated in TREC-3, TREC-4 (another 25 queries for the same Mexican corpus), and TREC-5 (where an European Spanish corpus was used). In TREC-5 a Chinese language track was introduced using both newspaper (People's Daily) and newswire (XinHua) sources from People's Republic of China and 25 Chinese queries with an English translation supplied. The TREC-5 corpus was represented with the GB character set for the simplified Chinese language of PRC. Chinese monolingual experiments on this collection were done in TREC-5 and TREC-6 and sparked serious research into Chinese text segmentation methods using dictionary methods as well as statistical methods using measures such as mutual information. Comparisons have been made with simple overlapping bigram segmentation methods for monolingual Chinese retrieval. TREC conferences TREC-6, TREC-7 and TREC-8 has cross language tracks which focussed upon European languages (English, French, German, and later Italian). Following TREC-8 the venue for evaluating European language retrieval moved to Europe with the Cross-Language Evaluation Forum (CLEF) first held in Lisbon in September 2000 [9].

2 Task Description

As in past TREC cross-language information retrieval evaluations, the task for each group was to match topics in one language (English in this case) against documents in another language (Chinese) and return a ranked list of the top 1000 documents associated with each topic. Multiple runs were allowed for each group but one run using only the title and description field was required. Evaluation then proceeded by pooling ranks and manual examination of the pools by human judges who decide upon the relevance or irrelevance of each document in the pool. Once relevance judgments were established the usual measures of recall and precision could be computed upon the ranked list of each entry.

2.1 Topics

Twenty-five topics in English (numbers CH55-CH79) were created at NIST. Two typical topics are Topic 56 (human rights violations) and Topic 79 (livestock in China):

```
<top>
<num> Number: CH56
<title> human rights violations
<desc> Description:
What human rights violations have occurred in countries
outside of China according to the Chinese press.
<narr> Narrative:
Reports of human rights violations in China are not
relevant.
</top>
```

```
<top>
<num> Number: CH79
<title> livestock in China
<desc> Description:
What kinds of livestock are being raised in China?
<narr> Narrative:
A document that discusses livestock farming in China,
but is not specific about the kind of livestock is
not relevant.
</top>
```

These topics demonstrate two kinds of difficulty. For topic CH56, the limitation of relevant human rights violations to 'countries outside China' is one of discrimination between human rights news stories concerned within China and those whose focus is other than China. Topic CH79 illustrates the use of a general term (livestock) while requesting specificity (e.g. pigs) within the documents returned.

2.2 Documents

The corpus for TREC-9's CLIR evaluation consisted of 126,937 documents (188 megabytes in size) with newspaper sources from Hong Kong for the periods 1998-1999. In distinction from the earlier TREC

Chinese corpus, these sources were written in the richer traditional Chinese character set, encoded in the BIG5 encoding. In particular the source documents came from:

- Hong Kong Commercial Daily (Aug 11, 1998 - Jul 31, 1999)
- Hong Kong Daily News (Feb 1, 1999 - Jul 31, 1999)
- Ta kung pao (Oct 21, 1998 - Mar 4, 1999)

3 Participants and General Approach

Sixteen groups participated in the TREC-9 Chinese evaluation, listed here in alphabetical order:

BBN Technologies
Fudan University (PRC)
IBM T.J. Watson Research Center
Johns Hopkins University (Applied Physics Laboratory)
Korea Advanced Institute of Science and Technology
Microsoft Research, China
MNIS-TextWise Labs
National Taiwan University
Queens College, CUNY
RMIT University (Australia)
Telecordia Technologies, Inc.
The Chinese University of Hong Kong
Trans-EZ Inc.
University of California at Berkeley
University of Maryland
University of Massachusetts

The majority of approaches utilized word or phrase translation from English to Chinese by lookup in bilingual dictionaries or word lists. A number of groups used the Linguistic Data Consortium's English-Mandarin word list of approximately 120,000 pairs of words. Other dictionaries included the CETA (Chinese-English Translation Assistance) dictionary and the KingSoft online bilingual dictionary as well as local (proprietary) dictionaries.

Other approaches (in particular BBN) made use of statistical association models to create bilingual dictionaries from the alignment of parallel English-Chinese Corpora. Corpora used for development of resources or pre/post query expansion included:

- LDC parallel Hong Kong SAR Law
- LDC parallel Hong Kong SAR News

- Academica Sinica Balanced Corpus (ASBC)
- TREC-6 (People's Daily and Xinhua News Agency)
- Foreign Broadcast Information Service (FBIS) data
- Bilingual data harvested from the WWW
- Other local (proprietary) mono and bilingual corpora

In addition a few commercial machine translation software packages were used and coupled with other resources.

Extensive experimentation was done by some groups with query expansion, both before query translation from English to Chinese and after translation using blind feedback from the top ranked documents of an initial retrieval.

4 Experimental and methodological details by group

This section provides a summary of experiments run and methodological approaches tested by the eight groups with best-performing English-Chinese crosslingual runs, according to the official results (see Results section below). Experiments and approaches which seem unique are given more description in this section. Readers are directed to the individual papers for more detail.

4.1 BBN

BBN [12] extended the hidden Markov model (HMM) for monolingual retrieval to cross-language retrieval by incorporating into the model the word translation probabilities. Two manually created lexicons (i.e, the LDC wordlist and CETA) and two parallel corpora (i.e, the Hong Kong News and Hong Kong Law) were used to translate English query words into Chinese. The parallel texts were first aligned at the sentence level iteratively using WEAVER, a statistical machine translation toolkit developed at Carnegie Mellon University. Then WEAVER was applied to the sentence-aligned parallel texts to estimate word translation probabilities. When the translation resources were used individually, the Hong Kong News corpus yielded the best performance, probably because of similarity in topics covered in the test documents and the Hong Kong News corpus. However when all four translation resources were combined, the overall precision was substantially better. Unlike many participating groups of the cross-language track, BBN did not attempt phrasal translation and disambiguation of translation terms. A number of retrieval runs were performed to test the impact of query expansion for three levels of query length. The results showed over 10% improvement for overall precision for both pre-translation and post-translation query expansion when either one was applied alone. But when both pre- and post-translation expansion were applied, the post-translation expansion did not further improve the overall precision execept for the short queries consisting of only titles. In their official monolingual run, the Chinese text was segmented into words using the built-in segmentor in BBN's IdentiFinder.

The monolingual performance was slightly lower than the best cross-language retrieval performance. However later it was found that when the bigrams and unigrams were used in indexing, the monolingual performance increaded from .2888 to .3779.

4.2 Microsoft Research, China

Microsoft Research China group used a slightly modified version of SMART as their retrieval system [4]. A series of experiments were carried out to test the impact on retrieval performance using different indexing units and their combination. The results show that combining word-indexing and character-indexing works well for Chinese monolingual retrieval. They also used NLWin, a natural language processing system developed by Microsoft, to identify multi-word phrases and unknown words in the Chinese texts. They developed a co-occurrence based method to disambiguate translation terms, and a phrase detection and translation technique which improved the retrieval performance over the primitive dictionary-based translation. The phrases were identified using NLWin, and complex phrases were translated into Chinese based on a statistical model that maximizes the probability of phrase translation patterns and bigram probabilities estimated from a bigram language model trained on a large Chinese corpus. An interesting feature of the phrasal translation was that the Chinese words were put in the approriate order which may differ from the order of the source English words. About 125 MB of Chinese and English parallel texts were automatically mined from the Internet. A statistical translation model which is a variant of the IBM model was applied to the sentence-aligned parallel text to estimate word translation probabilities. When translation disambiguation and phrasal translations were augmented by statistical translation, the cross-language retrieval performance was as good as that obtained using IBM HomePage Dictionary 2000, a commercial Engish-Chinese machine translation system. The best cross-language retrieval run MSRCN2 combined bilingual lexicon, parallel texts mined from the Internet, and the machine translation system. Both pre- and post-translation query expansion were tried, however the pre-translation query expansion did not improve the overall precision.

4.3 Fudan University, China

The document scoring function used by Fudan University group was based on the maximum likelihood ratio formula developed by MIT. A number of rule-based named entity extractors were used to identify words that are not in the segmentation dictionary. In addition, the occurrence frequency and mutual information between characters of unidentified strings were used to identify unknown words. The translation resources used for query translation consist of three dictionaries: a general English-Chinese dictionary, a technical terminology dictionary, and an idiom dictionary. The translated queries were further expanded using a Chinese thesaurus of nearly 70,000 entries. The best cross-language run was the one without pseudo relevance feedback [11].

4.4 Chinese University of Hong Kong

The CUHK group translated the queries into Chinese by considering two adjacent words each time. Among all possible translation pairs found in a bilingual dictionary, the one with the hightest similarity was chosen as the final translation for the source adjacent words. They experimented with pre- and post-translation query expansion using Rocchio relevance feedback within the SMART retrieval system. The pre-translation query expansion improved the overall precision from .1862 to .2642, an increase of 42% over the baseline run. However the post-translation did not gain any further improvement [5].

4.5 Queens College, New York

The Queens College group used a commercial machine translation software named HuaJian and the LDC wordlist augmented with an additional 6,000 translation pairs extracted from the Hong Kong Laws corpus to translate the queries. For dictionary lookup, up to six translation terms were kept for each query term. An equal weight was assigned to each translation terms of the same source query term. The final result for a cross-language run was produced by combining the result using the MT-translated queries and the result using the dictionary-translated queries. The best cross-language run named HxD combined MT and augmented LDC wordlist translations with pre- and post-translation query expansion and Chinese collection enrichment [6].

4.6 University of Massachusetts

The University of Massachusetts group used their INQUERY system with Local Context Analysis (LCA) technique for query expansion. The Chinese queries were translated into English by looking up multi-word phrases or words in a bilingual dictionary built by merging two Chinese-English dictionaries with an English-Chinese dictionary. Multiple translations were retained and treated as synonyms. Both pre- and post-translation query expansion using LCA were tried. The post-translation query expansion gained very little improvement [1].

4.7 IBM Research

The IBM group used a character-based statistical model to translate the English queries into Chinese, and a word-based statistical model supplemented with the LDC dictionary to translate the Chinese documents into English, both models being trained on Hong Kong News and Hong Kong Law parallel corpora. Their official run was a merging of the results from three runs based statistical query translation, commercial MT-based query translation, and statistical document translation [3].

4.8 Korea Advanced Institute for Science and Technology (KAIST)

The KAIST group experimented with query translations using bilingual dictionaries and machine translation systems. The cross-language run using two bilingual dictionaries of 50,000 and 15,000 entries respectively outperformed the one using two machine translation systems. They observed that some of the proper names in the queries are spelled out in Chinese Pinyin (e.g., Daya Wan) and attempted to obtain the Chinese names based on a Chinese pinyin table and the occurrence statistics of the characters in the Chinese collection [7].

5 Relevance judgments and pool contributions

In order to create a pool of documents for each topic for human evaluation of relevance, each participating group was invited to nominate a single entry run from the monolingual and/or cross-lingual tasks to be included in the judgments. This produced 39 cross-lingual runs and 13 monolingual runs. All but one of the runs was automatic. The top 50 ranked documents were taken from each nominated run and added to the pool to be evaluated. As usual duplicated documents from runs with overlap are removed to produce a unique list of documents for each topic. The resulting document pools had mean size of 598 documents (39 percent of the maximum pool size) to be read by the judges. The relevant documents over the pools came from the component run-types as follows: thirteen percent were only found by monolingual runs, twenty-eight percent came from crosslingual runs only, while the remaining 59 percent were found in both monolingual and crosslingual runs. Figure 1 shows the number of unique documents contributed by site.

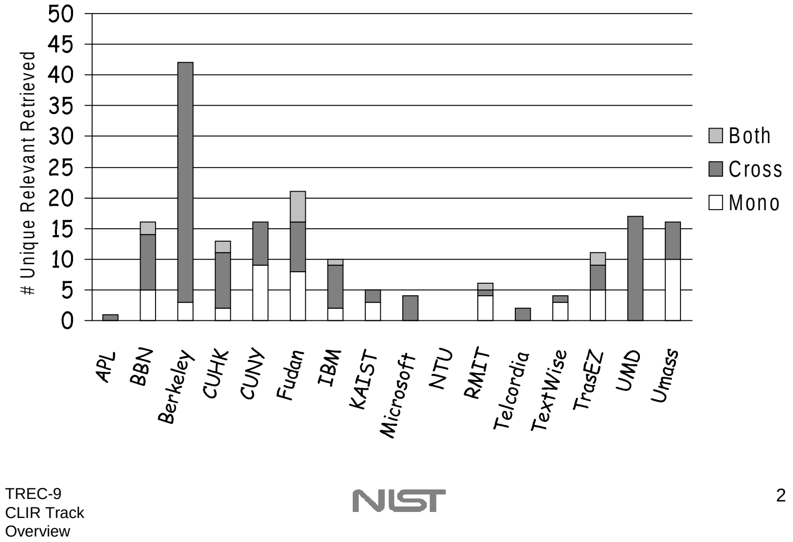

Figure 1: Unique Relevant Documents by Site

The only groups which contributed more than 20 relevant documents to the pool were Fudan University [11] and Berkeley [2].

6 Results

Figure 2 displays the recall-precision graph for the top eight best-performing crosslingual sites (the figure shows only the best run from each site). Although some groups seem to have clearly outperformed others, readers are cautioned that the evaluation only covered twenty-five queries, and it is unlikely that sufficient statistical signficance could be attained to confirm the rankings. The team from BBN outperformed all others with their hybrid combination of methods using a hidden markov ranking model, parallel corpora, and query expansion.

The reporting in the graph mixes run modes, since three runs used only title and description, while

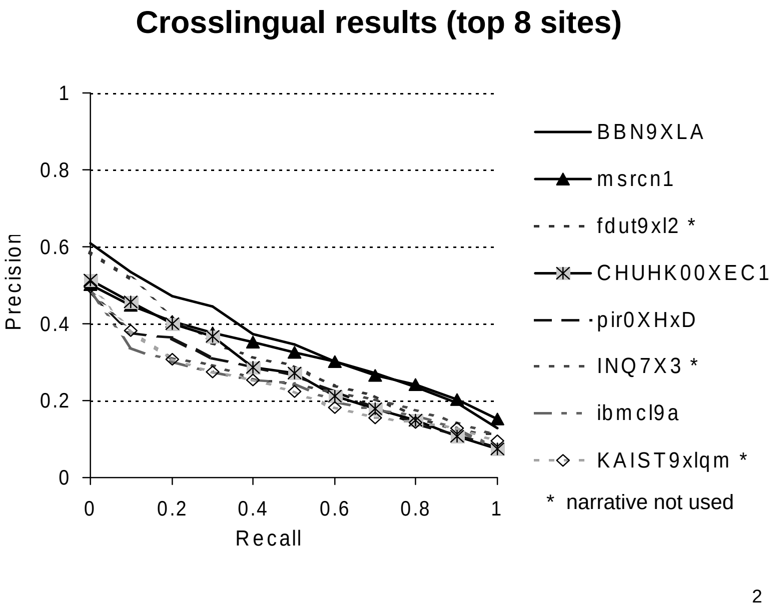

Figure 2: TREC-9 Cross-Language Retrieval Results

the others used the narrative portion of the topic as well.

The monolingual results for the top eight sights are displayed in figure 3. While they show very little difference between sites, two sites, Johns Hopkins – apl9xmon [8] and TextWise – TWmono [10] clearly performed significantly better on their monolingual than crosslingual runs. On the other hand, the best precision at 0.0 recall of 0.7079 from Berkeley (BRKCCA1) exceeded the best official CLIR run of 0.6078 by BBN. Finally, the overall average precision of the best official monolingual run (apl9xmon - 0.3085) trails the best crosslingual run (bbn9xla - 0.3485). BBN noted this discrepancy and attributed it in part to lack of query expansion in other bigram-based methods (BBN's official monolingual run used word-based indexing). BBN later implemented a bigram/unigram based monolingual algorithm with query expansion and achieved an overall monolingual precision of 0.3779 [12].

7 Summary and Outlook

The TREC-9 crosslingual information retreival task focussed this year on English-Chinese retrieval. Major experiments were undertaken using combinations of machine-readable dictionaries, machine translation software, and parallel corpora of news stories, legal documents, and bilingual sites mined from the

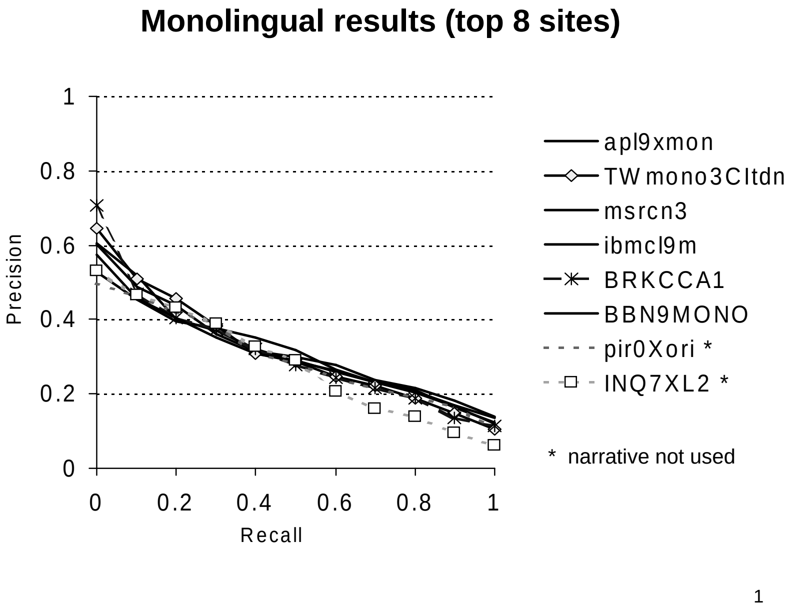

Figure 3: TREC-9 Monolingual Chinese Results

WWW. These were coupled with a variety of pre and post query expansion techniques. Many participating groups ran experiments which showed the contribution to overall precision of each component in the combination. The best performance was achieved by combining many of these techniques and by extensive use of the supportive resources.

In 2001 the TREC cross-language track will move from Chinese experiments to retrieving from a collection Arabic documents using either English or French queries. However, CLIR experiments with Chinese collections will continue with the NTCIR evaluation organized and hosted by the National Institute of Informatics of Japan (http://research.nii.ac.jp/ntcir/workshop/work-en.html).

We are grateful to Paul Over of NIST who supplied the basic information contained in our figures. His original Powerpoint presentation provided the basic outline from which this overview was written.

References

[1] J. Allan, M. Connell, W. B. Croft, F.-F. Feng, D. Fisher, , and X. Li. Inquery at trec-9. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[2] A. Chen, H. Jiang, and F. Gey. English-chinese cross-language ir using bilingual dictionaries. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[3] M. Franz, S. McCarly, and W.-J. Zhu. English-chinese information retrieval at ibm. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[4] J. Gao, J.-Y. Nie, J. Zhang, E. Xun, Y. Su, M. Zhou, and C. Huang. Trec-9 clir experiments at msrcn. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[5] H. Jin and K.-F. Wong. Trec-9 clir at cuhk, disabmiguation by similarity values between adjacent words. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[6] K. Kwok, L. Grunfeld, N. Dinstl, and M. Chan. Trec-9 cross-language, web and question answering track experiments using pircsi. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[7] K.-S. Lee, J.-H. Oh, J.-X. Huang, J.-H. Kim, and K.-S. Choi. Trec-9 experiments at kaist: Qa, clir and batch filtering. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[8] P. McNamee, J. Mayfield, and C. Piatko. The haircut system at trec-9. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[9] C. Peters, editor. Springer, Lecture Notes in Computer Science Series No. 2069, 2001.

[10] M. Ruiz, S. Rowe, M. Forrester, and P. Sheridan. Cindor trec-9 english chinese evaluation (mnis-textwise labs). In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[11] L. Wu, X. Huang, Y. Guo, B. Liu, and Y. Zhang. Fdu at trec-9: Clir, filtering and qa tasks. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.

[12] J. Xu and R. Weischedel. Trec-9 cross-lingual retrieval at bbn. In D. K. Harman and E. Voorhees, editors, *in this volume*, 2001.