TREC-9 Cross-lingual Retrieval at BBN

Jinxi Xu and Ralph Weischedel

BBN Technologies

70 Fawcett Street

Cambridge, MA 02021

1 Introduction

BBN participated only in the cross-language track at TREC-9. We extended the monolingual approach of Miller et al. (1999), which uses hidden Markov models (HMM), by incorporating translation probabilities from Chinese terms to English terms. We will describe our approach in detail in the next section.

This report will explore the following issues:

1. Whether our HMM-based retrieval model is a viable approach to cross-lingual IR. This is answered by its retrieval performance relative to monolingual retrieval performance.
2. The relative contribution of bilingual lexicons and parallel corpora.
3. The impact of query expansion on cross-lingual performance. We will use two types of query expansion: using English terms and Chinese terms.
4. The impact of query length on retrieval performance. We will use three versions of queries: short, which consist of only the title fields, medium, which consist of title and description fields and long, which consist of title, description and narrative fields of the TREC topics.
5. Whether indexing English words in Chinese documents helps cross-lingual IR. Even though the documents in the corpus are in Chinese, many of them also contain some English words. English words in the documents can directly match the query words.
6. Dialect issues. The Chinese language has many dialects. Cantonese, which is used by the TREC-9 corpus, is one example. Since we had lexical resources for Mandarin (standard Chinese) and for Cantonese, we could measure the impact of dialects on cross-lingual IR.

This report includes official results for our submitted runs and results for experimental runs that are designed to help us explore the issues above.

2 A Hidden Markov Retrieval System for Cross-lingual IR

In our approach, the IR system ranks documents by the probability that a Chinese document D is relevant given an English query Q, $P(D \textit{ is Rel} \mid Q)$. Using Bayes Rule, and the fact that $P(Q)$ is constant for a given query, and our initial assumption of a uniform a priori probability that a document is relevant, ranking documents according to $P(Q|D \textit{ is Rel})$ is the same as ranking them according to $P(D \textit{ is Rel}|Q)$. The approach therefore estimates the probability that a query Q is generated, given the document D is relevant. A two state Hidden Markov model approximates the query generation process given a document. One state is General English, denoted by GE, in which a term e is selected from the English vocabulary. General English words do not describe the content of the document. They are chosen simply because the user is creating a natural language query in English. The other state is the document state D in which a Chinese term c from the document is selected and translated to an English word e.

After a query is generated from a state, the HMM either stays at the current state or transits to the other state to generate the next query term. The process continues until all query terms are produced.

The following parameters specify the model:

1. General English word probabilities $P(e|GE)$, estimated by

   $$P(e|GE) = \text{frequency of } e \text{ in English corpus/size of English corpus.}$$

   Here e is an English word. English news articles in TREC disks 1-5 are used as an English corpus for this purpose.

2. Chinese word probabilities from the document D, $P(c|D)$, estimated by

   $$P(c|D) = \text{frequency of } c \text{ in } D\text{/size of } D$$

   Here c is a Chinese word.

3. Translation probabilities from Chinese words to English words, $P(e|c)$. We assume that translation probabilities are independent of the document. This is not true, but reduces the number of parameters. We used simple translation probabilities from a bilingual lexicon and more sophisticated estimates from parallel texts.

4. Transition probabilities from one state to the other. We assume

   $$P(GE\text{->}D) = P(D\text{->}D) = a \text{ and}$$

   $$P(D\text{->}GE) = P(GE\text{->}GE) = 1\text{-}a.$$

Further we assume a is independent of the document. Using TREC-5/6 queries (Chinese track) as training, we chose a=0.3.

Note we did not use the standard EM (Expectation-Maximization) procedure for parameter estimation, since using EM would require many training queries for each document.

In this model, we estimate the probability of a query given a document as

$$P(Q\ |D\ is\ rel) = \prod_{e\ in\ Q}(aP(e|GE) + (1-a)P(e|D))$$

and

$$P(e|D) = \sum_{all\ Chinese\ words\ c} P(c|D)P(e|c)$$

Our monolingual retrieval approach is the one proposed by Miller et al (1999). It ranks documents according to:

$$P(Q|D\ is\ rel) = \prod_{c\ in\ Q}(aP(c|GC) + (1-a)P(c|D))$$

where $P(c|GC)$ is general Chinese probability for word c, which was estimated from the TREC-9 Chinese corpus itself.

3 Lexical Resources

Two manually created bilingual lexicons were used in our experiments:

- one dealing with the Mandarin dialect from the Linguistic Data Consortium (LDC) and
- the CETA lexicon also dealing primarily with Mandarin.

In addition, two parallel corpora were used to generate bilingual lexicons. The parallel corpora are the Hong Kong SAR news (HKNews) and Hong Kong SAR laws (HKLaws), both from LDC. HKNews has around 18,000 pairs of documents in English and Chinese and has 6.3 million English words. HKLaws has 310,000 pairs of sentences in English and Chinese, with 6.6 million English words.

The following steps were taken to use each bilingual lexicon (whether manually generated or automatically derived from parallel copora):

1. Stem Chinese words via a simple algorithm to remove common Chinese suffixes and prefixes (such as "DE" and "BEI").
2. Use the Porter stemmer to stem the English words (Porter, 1980).
3. Split English phrases into words. If an English phrase is a translation for a Chinese word, each word in the phrase is taken as a separate translation for the Chinese word[1].
4. Estimate translation probabilities.

The resulting lexicons consist of a number of English-Chinese word pairs together with translation probabilities.

For those experiments where no parallel corpus was employed, we assumed a uniform distribution on a word's translations. If a Chinese word *c* has *n* translations e_1, e_2, …e_n, each of them will be assigned equal probability, i.e., $P(e_i|c)=1/n$.

For those experiments where a parallel corpus was employed, we used WEAVER to automatically extract additional translation pairs from the parallel corpora to improve the bilingual lexicons. WEAVER is a statistical machine translation toolkit developed by John Lafferty at Carnegie Mellon University. It has a component to automatically derive word translations based on sentence-aligned parallel text. The Chinese texts in the corpora were segmented by BBN's IdentiFinder™, an information extraction system with a built-in segementor. Since the HKNews corpus in its original form was only aligned at the document level, we developed a sentence alignment algorithm to align it at the sentence level. Our algorithm works by performing an initial alignment using a (potentially small) initial bilingual lexicon (the LDC lexicon). A bilingual lexicon was induced from the initial alignment using WEAVER. The induced lexicon supplements the initial lexicon in producing a better alignment, which in turn results in a better lexicon. The process eventually converges and outputs a list of term translations with translation probabilities.

The translations obtained by WEAVER are statistical in nature. In theory, any Chinese term can be translated to any English term with some probability; for the vast majority of word pairs, the probability approaches 0. For each Chinese term, we output up to 20 English terms and discard the rest, in order to keep the size of the lexicon manageable and to save retrieval time. Table 1 shows some statistics about the lexicons used in our experiments.

The lexicon used for our submitted runs is labeled "ALL" in Table 1. It is a combination of all lexical resources described before, LDC, CETA, HKNews and HKLaws. The sets of English-Chinese word pairs in the individual lexicons were unioned and the translation probabilities linearly combined, with coefficients 0.2, 0.4, 0.3 and 0.1 for LDC, CETA, HKNews and HKLaws respectively. The weights were chosen to reflect the value of each lexical source based on the training queries (TREC-5/6 Chinese). To utilize English words in the documents for cross-lingual retrieval, we include an English word as a

---

[1] This is incorrect, but greatly simplified implementation. The correct method would be to treat phrases in the lexicon and in the queries as single tokens. Research in monolingual IR demonstrated that phrase processing is prone to error and does not conclusively improve retrieval performance.

translation of itself with probability 1 and add such "translations" to our lexicons. (Such translations are not included in the statistics in Table 1).

Lexicon Name	# of English terms	# of Chinese terms	# of translation pairs
LDC	86,000	137,000	240,000
CETA	35,000	202,000	517,000
HKNews	21,000	75,000	1,266,000
HKLaws	14,000	38,000	543,000
ALL	108,000	371,000	2,470,000

Table 1: Lexicon statistics. All = combination of all four sources

4 Indexing

One problem in indexing Chinese is segmenting the text, since Chinese has no spaces between words. Instead of using a Chinese segmentor, we used a sub-string match algorithm to extract words from a string of Chinese characters. The algorithm examines any sub-string of length 2 or greater and treats it as a Chinese word if it is in our bilingual lexicons. In addition, any single character that is not part of any of the recognized Chinese words in the first step is also treated as a Chinese word. Note that this algorithm can extract a compound Chinese word as well as its components. For example, the Chinese word “LiZhiWuLi” (“particle physics”) as well as the Chinese words “LiZhi” (“particle”) and “WuLi” (“physics”) will be extracted. This seems desirable because it ensures the retrieval algorithm will match both the compound words as well as their components. The reason for using substring match instead of a more sophisticated segmentor is to improve the chance of mapping words in the Chinese document to an English term via the bilingual lexicons. A segmentor may mis-segment (e.g., a segmentation unit may cover the ending of one word and the beginning of another word). It may over-segment (e.g., producing a compound word while the lexicon only defines the components). It may also under-segment (e.g., producing individual words not defined by the lexicon). Substring matching may result in spurious matches, but we believe it is a less serious problem than being unable to map from Chinese to English due to segmentation errors. Of course, Chinese stop words are removed.

5 Query Processing and Query Expansion Issues

Our first step in query processing is to remove stop words from the queries. These include functional words such as “of” and “the” as well as red herrings in TREC topics such as “relevant” and “document”.

Our query expansion procedure works as follows:

1. For each query, retrieve 10 top ranked documents by an initial retrieval
2. Choose 50 expansion terms from the top ranked document. First, terms that only occur in one top ranked document were discarded. Then expansion terms were ranked by their average *tfidf* weight in the top ranked documents. The *tfidf* formula is the one reported in the UMass TREC6 report (Allan et al, 1998). The top 50 terms were added to the query. The expansion terms, as well as the original query terms were “weighted” by the formula

$$wt(t,Q) = wt_{old}(t,Q) + 4/10 \sum_{1<=i<=10} tfidf(t,d_i)$$

   Q is a query, $wt_{old}(t, Q)$ is the weight of term t in the original query; $tfidf(t, d)$ is the *tfidf* value of t in document d; and d_i’s are the retrieved documents. We interpret the "weight" of a query term in the context of our HMM retrieval approach to be the frequency with which the term is generated by the user. Therefore, the weight was used as an exponent in the retrieval function.

We submitted one monolingual run and three cross-lingual runs:

- BBN9MONO: a monolingual run with automatic query expansion. Final queries consist of the original Chinese queries and 50 expansion terms, using the query expansion procedure above.
- BBN9XLC: Cross-lingual without query expansion.
- BBN9XLB: Cross-lingual run with automatic Chinese query expansion. An initial cross-lingual retrieval was performed using the original English queries. Final queries consist of the original English queries and 50 Chinese expansion terms.
- BBN9XLA: Cross-lingual run with English query expansion and Chinese expansion. English terms were selected from top documents retrieved from an English corpus. Then the expanded English queries were run against the Chinese corpus to get 50 Chinese expansion terms. Final queries consist of the original English queries, 50 English expansion terms and 50 Chinese expansion terms.

The English corpus used for query expansion consists of news articles from TREC disks 1-5 (AP, WSJ, SJMN, FT, L. A. TIMES and FBIS) and 400,000 recent articles collected by FBIS in years 1999 and 2000.

Note queries in BBN9XLA and BBN9XLB contain both English terms and Chinese terms. To score a document against a query, two HMM scores were computed, one for the English query terms using the cross-lingual retrieval function, the other for the Chinese terms using the monolingual retrieval function. The two scores were multiplied to produce the final score for the document.

6 Official Retrieval Results

Table 2 shows the average precision for our submitted runs. What is striking is that all our cross-lingual runs have a higher score than our monolingual run. The results demonstrate that query expansion (BBN9XLA and BBN9XLB) improves retrieval performance, consistent with previous studies (Ballesteros and Croft, 1997).

BBN9MONO	BBN9XLA	BBN9XLB	BBN9XLC
0.2888	0.3401	0.3326	0.3099

Table 2: Retrieval results of submitted runs

7 Impact of query length and query expansion

Table 3 shows the impact of query expansion on cross-lingual retrieval performance. We show three versions of queries, short, medium and long. Short queries only use the words in the title field of the topics. Medium queries use title and description fields. Long queries use title, description and narrative. Query expansion improves performance for all query lengths. As expected, query expansion is more useful for short queries, and less useful for long queries. Three things are worth mentioning about the results. First, query expansion seems to neutralize the effect of query length. Without query expansion, the difference between short and long queries is 0.0669. After query expansion, it is reduced to 0.017. Second, English query expansion adds more than Chinese; apparently the benefit of a far larger corpus outweighs translation ambiguity. Third, while English expansion and Chinese expansion both improve retrieval performance, their combination does not improve performance further, except on the short queries. In fact, it is worse than either English expansion or Chinese Expansion alone for the medium queries. However, a query by query analysis shows that the surprising result is due to a statistical outlier in the retrieval results. The retrieval performance for topic 62 is 1.000 using English expansion and 0.3333 using both English and Chinese expansion. That query alone causes a difference of 0.0267 in average retrieval performance. Furthermore, topic 62 has only one relevant document; A small perturbation in the ranked output can cause a big change in retrieval performance. Under these circumstances, we cannot rule out the retrieval advantage of using both English and Chinese query expansion.

	No Expansion	Only English expansion terms	Only Chinese expansion terms	Both English & Chinese expansion terms
Short	0.2430	0.2991	0.2871	0.3231
Medium	0.2869	0.3282	0.3183	0.3038
Long	0.3099	0.3420	0.3326	0.3401

Table 3: Impact of query expansion on crosslingual retrieval performance

Table 4 shows the impact of query expansion on monolingual retrieval performance. As in cross-lingual retrieval, query expansion improves retrieval performance, but the amount of improvement is smaller.

	No Expansion	Expansion
Short	0.2299	0.2469
Medium	0.2476	0.2668
Long	0.2618	0.2888

Table 4: Impact of query expansion on monolingual performance

8 Impact of lexical sources on retrieval performance

The lexicon we used in our official runs is a combination of 4 lexical sources. Table 5 shows the contribution of each lexical source independently by reporting average precision without query expansion. The results show that the lexicon derived from the parallel corpus HKNews is the single most useful lexical resource; second is CETA, then LDC and last HKLaws. Each of these sources alone is significantly worse than the combined lexicon.

The experiment shows that different lexical sources can complement each other nicely. For our HMM-based approach, the results also show that the issue of lexicon completeness overrides that of translation ambiguity. On average, the combined lexicon has more than 1,000 Chinese translations per English query term. Even though this large figure is partly due to a few outliers, it does indicate there is a lot of translation ambiguity. The results indicate this does not have a big negative effect on retrieval performance.

	LDC only	CETA only	HKNews only	HKLaws only	ALL
Short	0.1491	0.1517	0.1875	0.1386	0.2430
Medium	0.1839	0.1944	0.2285	0.1395	0.2869
Long	0.1725	0.2126	0.2418	0.1441	0.3099

Table 5: Impact of lexical sources on average precision of retrieval. These results are without query expansion.

Another way to determine the value of a lexical source is to measure how much it contributes to the combined lexicon by removing the source from the combined lexicon and showing the impact on retrieval performance. The remaining sources were given equal weight. Table 6 shows that the most useful source is HKNews and the least useful HKLaws. In fact, removing HKLaws from the lexicon improves retrieval performance slightly. We think the reason is the domain mismatch between HKLaws and the TREC-9 Chinese corpus of news articles.

8.1 Comparing with TREC5 and TREC6 Queries

Although the TREC-5/6 Chinese corpus and TREC-9 corpus are both in Chinese, the former is in standard Chinese (Mandarin) and the latter in Cantonese. There are many differences in vocabulary between the two. As a result, using a bilingual lexicon for one dialect is sub-optimal for retrieval on a corpus in the other dialect. This effect can be seen when we compare retrieval performance using TREC-5/6 queries with TREC-9, as in Table 7. The LDC and CETA lexicons are better lexical resources than HKNews for TREC-5/6 but the opposite is true for TREC-9, probably because of the difference between the vocabularies in Mandarin and Cantonese. Had we had a bilingual lexicon for Cantonese, better retrieval results on TREC-9 may have been possible.

	ALL but LDC	All but CETA	ALL but HKNews	ALL but HKlaws
Short	0.2400	0.2298	0.1967	0.2462
Medium	0.2816	0.2678	0.2252	0.2950
Long	0.2924	0.2802	0.2506	0.3100

Table 6: Impact of removing a lexical source on average precision of retrieval. These results are without query expansion.

	LDC only	CETA only	HKNews only	HKLaws only
TREC-5 & 6 Medium	0.2897	0.3400	0.2496	0.1684
TREC-9 Medium	0.1839	0.1944	0.2285	0.1395

Table 7: Comparing TREC-5/6 and TREC-9

9 Utilizing English words in Chinese documents

Some Chinese documents in the TREC-9 corpus contain both English words and Chinese words. The English words are very useful for retrieval, for two reasons. First, they provide additional information about the content of the documents. Second, they can be utilized directly without translation, which invariably introduces errors. Such words were used in our submitted cross-lingual runs in the hope of improving retrieval. If we turned off this feature, the retrieval performance for BBN9XLC would be 0.3077 instead of 0.3099. Even though the difference is very small, we still think it is a desirable feature that can make a difference in a retrieval environment where such documents are common.

10 Monolingual Retrieval using Bigrams and Unigrams

Our official cross-lingual results are significantly better than our monolingual results. This anomalous result can be partly explained by the use of word-based indexing. As we discussed earlier, a word-based index is geared toward maximizing cross-lingual performance. For monolingual retrieval in Chinese, previous studies (Kwok, 1997) suggested that the best strategy may be to use bigrams. For comparison, we indexed the TREC-9 corpus using bigrams of Chinese characters and unigrams. Assuming a Chinese document is a sequence of Chinese characters, at each character position, we treat the bigram (current and the next characters) as a token. In addition, we also treat each character as a token. The resulting document is a bag of bigrams and unigrams. Stop words were discarded in the process. In a similar way, we processed the Chinese queries. Table 8 shows the monolingual results using bigrams and unigrams, together with our submitted results. Using bigrams and unigrams results in a huge improvement in monolingual performance. The results are also better than cross-lingual performance.

Bigrams. No Expansion	Bigrams. Query Expansion	BBN9MONO	BBN8XLA	BBN9XLB	BBN9XLC
0.3362	0.3779	0.2888	0.3401	0.3326	0.3099

Table 8: Using bigrams for monolingual retrieval

11 Conclusions

Our work was based on a previously reported HMM for retrieval (Miller et al., 1999); we extended that model from monolingual to cross-lingual retrieval. Several conclusions are suggested by the experiment:

1. As expected, query expansion improved short queries more than long queries. For this set of queries, it is interesting that the query expansion reduced the gap in (cross-lingual) performance between short and long queries from 25% relative without expansion to only 5% relative.

2. Quite surprisingly, with word-based indexing, all our cross-lingual runs were better than monolingual; the best cross-lingual run was 118% of monolingual. If we had used bigram indexing for monolingual performance, the best cross-lingual (word-based indexing) would have been 90% of the best monolingual (bigram based indexing).

3. Not surprisingly, the best bilingual resource was the one closest in dialect (Cantonese) and genre (news) to the document collection, even though it was automatically derived from a parallel corpus and highly ambiguous.

4. For our probabilistic model, coverage of the bilingual lexicon seems far more important than the degree of ambiguity in the lexicon.

5. Query expansion in English proved more valuable than query expansion in Chinese, in spite of the added ambiguity, perhaps because the English corpus for unsupervised relevance feedback was so much larger in English than for Chinese.

References

D. Miller, T. Leek and R. Schwartz, 1999. "A Hidden Markov Model Information Retrieval System." *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214-221, 1999.

J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, J. Xu, 1998. INQUERY Does Battle With TREC-6. In *TREC6 Proceedings*.

K. L. Kwok, 1997. Comparing Representations in Chinese Information Retrieval. *Proceedings of the 20th ACM SIGIR International Conference on Research and Development in Information Retrieval.*

L. Ballesteros and W.B. Croft, 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. *Proceedings of the 20th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 84--91.

M. Porter, 1980. An algorithm for suffix stripping. In *Program*, 14(3), 130-137.

Acknowledgements

We wish to thank John Lafferty, whose WEAVER software was used to derive bilingual lexicons from parallel corpora. We also would like to thank Rich Schwartz for his advice.

The work reported here was supported in part by the Defense Advanced Research Projects Agency under contract number N66001-00-C8008. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

Appendix

Table 9 summarizes monolingual results in this report.

Name	**Title**	**Desc**	**Narr**	**Query Exp**	**Words**	**Bigrams**	**Avg precision**
BBN9MONO	x	x	x	x	x		0.2888
	x				x		0.2299
	x			x	x		0.2469
	x	x			x		0.2476
	x	x		x	x		0.2668
	x	x	x		x		0.2618
	x	x	x			x	0.3362
	x	x	x	x		x	0.3779

Table 9: Monolingual results. Words = Word-based index. Bigrams = index using bigrams and unigrams of Chinese characters.

Table 10 summarizes cross-lingual results in this report.

Name	Title	Desc	Narr	ChQE	EnQE	LDC	CETA	HK News	HK Laws	Avg Precision
BBN9XLA	x	x	x	x	x	x	x	x	x	0.3401
BBN9XLB	x	x	x	x		x	x	x	x	0.3326
BBN9XLC	x	x	x			x	x	x	x	0.3099
	x					x	x	x	x	0.2430
	x				x	x	x	x	x	0.2991
	x			x		x	x	x	x	0.2871
	x			x	x	x	x	x	x	0.3231
	x	x				x	x	x	x	0.2869
	x	x			x	x	x	x	x	0.3282
	x	x		x		x	x	x	x	0.3183
	x	x		x	x	x	x	x	x	0.3038
	x	x	x		x	x	x	x	x	0.3420
	x	x	x	x		x	x	x	x	0.3326
	x					x				0.1491
	x						x			0.1517
	x							x		0.1875
	x								x	0.1386
	x					x	x	x	x	0.2430
	x	x				x				0.1839
	x	x					x			0.1944
	x	x						x		0.2285
	x	x							x	0.1395
	x	x				x	x	x	x	0.2869
	x	x	x			x				0.1725
	x	x	x				x			0.2126
	x	x	x					x		0.2418
	x	x	x						x	0.1441
	x						x	x	x	0.2400
	x					x		x	x	0.2298
	x					x	x		x	0.1967
	x					x	x	x		0.2462
	x	x					x	x	x	0.2816
	x	x				x		x	x	0.2678
	x	x				x	x		x	0.2252
	x	x				x	x	x		0.2950
	x	x	x				x	x	x	0.2924
	x	x	x			x		x	x	0.2802
	x	x	x			x	x		x	0.2506
	x	x	x			x	x	x		0.3100

Table 10: Crosslingual results. Title=the title field, Desc=the description field, Narr=the narrative field, ChQE=Chinese expansion terms, EnQE=English expansion terms, LDC= the LDC lexicon, CETA= the CETA lexicon, HKNews=lexicon extracted from HKNews, HKLaws=lexicon extracted from HKLaws. A "x" indicates a topic filed, a lexical resource, or a query expansion type is used.