TREC 2002 Cross-lingual Retrieval at BBN

Alexander Fraser[1], Jinxi Xu and Ralph Weischedel
BBN Technologies
50 Moulton Street
Cambridge, MA 02138

1. Introduction

We used basically the same retrieval system we used in TREC 2001. Our experiments featured a different method for estimating general English probabilities, two additional Arabic stemmers, a more complex model for lexicon extraction from parallel texts and a slightly different method for query expansion. To our disappointment, these changes did not improve retrieval performance.

1.1 Retrieval System

Our retrieval system was documented in (Xu et al, 2001). It ranks documents based on the probability that a query is generated from a document:

$$p(Q \mid D) = \prod_{t_q \ in\ Q} (\alpha P(t_q \mid GE) + (1-\alpha) \sum_{t_d \ in\ D} p(t_d \mid D)\, p(t_q \mid t_d))$$

$$p(t_d \mid D) = \frac{frequency\ of\ t_d\ in\ D}{size\ of\ D}$$

Where Q is a query, D is a document, t_q's are query terms, t_d's terms in the document. The mixture weight α is fixed to 0.3.

Two sets of parameters are important in the retrieval model. One is the translation probabilities $P(t_q|t_d)$. In TREC 2001, we used model 1 of Brown's statistical MT work (Brown et al, 1993) for estimating term translation probabilities from a parallel corpus due to efficiency considerations. With more computer power at disposal, for TREC 2002 we used the more complex but potentially more accurate model 4 for the same purpose. Differences between the two models were discussed by Brown et al, 1993.

The other is the general English probabilities $P(t_q|GE)$, which model the importance of the query terms for retrieval. In TREC 2001, we used a large English corpus (news stories in TREC English vols 1-5) for estimating $P(t_q|GE)$, by dividing the frequency of the term by the size of the English corpus, based on the assumption that the English and the Arabic corpora are sufficiently close in content and genre. This assumption is clearly not true. The English corpus consists of stories published in the late 80's and early 90's while the Arabic AFP corpus consists of articles published in the late 90's and 2000. Second, the two corpora focus on different geographic regions (AFP on Middle East

[1] Alexander Fraser is currently with Information Sciences Institute, University of South California

while TREC English on U.S.). Since finding a closely matched English corpus for AFP is hard, for TREC 2002 we computed GE probabilities based on the statistics of the Arabic translations of the English terms, using a technique proposed by Hiemstra et al, 1999:

$$p(e \mid GE) = \sum_{\text{Arabic words } a} p(e \mid a) p(a \mid GA)$$

where $p(a|GA)$ was computed based on the frequency counts of the Arabic terms in the AFP corpus.

1.2 Lexical Resources

We used two lexical resources for term translation, a parallel corpus and a manual lexicon. The parallel corpus was obtained from the United Nations (UN). The United Nations web site (http://www.un.org) publishes all UN official documents under a document repository, which is accessible by paying a monthly fee. A special purpose crawler was used to extract documents that have versions in English and Arabic. After a series of clean-ups, we obtained 38,000 document pairs with over 50 million English words. For sentence alignment, a simple BBN alignment algorithm was used. Translation probabilities were obtained by applying a statistical machine translation toolkit, GIZA++ (Och and Ney, 2000) on the UN corpus. GIZA++ is based on the statistical translation work pioneered by (Brown et al, 1993). We experimented with both model 1 and model 4 for lexicon extraction. The manual lexicon was obtained from Tim Buckwater (Buckwalter, 2001). It contains about 86,000 word pairs.

1.3 Arabic Stemmming

In TREC 2001 CLIR, we used the Buckwalter stemmer (Buckwalter 2001) for stemming Arabic words. It is table-driven, employing a number of tables that define all valid prefixes, stems, suffixes, and their valid combinations. Given an Arabic word *w*, the stemmer tries every segmentation of *w* into three sub-strings, $w=x+y+z$. If *x* is a valid prefix, *y* a valid stem and *z* a valid suffix, and if the combination is valid, then *y* is considered a stem. We modified the stemmer so that it only stems a word if the word has exactly one possible stem. Otherwise, the original word is returned. The performance of the Buckwalter stemmer depends on the coverage of the stem table: Words whose stems are not in the stem table cannot be stemmed by the stemmer.

In TREC2002, we experimented with two new Arabic stemmers as well as Buckwalter. One is UMass Light 8 (Larkey et al, 2002). The other is Al-Stem (Darwish, 2002), the standard stemmer for TREC 2002 CLIR. Both are rule-based and as such are not affected by lexicon coverage. Recent studies (Larkey et al, 2002; Darwish and Oard, 2002) demonstrated that rule-based stemmers are suitable for Arabic retrieval.

1.4 Query Expansion

In our TREC 2001 experiments, English and Arabic query expansions were performed sequentially: We performed English expansion first and then used the expanded English queries to retrieve the top documents for Arabic expansion. A potential problem with sequential expansion is that it can propagate errors made in the English expansion to the

Arabic expansion. In TREC 2002, we experimented with parallel expansion: We performed English and Arabic expansions independently, using the original unexpanded queries in the initial retrieval for expansion of both languages.

For English query expansion, we used a corpus of 1.2 million articles from sources AP, Reuters and FBIS. For Arabic query expansion, we used the AFP corpus and optionally additional articles from two newspaper sources Al-Hayat and An-Nahar. The expansion parameters are identical for both languages (English and Arabic): 50 terms were selected from 10 top retrieved documents based on their total TF.IDF in the top documents. The expansion terms and the original query terms were weighted as follows:

$$weight(t) \; = \; old\_weight(t)+0.4*\sum TFIDF(t, D_i)$$

where D_i's are the top retrieved documents.

1.5 Spelling Normalization

We used the same procedure we used last year to normalize spelling variations in Arabic words. Two kinds of spelling variations were considered. The first is the confusing of the letter YEH (ي) and the letter ALEF MAKSURA (ى) at the end of a word. Since variations of this kind usually result in an "invalid" word that is un-stemmable by the Buckwalter stemmer, our solution is to detect such "errors" using the stemmer and restore the correct word ending. The second is to write diacritical ALEFs (e.g. أ , إ and آ) as the plain ALEF (ا). In our experiments, we replaced all occurrences of the diacritical ALEFs by the plain ALEF.

2. Results of Submitted Runs

We submitted four runs—all are cross-lingual runs. The runs differ in the following aspects:

- The model used for lexicon extraction from the parallel corpus, *model 1 vs model 4*
- The lexical resources used for term translation
- The Arabic stemmer(s) used
- The Arabic corpus used for query expansion
- The query expansion method, *sequential vs parallel*
- The method the GE probabilities was calculated, *old vs new*. The old method computed the GE probabilities from the TREC English corpus while the new method computed them from the Arabic AFP corpus.

Table 1 shows the features of our submitted runs. BBN11XLS is our standard resource run. BBN11XLC essentially repeated our TREC 2001 work on the TREC 2002 query set. To our disappointments, the changes we made to last year's work did not produce better retrieval results, as shown by Table 2. In fact, the collective effect of the changes is a noticeable degradation in the retrieval performance (BBN11XLA and BBN11XLB vs

BBN11XLC). We are currently analyzing the impacts of the individual changes on retrieval.

	Model for lexicon extraction	Lexical resources	Arabic stemmer	Arabic Expansion Corpus	Query expansion	GE probabilities
BBN11XLA	Model 4	Parallel corpus and manual lexicon	Buckwalter and UMass Light 8	AFP, Al-Hayat, An-Nahar	Parallel	New
BBN11XLB	Model 4	Parallel corpus and manual lexicon	UMass Light 8	AFP, Al-Hayat, An-Nahar	Parallel	New
BBN11XLC	Model 1	Parallel corpus and manual lexicon	Buckwalter	AFP, Al-Hayat, An-Nahar	Sequential	Old
BBN11XLS	Model 1	Parallel corpus	Al-Stem	AFP	Parallel	New

Table 1: Description of sumbitted runs for TREC 2002 CLIR. BBN11XLA used two stemmers: Buckwalter for term translation and UMass Light 8 for stemming the Arabic expansion terms.

	BBN11XLA	BBN11XLB	BBN11XLC	BBN11XLS
Average Precision	0.3444	0.3514	0.3756	0.3473

Table 2: Retrieval results of submitted runs

Acknowledgements

This research is sponsored by the Defense Advanced Research Projects Agency and managed by SPAWAR Systems Center under contract N66001-00-C-8008. 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, the SPAWAR Systems Center, or the United States Government.

References

P. Brown, S. Della Pietra, V. Della Pietra, J. Lafferty and R. Mercer, 1993. “The Mathematics of Statistical Machine Translation: Parameter Estimation”. In *Computation Linguistics*, 19(2), 1993.

T. Buckwalter, 2001. Personal Communications.

K. Darwish. http://www.glue.umd.edu/~kareem/research/.

K. Darwish and D. Oard, 2002. “Term Selection for Searching Printed Arabic.” In ACM SIGIR 2002.

Hiemstra, D. and de Jong, F. 1999. “Disambiguation strategies for cross-language information retrieval.” In Proceedings of the third European Conference on Research and Advanced Technology for Digital Libraries, pages 274-293, 1999.

L. Larkey, L. Ballesteros and M. Connell, 2002. “Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis.” In ACM SIGIR 2002.

F. Och and H. Ney, 2000. “Improved Statistical Alignment Models.” In proceedings of *ACL 2000*.

J. Xu, R. Weischedel and C. Nguyen, 2001. “Evaluating a Probabilistic Model for Crosslingual Retrieval.” In proceedings of ACM SIGIR 2001.