Arabic Information Retrieval at IBM

Martin Franz, J. Scott McCarley
IBM T.J. Watson Research Center
P.O. Box 218, Yorktown Heights, NY 10598
<franzm,jsmc>@watson.ibm.com

February 6, 2003

1 Introduction

IBM built two systems for crosslanguage experiments with English queries and Arabic documents. One system approached translation and retrieval as entirely separate tasks: we used a machine translation system to translate the Arabic documents into English, and then did the retrieval with a standard English IR system; the other system incorporated the parameters of a machine translation model directly into an IR scoring formula. A further experiment combined both models.

For processing Arabic text, we had access to an innovative Arabic morphological analyzer, whose details will be described elsewhere. We incorporated well-known text normalizations [1] into the Arabic text processing. Our monolingual baseline system was similar to the system we have used in previous ad hoc tracks [3], and consisted of an Okapi [4] first pass followed by LCA-style [5] query expansion, applied to the normalized Arabic stems.

Translation model parameters were estimated from the U.N. parallel corpus. The English half was morphologically analyzed (as were the English queries in our submissions); the Arabic half was morphologically analyzed and text normalizations were applied. We built separate translation models relating normalized Arabic morphs to English morphs and relating Arabic words to English morphs.

2 Convolutional Model

Following the approach described in [2], we model $p(q|d)$, the probability of generating an English query q given an Arabic document d as a smoothed *convolution* of an English to Arabic translation probability with a probability of sampling Arabic text from the document. More specifically, we write

$$p(q|d) = \prod_i \Bigg(\alpha_1 p_0(q_i) + \alpha_2 \sum_j t_w(q_i|w_j) p(w_j|d) + \alpha_3 \sum_j t_s(q_i|s_j) p(s_j|d) \Bigg) .$$

Here the q_i are the morphological stems of the English query words, w_j are the (inflected) Arabic words, and s_j are the morphological stems of the Arabic words. We estimate the sampling probabilities $p(w_j|d)$ and $p(s_j|d)$ as the appropriate token count divided by the document length. We estimate the translation probabilities $t_w(q|w)$ of English stems given Arabic words and $t_w(q|s)$ of English stems given Arabic stems using the methods of [6], with the U.N. parallel corpus as training data. These estimates are smoothed with $p_0(q_i)$, the background probability of the English word which we estimate from the English half of the U.N. parallel corpus. For each query, the top documents were selected according to $p(q|d)$, and then the top Arabic terms were selected by tf-idf (similar to [1].) The expanded query was then rescored with the Okapi formula. Although this system uses the technology of statistical machine translation, it does not result in a translation of the corpus. In particular it only predicts the "bag of words" of which an English translation of a given document would be composed. It does not try to predict the order of the words - essential for a human-readable result.

3 Document Translation

An alternative approach is to use machine translation to translate the documents into English, and then use an English monolingual retrieval system similar to our previous TREC adhoc submissions [3, 4, 5] to retrieve the documents. The Arabic-to-English statistical machine translation system heavily draws upon Arabic morphological processing modules including word segmentation, part-of-speech tagging, and a novel technique of identifying optimal word units in the source and target languages inducing a higher quality word-to-word alignments. The morphologically processed corpus is used for IBM Model 1 [6] training and decoding. Further details will be published elsewhere.

Although both of these statistical models are trained on the same training corpus, they differ in several important aspects:

(1) the convolutional model "translates" on a document-by-document basis - words arbitrarily far apart in the document influence each other's translation; the machine translation system translates on a sentence-by-sentence basis - no information propagates across sentence boundaries.

(2) the convolutional model is based on a directly trained $p(english|arabic)$; the translation model is a source-channel model and uses $p(arabic|english)p(english)$

(3) the convolutional model sums over all possible translations of each Arabic word; the translation model makes a hard decision about each word's translation(s).

system	method	AveP	P20
ibmy02a	convolution	0.3509	0.4170
ibmy02b	doc. trans.	0.2705	0.3760
ibmy02c	merged a,b	0.3563	0.4290
ibmy02d	monolingual	0.3030	0.3820

Table 1: Results - mean average precision and precision at rank 20 of official submissions

4 Results

We submitted four experiments (three cross-lingual and one monolingual) for the evaluation. The results are shown in Fig. (1) The convolutional model (ibmy02a) had noticeably better performance than the document translation (ibmy02b) model. It is not clear whether this difference is due to summing over all possible translations or to other differences in the model. We also submitted a run that combined both methods (ibmy02c) [7], for a slight improvement in performance. In Fig. (1) we show a scatter plot of the query-by-query scores of the two methods.

5 Acknowlegments

This work is supported by DARPA under SPAWAR contract number N66001-99-2-8916. The authors would like to thank Young-Suk

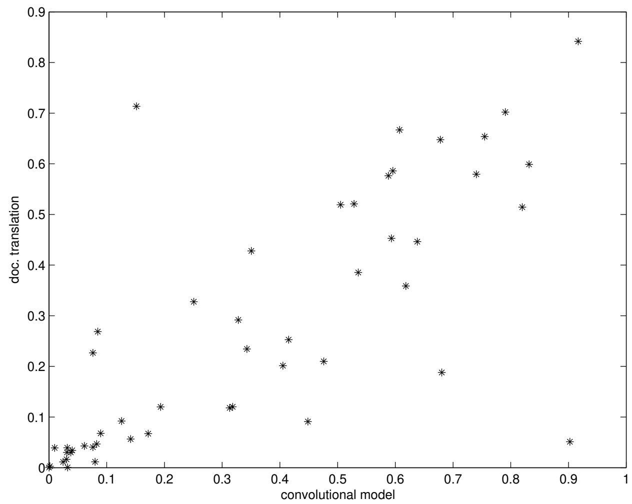

Figure 1: Scatterplot of the average precision of each query in the two IR systems.

Lee for the Arabic morphological analysis and Arabic to English translation system.

References

[1] J.Xu, A.Fraser, and R.Weischedel "TREC 2001 Cross-lingual Retrieval at BBN" in *Proceedings of the Tenth Text REtrieval Conference (TREC-10)* ed. by E.M. Voorhees and D.K.Harman, 2001.

[2] J.Xu,, R. Weischedel, and C.Nguyen, "Evaluating a Probabilistic Model for Cross-lingual Information Retrieval", In *SIGIR 2001*, pp. 105-110.

[3] M. Franz, J.S.McCarley, and R.T. Ward, "Ad hoc, Cross-language and Spoken Document Information Retrieval at IBM", in *Proceedings of the Eight Text REtrieval Conference (TREC-8)* ed. by E.M. Voorhees and D.K.Harman, p.391, 2000.

[4] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, M. Gatford, "Okapi at TREC-3" in *Proceedings of the Third Text REtrieval Conference (TREC-3)* ed. by D.K. Harman. NIST Special Publication 500-225, 1995.

[5] J. Xu and W. B. Croft 1996 Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Zurich, Switzerland, pp. 4-11.

[6] P. F. Brown et al. "The mathematics of statistical machine translation: Parameter estimation", *Computational Lingustics*, 19 (2), 263-311, June 1993.

[7] J.S. McCarley, "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?", in *37th Annual Meeting of the Association for Compuational Linguistics* College Park, MD, 1999.