TREC 2001 Cross-lingual Retrieval at BBN

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

1 INTRODUCTION

BBN only participated in the cross-lingual track in TREC 2001. Arabic, the language of the TREC 2001 corpus, presents a number of challenges to both monolingual and cross-lingual IR. First, many inflected Arabic words can correspond to multiple uninflected words, requiring context to disambiguate them. Second, orthographic variations are prevalent; certain glyphs are sometimes written as different, but similar looking glyphs. Third, broken plurals, analogous to irregular nouns in English, are very common. Such nouns cannot be easily reduced to their singular forms using a rule-based approach. Fourth, Arabic words are highly ambiguous due to the tri-literal root system and the omission of short vowels in written Arabic. The focus of this report is to explore the impact of these issues on Arabic monolingual and cross-lingual retrieval.

2 ISSUES IN ARABIC RETRIEVAL

2.1 Stemming

We used a modified version of Buckwalter's 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 re-implemented the stemmer to make it faster and compatible with UTF8 encoding. Also, we modified it so that if no valid combination of prefix-stem-suffix is found, the word itself is returned as the stem.

Ambiguities arise when a word has several stems. We used two techniques to deal with this problem. With the *sure-stem* technique, we only stem a word if it has exactly one stem. Otherwise, the word is left alone. With the *all-stems* technique, we probabilistically map a word to all possible stems. Since our retrieval system is based on a probabilistic generative model, such ambiguities can be easily accommodated. In the absence of training data, we assume that all possible stems are equally probable. That is, if a word has *n* possible stems, each stem gets *1/n* probability. The advantage of sure-stem is that it does not introduce additional ambiguity, while the advantage of all-stems is that it always finds a stem for a word when one exists.

---

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

2.2 Orthographic Variation

Arabic orthography is highly variable. For instance, changing the letter YEH (ي) to the letter ALEF MAKSURA (ى) at the end of a word is very common. (Not surprisingly, the shapes of the two letters are very similar.) 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.

A much trickier type of orthographic variation is when certain diacritical ALEFs (e.g. أ, إ and آ) are written as the plain ALEF (ا). Often, both the intended word and what is actually written are valid words. This is much like confusing "résumé" with "resume" in English. We explored two techniques to address the problem. With the *normalization* technique, we replace all occurrences of the diacritical ALEFs by the plain ALEF. With the *mapping* technique, we map a word with the plain ALEF to a set of words that can potentially be written as that word by changing diacritical ALEFs to the plain ALEF. In this absence of training data, we will assume that all the words in the set are equally probable. Both techniques have pros and cons. The normalization technique is simple, but it increases ambiguity. The mapping technique, on the other hand, does not introduce additional ambiguity, but it is more complex. Another problem is that the uniform probability assignment may deviate from the true probability distributions.

2.3 Broken Plurals

Broken plurals, analogous to irregular nouns in English (e.g. "woman/women"), are very common in Arabic. It is hard if not impossible to write a rule-based algorithm to reduce them to singulars. As such, broken plurals are not dealt with by the Buckwalter stemmer.

The problem is primarily a concern for monolingual retrieval. For CLIR, it is not a major problem because plurals and singulars can be translated separately. For monolingual IR, we use a statistical thesaurus, derived from the UN parallel corpus, to address the problem of broken plurals. The basic idea is that the singular and the plural forms of the same Arabic word should have the same stemmed translations in English. The problem can be formalized as the problem of estimating the probability that a user uses one Arabic word *b* to describe another Arabic word *a*. That is achieved by translating *a* to an English word *x* and then translating *x* to *b*. Translation probabilities from *a* to *x* and *x* to *b* are estimated by applying a statistical machine translation tool-kit, GIZA++ (to be described later), on the UN parallel corpus. It is easy to verify that

$$P_{thesaurus}(b \mid a) = \sum_{English\ words\ x} p(x \mid a)\, p(b \mid x)$$

A mixture model was used to emphasize the original words in the translation:

$$p(b \mid a) = 0.4\, p_{diag}(b \mid a) + 0.6\, p_{thesaurus}(b \mid a)$$

where $p_{diag}(b|a)=1$ if $a=b$ and 0 otherwise. The mixture weights were chosen based on experiments using the TREC-8 English monolingual test queries.

2.4 Tri-literal root system and omission of vowels

Most Arabic words can be derived from a small number (e.g. a few thousands) of roots. Most roots consist of only three consonants. Making things worse, short vowels are normally omitted in written Arabic. As a result, Arabic words tend to have a high level of ambiguity. If not addressed, this problem will hurt cross-lingual retrieval, because an Arabic word would have many translations.

Instead of explicit disambiguation, which weeds translations out based on context, we use a probabilistic solution that differentiates likely and unlikely translations. Although an Arabic word may have many translations, certain translations are more likely than others; hence, probability estimates limit the impact of ambiguity. In our CLIR experiments, we estimate translation probabilities from a large parallel corpus (the UN corpus) in addition to a manual bilingual lexicon.

3 BILINGUAL RESOURCES

We used a manual lexicon and a parallel corpus for estimating term translation probabilities. The manual lexicon consists of word pairs from three sources:

- A bilingual term list from Buckwalter (Buckwalter, 2001), with 86,000 word pairs.
- 20,000 word pairs, derived by applying the Sakhr machine translation system (http://www.sakhr.com/) on a list of frequent English words
- 10,000 word pairs gleaned from NMSU's named entity lexicon (http://crl.nmsu.edu/~ahmed/downloads.html).

Uniform translation probabilities are assumed for the English translations in the lexicon. That is, if an Arabic word has *n* English translations, each translation gets probability *1/n*.

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). Model 1 in Brown's work was used in this work for its efficiency.

The translation probabilities for the two sources were linearly combined to produce a single probability estimate for each word pair:

$$p(e \mid a) = 0.8\, p_{un}(e \mid a) + 0.2\, p_{lexicon}(e \mid a)$$

where *e* is an English word, *a* is an Arabic word, p_{un} and $p_{lexicon}$ are probabilities from the UN corpus and the manual lexicon respectively. We gave a higher weight to the UN corpus because it appears to be of higher quality.

4 OUR RETRIEVAL SYSTEM

Our retrieval system was documented in (Xu and Weischedel 2000; Xu et al, 2001). Our system ranks documents based on the probability that a query is generated from a document:

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

Where Q is a query, D is a document, t_q's are query terms, t_d's terms in the document. GL is a background corpus of the query language. The mixture weight α is fixed to 0.3. $p(t_q|t_d)$ is the translation probability from t_d to t_q. We estimate $p(t_q|GL)$ and $p(t_d|D)$ as:

$$p(t_q \mid GL) = \frac{\textit{frequency of } t_q \textit{ in } GL}{\textit{size of } GL}$$

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

In our cross-lingual experiments, the general English corpus (i.e. *GL* in the formulas) consists of newspaper articles in the TREC English disks 1-5 and more recent articles from FBIS. Translation probabilities were estimated as described in the previous section.

Because monolingual retrieval is a special case of cross-lingual IR, where document terms and query terms happen to be of the same language, the same system was used for both cross-lingual and monolingual IR. For simple monolingual IR, the translation matrix is an identity matrix (a diagonal matrix with 1's on the diagonal). In that case, the retrieval model is the same as the one proposed by (Miller, Leek and Schwartz, 1999). For thesaurus-based retrieval, the translation matrix is the thesaurus.

Our system can easily accommodate the all-stems technique for stemming and the mapping technique for orthographic resolution, since both are simple probabilistic translations. In CLIR, these translations are applied before the translations to English terms. In other words, the translation from a document term to a query term consists of a number of intermediate translations. It is easy to verify that the translation matrix from document terms to query terms is the product of the intermediate translation matrixes.

5 OFFICIAL RESULTS

In all submitted runs, the document terms are unstemmed Arabic words. Words with apparently incorrect endings such as substitution of ALEF MAKSURA (ى) for YEH (ي) were handled automatically as described in Section 2.2. We submitted one official monolingual run and four official cross-lingual runs as follows:

- BBN10MON. Our monolingual run. Only the title and description fields were used for query formulation. Queries consist of Arabic stems. In query processing, each Arabic word is replaced by its stem(s). The statistical thesaurus described before was used for translations between Arabic stems.

Stop words were removed. Our stop word list was obtained from Yaser Al-Onaizan at ISI (http://www.isi.usc.edu). That list was augmented with a few manually selected high frequency words from the AFP corpus.

The mapping technique was used for orthographic resolution. The all-stems technique was used for stemming. Both were applied before the thesaurus translations of Arabic stems.

Automatic query expansion was used to add additional terms to the queries. An initial retrieval was performed on an Arabic corpus consisting of AFP (i.e. the TREC 2001 corpus) and additional articles from newspaper sources Al-Hayat and An-Nahar. For each query, 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 re-weighted:

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

where D_i's are the top retrieved documents.

- BBN10XLC. Cross-lingual run without query expansion. Only the title and description fields of the English topics were used for query formulation. Term translation used both the manual bilingual dictionary and the statistical bilingual dictionary described in the previous section.

- BBN10XLB. Cross-lingual run with Arabic expansion. In addition to BBN10XLC, Arabic query expansion terms were used. The same query expansion procedure in BBN10MON was used here.

- BBN10XLA. Cross-lingual run with Arabic and English expansions. In addition to BBN10XLB, English expansion terms were used. English documents were retrieved from a newspaper corpus with 1.2 million articles from sources AP, Reuters and FBIS.

- BBN10XLD. Cross-lingual run with long queries. Same as BBN10XLA, except the narrative field was also used in query formulation. Arabic and English expansions were used.

The mapping technique was used for orthographic resolution and the all-stems technique was used for stemming in BBN10MONO. In contrast, in the cross-lingual runs, normalization and sure-stem were used in deriving term translations from the UN corpus.

Table 1 shows the TREC average precision for each run. In addition, it shows the number of queries in each run that achieved the best monolingual or cross-lingual performance among all submitted runs and the number of queries above the median. Overall, all our runs achieved very good performance.

Table 1 Retrieval results for official runs

	Average Precision	=best(out of 25)	>median(out of 25)
BBN10MON	0.4537	14	21
BBN10XLA	0.4382	6	23
BBN10XLB	0.4639	8	24
BBN10XLC	0.3604	0	22
BBN10XLD	0.4453	3	22

6 EXPERIMENTS USING SHORT QUERIES

The TREC 2001 topics are very long. Excluding stop words, the full topics have 26 English words per topic. Without the narrative field, the average query length is 12 words per query, still too long for typical ad hoc retrieval. The title field, which has an average of 6.6 words per topic, is more realistic. Table 2 shows the scores our official runs would have achieved h`ad we used only the title field in query formulation. The degradation due to the shortened queries is modest, except for BBN10XLA, for which the degradation is very large.

Table 2 Title and description vs title-only for query formulation

	BBN10MON	BBN10XLA	BBN10XLB	BBN10XLC
Title+Desc words (official runs)	0.4537	0.4382	0.4639	0.3604
Title words	0.4222	0.3699	0.4475	0.3441

7 MONOLINGUAL EXPERIMENTS

The goal of the following experiments is to demonstrate the impact of a number of issues on monolingual retrieval. In all experiments, query formulation used the title and description fields of the topics.

a. No text processing except for the removal of stop words in query and indexing. The translation matrix is an identity matrix.

b. All-stems stemming was used in addition to the removal of stop words. Elements in the translation matrix are “translation” probabilities from unstemmed words to stems.

c. The difference from b is that sure-stem stemming was used.

d. In addition to b, the mapping technique was used for orthographic resolution.

e. The only difference from d is that normalization instead of mapping was used for orthographic resolution.

f. Same as d, except that the statistical thesaurus was used for term translation

g. In addition to f, query expansion was used, based on AFP, Al-Hayat, and An-Nahar. This is our official monolingual run, BBN10MON.

h. Same as g, except that query expansion used only the AFP corpus.

Table 3 Monolingual results

a	b	c	d	e	f	g	h
0.1873	0.2388	0.2492	0.3145	0.3131	0.3682	0.4537	0.4571

Retrieval scores in Table 3 show that:

- Stemming is very useful for Arabic retrieval (a->b). The absolute change in performance is 0.05. The value of stemming seems to be even greater for Arabic than for English monolingual retrieval. This is not surprising given the fact that Arabic has more complex morphologies.

- There is a small difference between all-stems and sure-stem (b->c), the latter being slightly better. The difference is not statistically significant.

- Orthographic resolution is very important (b->d). The change in performance is 0.075. This suggests that word spellings in the documents are very different from those in the queries.

- There is little difference between the mapping and the normalization techniques for orthographic resolution (d->e). More research is needed to determine whether a better probability estimation procedure will improve the mapping technique.

- The automatically derived thesaurus is very useful (d->f). The performance change is 0.05. We believe that most of the improvement is due to the broken plurals successfully resolved by the thesaurus. The rest of the improvement is probably due to general synonyms captured by the thesaurus.

- Query expansion is very useful for TREC 2001 queries (f->g). The performance change is 0.085. This is not very surprising given the success of query expansion techniques in earlier TRECs.

- Query expansion using only AFP is as effective as using the combined corpus of AFP, Al-Hayat and An-Nahar (g->h). The advantage of using a larger corpus for query expansion suggested by earlier studies (e.g. Kwok and Chan, 1998) is not observed. The probable reason is that the AFP corpus already has enough relevant documents for the queries (165 relevant documents per query on average). The additional relevant documents in Al-Hayat and An-Nahar did not improve the worthiness of the top documents for the purpose of query expansion.

8 CROSS-LINGUAL EXPERIMENTS

8.1 Impact of Orthographic Variations

We compared BBN10XLC with an unofficial run where orthographic variations were not handled. Other conditions are the same for both runs. We found that there is little difference between the two runs (0.3604 vs 0.3584). This is very different from monolingual retrieval, where orthographic resolution is critical. However, the result is not surprising given the fact that different variants of the same word can be translated individually. Indeed, a casual inspection of the Buckwalter lexicon indicates that it often has separate entries for different spellings of the same word. It appears that the UN corpus also contains such spelling variations.

8.2 Effect of Arabic Stemming in Inducing a Bilingual Lexicon from a Parallel Corpus

We have compared three modes of learning term translations from the UN corpus. The first did not use stemming. The second used sure-stem. The third used all-stems. All three have pros and cons. The first keeps the maximum amount of word distinction, but requires more training data. The third requires less training data due to the reduced dimensionality, but increases word ambiguity, and the probability estimates are affected due to the one-to-many mapping from words to stems. The second is a compromise.

The retrieval scores in Table 4 show that no-stem is slightly better than sure-stem, which is slightly better than all-stems. While the differences are too small to make firm conclusions, they suggest that Arabic stemming is not an important issue in CLIR.

Table 4 Three modes of GIZA++ training: no-stem, sure-stem and all-stems

No-stem	Sure-stem	All-stems
0.3106	0.2994	0.2895

8.3 Impact of Resource Combination

Table 5 shows the retrieval scores when:

- The Buckwalter lexicon was used for term translation.
- The augmented Buckwalter lexicon (with additional word pairs from Sakhr and NMSU) was used.
- The UN corpus was used.
- All resources were combined.

Table 5 Impact of resource combination

Buckwater only	Augmented Buckwalter	UN only	ALL (BBN10XLA)
0.2695	0.2697	0.2994	0.3604

The scores indicate that the additional translation pairs from Sakhr and NMSU are not helpful. The combination of the UN and the manual lexicon significantly outperforms either resource alone, suggesting that the word ambiguity problem in Arabic is satisfactorily handled by complementing a manual lexicon with a parallel corpus. The result is consistent with our TREC9 Chinese CLIR work (Xu and Weischedel 2000).

8.4 Query Expansion

Table 6 shows that both English and Arabic expansion terms improve retrieval scores. The Arabic expansion terms are more effective than English expansion terms. This is expected because we know that the particular English corpus we used for query expansion is not a very good match for the Arabic test corpus. It is disappointing that using both sources of expansion terms does not improve retrieval further. In fact, it is worse than using Arabic expansion alone. One possible reason is that the weights for English expansion terms are larger than they should be. That suggests that reducing the weight on English expansion terms may result in better retrieval.

Table 6 Effect of query expansion on CLIR retrieval

No expansion (BBN10XLC)	English expansion	Arabic expansion (BBN10XLB)	English & Arabic expansions (BBN10XLA)
0.3604	0.4060	0.4639	0.4382

9 CONCLUSIONS

Concerning monolingual Arabic retrieval, the following proved true empirically:

- As in other languages, stemming is very important.
- Proper handling of orthographic variations is critical; the probabilistic model handled this type of ambiguity.
- A statistically derived thesaurus from a parallel corpus can effectively cope with the broken plural problem.
- Automatic query expansion by unsupervised relevance feedback proved very helpful, just as it has in other languages.

Concerning cross-lingual IR, the following was demonstrated empirically:

- Combining manual lexicons and parallel corpora in a probabilistic model gave much better performance than either alone.
- Stemming and handling of orthographic variations proved less critical for CLIR than for monolingual IR.
- Query expansion significantly improved retrieval performance, though query expansion in Arabic alone proved most effective.
- Cross-lingual retrieval outperformed monolingual retrieval, as it had in our Chinese experiments in TREC-9.

Acknowledgement: We would like to thank Ghada Osman, Mohamed Noamany and John Makhoul for their invaluable help.

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. L. Kwok and M. Chan, 1998. “Improving Two-Stage Ad-Hoc Retrieval for Short Queries.” In proceedings of SIGIR 1998.

D. Miller, T. Leek, and R. Schwartz, 1999. “A Hidden Markov Model Information Retrieval System.” In Proceedings of ACM SIGIR 1999.

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

J. Xu and R. Weischedel, 2000. “TREC9 Crosslingual Retrieval at BBN”, *TREC9 Proceedings*.

J. Xu, R. Weischedel, and C. Nguyen, 2001. “Evaluating a Probabilistic Model for Cross-lingual Retrieval.” In proceedings of ACM SIGIR 2001, pp. 105-110.