Empirical Studies in Strategies for Arabic Retrieval

Jinxi Xu
BBN Technologies
50 Moulton Street
Cambridge, MA 02138
jxu@bbn.com

Alexander Fraser
USC/ISI
4676 Admiralty Way
Marina del Rey, CA 90292
fraser@isi.edu

Ralph Weischedel
BBN Technologies
50 Moulton Street
Cambridge, MA 02138
weisched@bbn.com

ABSTRACT

This work evaluates a few search strategies for Arabic monolingual and cross-lingual retrieval, using the TREC Arabic corpus as the test-bed. The release by NIST in 2001 of an Arabic corpus of nearly 400k documents with both monolingual and cross-lingual queries and relevance judgments has been a new enabler for empirical studies. Experimental results show that spelling normalization and stemming can significantly improve Arabic monolingual retrieval. Character tri-grams from stems improved retrieval modestly on the test corpus, but the improvement is not statistically significant. To further improve retrieval, we propose a novel thesaurus-based technique. Different from existing approaches to thesaurus-based retrieval, ours formulates word synonyms as probabilistic term translations that can be automatically derived from a parallel corpus. Retrieval results show that the thesaurus can significantly improve Arabic monolingual retrieval. For cross-lingual retrieval (CLIR), we found that spelling normalization and stemming have little impact.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing--*Dictionaries, Indexing methods, Linguistic processing, Thesauruses*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval--*Relevance feedback, Retrieval models*

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Arabic retrieval, stemming, spelling normalization, thesaurus, n-grams, broken plurals, parallel corpora, cross-lingual retrieval

1 INTRODUCTION

Arabic is one of the most widely used languages in the world, yet there are relatively few studies on the retrieval of Arabic documents in the literature. Furthermore, the lack of a realistically large test corpus has been a problem in past studies on Arabic retrieval. This work will explore a few strategies for the retrieval of Arabic documents, using the recently available TREC Arabic corpus (Voorhees, 2001) for evaluation.

Arabic is a challenging language for information retrieval (IR) for a number of reasons. First, orthographic variations are prevalent in Arabic; certain combinations of characters can be written in different ways. For example, sometimes in glyphs combining HAMZA or MADDA with ALEF the HAMZA or MADDA is dropped, rendering it ambiguous as to whether the HAMZA or MADDA is present. Second, Arabic has a very complex morphology. Third, broken plurals are common. Broken plurals are somewhat like irregular English plurals except that they often do not resemble the singular form as closely as irregular plurals resemble the singular in English. Because broken plurals do not obey normal morphological rules, they are not handled by existing stemmers. Fourth, Arabic words are often ambiguous due to the tri-literal root system. In Arabic, a word is usually derived from a root, which usually contains three letters. In some derivations one or more of the root letters may be dropped, rendering many Arabic words highly ambiguous with one another. Fifth, short vowels are omitted in written Arabic. Sixth, synonyms are widespread, perhaps because variety in expression is appreciated as part of a good writing style by Arabic speakers (Noamany, 2001).

Those problems make exact keyword match inadequate for Arabic retrieval. We will explore a few search strategies to address the problems. Two techniques, spelling normalization and stemming, are well-known techniques for IR. Our experiments show that while these techniques can significantly improve retrieval, they are not adequate. The third technique, retrieval based on character n-grams, has been used by a few studies (Darwish et al, 2001; Mayfield et al, 2001; Kwok et al, 2001). We found that tri-grams from stems modestly improved retrieval on the test corpus, but the improvement is not statistically significant. To further improve Arabic retrieval, we propose a statistical thesaurus to deal with the large number of broken plurals and synonyms in Arabic. Our approach differs from existing techniques in that it is probabilistically motivated and employs a parallel corpus rather than a monolingual corpus for determining word associations. Experiments show that the thesaurus can significantly improve monolingual retrieval.

We also studied the effect of spelling normalization and Arabic stemming on cross-lingual retrieval, where English queries were used to retrieve Arabic documents. Interestingly, they had little impact on our CLIR experiments, as we had sufficient data to learn translations of each of the variants.

SIGIR '02, August 11-15, 2002, Tampere, Finland.

2 RETRIEVAL STRATEGIES

2.1 Spelling Normalization

Arabic orthography is highly variable. For instance, changing the letter YEH (ي) to 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, in our experiments we detected such "errors" using a stemmer (the Buckwalter Stemmer, to be discussed later) and restored the correct word ending.

A more problematic type of spelling variation is that certain glyphs combining HAMZA or MADDA with ALEF (e.g. أ , إ and آ) are sometimes written as a plain ALEF (ا), possibly because of their similarity in appearance. Often, both the intended word and what is actually written are valid words. This is much like confusing "résumé" with "resume" in English. Since both the intended word and the written form are correct words, it is impossible to correct the spellings without the use of context. In our experiments, we converted every occurrence of these glyphs to a plain ALEF.

2.2 Arabic Stemming

Arabic has a complex morphology. Most Arabic words (except some proper nouns and words borrowed from other languages) are derived from a *root*. A root usually consists of three letters. We can view a word as derived by first applying a *pattern* to a root to generate a stem and then attaching prefixes and suffixes to the stem to generate the word (Khoja and Garside, 2001). For this reason, an Arabic stemmer can be either root-based or stem-based.

Experiments in this work used a stem-based stemmer, the Buckwalter Stemmer (Buckwalter, 2001), for two reasons. First, it is a simple algorithm and can be easily re-implemented in a way usable in our retrieval system. Second, judging from the published results in the TREC 2001 Proceedings (Voorhees, 2001), it is at par with other stemmers for retrieval purposes. The algorithm 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. If several valid combinations are found, it returns all of the stems. We re-implemented the stemmer to make it faster and compatible with the UTF8 encoding. We also modified it so that if no valid combination of prefix-stem-suffix is found, the word itself is returned as the stem.

2.3 Character N-grams

Broken plurals, analogous to irregular nouns in English (e.g. "woman/women"), are very common in Arabic. There is no existing rule-based algorithm to reduce them to their singular forms, and it seems that it would be not be straight-forward to create such an algorithm. As such, broken plurals are not handled by current Arabic stemmers.

One technique to address this problem is to use character n-grams. Although broken plurals are not derived by attaching word affixes, many of the letters in broken plurals are the same as in the singular forms (though sometimes in a different order). If words are divided into character n-grams, some of the n-grams from the singular and plural forms will probably match. This technique can also handle words that have a stem but cannot be stemmed by a stemmer for various reasons. For example, the Buckwalter stemmer employs a list of valid stems to ensure the validity of the resulting stems. Although the list is quite large, it is still not complete. N-grams in this case provide a fallback where exact word match fails.

In this work, we have experimented with n-grams created from stems as well as n-grams from words. N-grams were created by applying a shifting window of *n* characters over a word or stem. If the word or stem has fewer than *n* characters, the whole word or stem was returned.

2.4 Deriving an Arabic Thesaurus from a Parallel Corpus

In all natural languages, synonyms present a challenge to IR. As discussed before, this problem is especially serious for Arabic. A common technique to address synonyms is to use a thesaurus. Here we discuss how to automatically derive an Arabic thesaurus from a parallel corpus, based on the intuition that synonyms in one language tend to be translated to the same words in the other language.

We treat synonyms as probabilistic translations between words. Our approach therefore attempts to estimate $p(b|a)$, the translation probability from one Arabic word *a* to another Arabic word *b*. We can imagine the user first translates *a* to some English word *x* and then translates *x* to *b*. Theoretically any English word could be the intermediate translation *x*. Therefore, $p(b|a)$ can be expressed as:

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

To overcome the problem of data sparseness, the probability was smoothed using a mixture model:

$$p(b \mid a) = \beta\, p_{diag}(b \mid a) + (1-\beta)\ p_{thesaurus}(b \mid a)$$

where $p_{diag}(b|a)=1$ if $a=b$ and 0 otherwise. The smoothing parameter β controls how much confidence we have in the original word and how much confidence we have in the thesaurus.

In our experiments, the probability estimates $p(x|a)$ and $p(b|x)$ were estimated from a parallel corpus using GIZA++ (Och and Ney, 2000). GIZA++ is a freely available statistical machine translation toolkit whose theory was based on the statistical translation work pioneered by (Brown et al, 1993). GIZA++ implemented several models proposed by Brown for estimating term translation probabilities from sentence aligned parallel corpora; Model 1 was used in this work for its efficiency.

The parallel corpus used in our experiments was obtained from the United Nations (UN). The UN website (http://www.ods.un.org) publishes all UN official documents under a document repository, which is accessible by paying a monthly fee. We extracted around 38,000 document pairs from the UN archive, with over 50 million English words and a similar number of Arabic words. An algorithm developed in-house was used to align the corpus at the sentence level.

While thesaurus-based retrieval has been extensively studied over the decades (Spark Jones, 1971; Deerwester et al, 1990; Jing and Croft, 1994; Schütze and Pedersen, 1994), our approach is different from existing ones in two ways. Our approach extracts word associations from a parallel corpus, while existing techniques employed a monolingual corpus. We believe that a parallel corpus contains stronger semantic clues and therefore can result in more reliable word associations than a monolingual corpus. Second, word associations in our technique have a well-defined probabilistic interpretation. This enables a principled integration of the thesaurus model and a probabilistic retrieval model. An effective thesaurus-based technique must deal with the problem of word polysemy or ambiguity, which is particularly serious for Arabic retrieval. For words with multiple meanings, a probabilistic technique like ours can emphasize probable meanings with high probabilities and discount unlikely ones with low probabilities. This curbs the impact of spurious word associations on retrieval.

3 RETRIEVAL SYSTEM

Our retrieval system was based on the probabilistic generative model described in (Xu, Weischedel and Ngyuen, 2001). It ranks documents according to the probability that a query Q is generated from a document D:

$$p(Q|D) = \prod_{t_q \text{ in } Q} \left[\alpha\, p(t_q|GL) + (1-\alpha) \sum_{t_d \text{ in } D} p(t_d \mid D)\, p(t_q \mid t_d)\right]^{f(t_q,Q)}$$

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

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

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

The retrieval model was originally proposed for CLIR. Since monolingual retrieval is a special case of CLIR, where the query terms and document terms happen to be of the same language (e.g. Arabic), the same retrieval system was also used for monolingual experiments. For simple monolingual IR, the lexicon used for term "translation" is an identity matrix, where $p(a|b)$=1 if a=b and 0 otherwise. For thesaurus-based monolingual retrieval, the translation probabilities were calculated from the UN parallel corpus as discussed in the previous section.

For CLIR, translation probabilities were estimated from the same UN parallel corpus and were combined with a manual bilingual lexicon, the Buckwalter lexicon (Buckwalter, 2001), with around 86,000 Arabic-English word pairs. We assume that translation probabilities in the manual lexicon are uniformly distributed. That is, if an Arabic term has n English translations, each translation gets 1/n probability. The two lexical resources were combined using a mixture model with equal weights to produce a single probabilistic bilingual lexicon for term translation.

In our monolingual experiments, the background corpus GL is the TREC 2001 Arabic corpus. In our CLIR experiments, the background corpus consists of newspaper articles in TREC English disks 1-5.

4 EXPERIMENTAL SETUP

Our experiments were performed on the TREC 2001 Arabic corpus (Voorhees, 2001). That corpus has 383,872 Arabic documents from Agence France Presse (AFP) with 25 test topics. Each topic has three versions, Arabic, English and French. The Arabic topics were used in our monolingual experiments and the English topics in our CLIR experiments. Only the title and description fields of the topics were used in query formulation. Retrieval performance was measured using the TREC non-interpolated average precision (Voorhees, 2001).

In addition to average precision, standard *t*-test (Hull, 1993) was used to determine the statistical significance of the retrieval difference between two retrieval runs. A difference is considered to be statistically significant if the *p*_value is less than 0.05.

We used the Arabic stop word list compiled by Yaser Al-Onaizan (http://www.isi.edu/~yaser/arabic/arabic-stop-words.html). That list was augmented with a handful of manually selected high frequency words from the AFP corpus. In our experiments, English words were stemmed using the Porter stemmer (Porter, 1980).

5 BASELINE FOR ARABIC MONOLINGUAL RETRIEVAL

Experiments in this section will establish a baseline for Arabic monolingual retrieval. Since spelling normalization and stemming are well-studied IR techniques, they were employed in the baseline.

Ambiguities arise when the Buckwalter stemmer returns several stems for a word. We considered two alternatives, *sure-stem* and *all-stems*. With *sure-stem*, we only stemmed a word if it has exactly one possible stem. Otherwise, the word was left alone. Both the documents and the queries were processed in the same manner. With *all-stems*, we did not stem the words in the documents but instead probabilistically "translated" them to stems. The query words were stemmed though, by replacing each word by its possible stem(s). In the absence of training data, we assume that all possible stems are equally-probable. That is, if a word w has n possible stems s_1, s_2, ...s_n, then $p(s_i|w)$=1/n. 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.

To show the impact of spelling normalizations (see Section 2.1) and stemming, four retrieval runs were carried out:

1. There was no text processing except for the removal of the stop words from the documents and the queries.
2. Spelling normalization was used in addition to stop word removal
3. Sure-stem was used in addition to stop word removal and spelling normalization
4. All-stems was used in addition to stop word removal and spelling normalization

In 1-3, translation probability $p(t_q|t_d)$=1 if t_q=t_d and 0 otherwise. That is, a term was only translated to itself.

Results in Table 1 show that spelling normalization produced a 22% relative improvement in retrieval performance. The improvement is statistically significant (*p*_value=0.017). This is not surprising because spelling variations are prevalent in the test corpus. Statistics from the test corpus indicate that about 80% of the words containing a glyph combining HAMZA or MADDA with ALEF also have a variant with just a plain ALEF in the corpus. A human assessment of a few hundred sentences indicates that if two words exist which are only different in that one has such a glyph and the other has a plain ALEF, most occurrences of the word with a plain ALEF would be written as the word containing the glyph under a strict writing standard.

Sure-stem stemming produced a remarkable 40% relative improvement in performance. The improvement is statistically significant (*p*_value=0.003). The impact of stemming on Arabic retrieval is far greater than the impact on English retrieval (Harman, 1991). The complex morphology of Arabic causes a high level of synonymy in its vocabulary. For example, the TREC Arabic corpus has over 500,000 unique unstemmed words. In comparison, an English corpus of comparable size (Wall Street Journal in TREC disks1&2) has about 200,000 unstemmed words. Many Arabic words can be conflated to the same stem. Statistics collected from the test corpus indicate that 1,300 stems have 50 or more unstemmed words, and 300 stems have over 100 unstemmed words. Given such data, it is understandable that stemming has such a big impact on retrieval.

There is little difference between the retrieval scores of sure-stem and all-stems. The difference is not statistically significant. Statistics show that 7% of words in the test corpus have two or more possible stems. The percentage is probably too small to have an impact on retrieval. We will use the sure-stem result as our monolingual baseline.

Table 1: Impact of spelling normalization and stemming on Arabic monolingual retrieval

Stop words removal	Normalization	Sure-stem	All-stems
0.1873	0.2291	0.3208	0.3131

6 IMPROVING ON THE MONOLINGUAL BASELINE

6.1 N-gram-based Retrieval

Two methods of creating n-grams were tried: from words and from stems. Retrieval scores in Table 2 show that stem-based n-grams are better than word-based n-grams for retrieval. The probable reason is that some of the word-based n-grams are prefixes or suffixes, which can cause false matches between documents and queries. The best results were obtained with trigrams, suggesting that bigrams carry too little contextual information while 4-grams and longer ones simply simulate word or stem-based retrieval.

Trigrams from stems produced the best result, with a 5% relative improvement over the baseline (from 0.3208 to 0.3365). However, the improvement is not statistically significant, meaning the benefit of using n-grams is not conclusive.

Table 2 : Retrieval results using n-grams

	Bigrams	Trigrams	4-grams
Words	0.1461	0.2990	0.2900
Stems	0.1655	0.3365	0.3165

6.2 Thesaurus-based Retrieval

Table 3 compares the retrieval performance of the statistical thesaurus described in section 2.4 with the baseline and the trigram results. The smoothing parameter β in the mixture model was set to 0.1 in the experiment. The relative improvement over the baseline is 18%. The improvement over trigrams is 13%. The improvements in both cases are statistically significant (*p*_value=0.006 and 0.031 respectively). The results clearly show that the thesaurus is a better technique than the use of n-grams for improving on the monolingual baseline. While it appears that both broken plurals and general synonyms have contributed to the improved retrieval, a breakdown of the two factors is not available because we do not have a human assessment on the word pairs in the thesaurus. This is left for future work.

Table 3 : Comparing baseline, trigrams and thesaurus for Arabic monolingual retrieval

Baseline	Trigrams	Thesaurus
0.3208	0.3365	0.3790

Figure 1 shows the retrieval performance as a function of the smoothing parameter β. As we discussed in Section 2.4, a larger β places more confidence in the original terms while a smaller β places more confidence in the translations learned from the parallel corpus. Retrieval performance peaks when β=0.1. Overall, retrieval performance is not sensitive to the choice of β: Any value between 0 and 0.4 works fine.

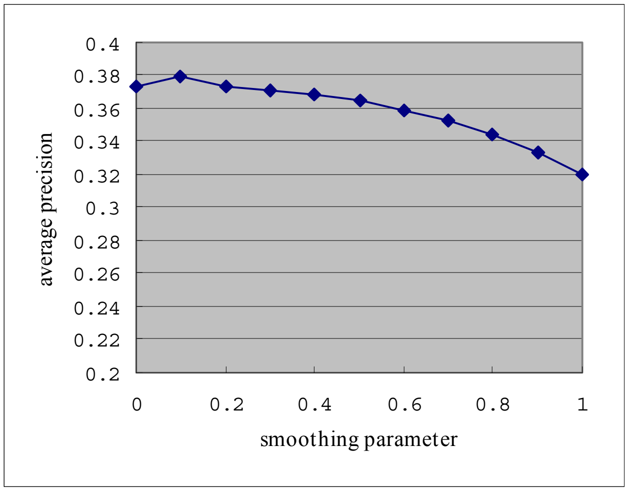

Figure 1: The effect of smoothing parameter β on thesaurus-based retrieval.

Using a thesaurus can be viewed as a query expansion technique. Another commonly used query expansion technique in IR is local feedback (Buckley et al, 1996). Local feedback selects terms from top retrieved documents and adds them to the initial query. The expanded queries usually significantly improve retrieval performance. One may wonder whether thesaurus and local feedback overlap, and whether using one eliminates the need for the other.

To address that concern, we compared two versions of local feedback. In one, the statistical thesaurus was used in both the initial retrieval and the final retrieval with the expanded queries.

In the other, the thesaurus was not used at all. In our experiments, local feedback selected 50 terms from 10 top retrieved documents based on their total *tf×idf* weight in the top documents. The expansion terms and the original query terms were re-weighted. In the probabilistic retrieval model used in this work, we interpret the weight of a query term to be the frequency of the term being generated in query generation. As described in Section 3, the frequency is used as an exponent in the retrieval function. With local feedback, the frequency of a term *t* in a query *Q* is calculated:

$$f(t,Q) = f_{old}(t,Q) + 0.4 \sum_{1 \le i \le 10} tfidf(t, D_i)$$

where D_i is a top retrieved document, *tfidf*(t, D_i) is the *tf×idf* weight of term *t* in D_i, $f_{old}(t, Q)$ and $f(t, Q)$ are the old and new frequencies of *t* in the query. The *tf* and *idf* functions were based on the ones described in (Allan et al, 2000).

Results in Table 4 show that using local feedback and the thesaurus together is 15% better than local feedback alone. This is comparable to the 18% improvement produced by the thesaurus when local feedback was not used. The improvement is statistically significant (*p*_value=0.015). The results suggest that local feedback and the thesaurus are two different types of query expansion techniques.

Table 4: The impact of the thesaurus when used together with local feedback

Local feedback only	Local feedback + Thesaurus
0.4020	0.4630

7 CROSS-LINGUAL EXPERIMENTS

To explore the impact of spelling normalization and Arabic stemming on CLIR, we have compared three versions of bilingual lexicon creation for term translation. All three were formed from the UN parallel corpus and the Buckwalter lexicon using the same procedure described in Section 3. The only difference is that the Arabic terms are non-normalized words in the first lexicon, are non-normalized stems in the second, and are normalized stems in the third.

Results in Table 5 show that the differences between the three versions of lexicon are very small. The differences are not statistically significant. This is different from monolingual IR, where spelling normalization and stemming had a very big impact. The explanation is that the UN parallel corpus, with over 50 million words in each language, has enough data to enable GIZA++ to reliably learn the English translations for most Arabic words. It appears that the Buckwalter lexicon also lists the most common Arabic spelling variants. Therefore, the advantage of translating words by groups over translating them individually is very small. Besides, stemming and normalization invariably introduce ambiguity. Apparently, the small benefit of stemming and spelling normalization was canceled by the introduced ambiguity. Although their impact on CLIR performance is small, spelling normalization and stemming are still useful because they reduce the need for memory because there are fewer entries in the lexicon and they improve the retrieval speed by simplifying the score computation.

Table 5: The effect of spelling normalization and stemming of Arabic words on CLIR

Translation of non-normalized words	Translation of non-normalized stems	Translation of normalized stems
0.3447	0.3584	0.3604

One might wonder whether we can use the Arabic monolingual thesaurus to improve CLIR. The assumption is that the thesaurus would be useful for cases where we do not know how to translate a word but we do know how to translate its synonyms. We did not run such an experiment because it is computationally prohibitive, requiring multiplying three large matrices. But based on a similar argument to the one we just made, it is unlikely such an experiment would produce better retrieval results given that we already have enough resources for effective term translation.

8 RELATED WORK

A key enabling element for our work has been the release by NIST of a large Arabic corpus with relevance judgments for both monolingual and cross-lingual retrieval (Voorhees, 2001). Arabic stemming algorithms can be roughly classified as either stem-based or root-based. While stem-based algorithms such as the Buckwalter stemmer (Buckwalter, 2001) remove prefixes and suffixes from Arabic words, root-based algorithms (Beesley 1996; Khoja and Garside, 2001) further reduce stems to roots. A study reported better results using roots than stems on a very small test corpus (Abu-Salem et al, 1999). However, a later study on the TREC corpus showed that stem-based retrieval is more effective than root-based retrieval (Aljlayl et al, 2001). The idea of using n-grams for Arabic retrieval appeared in several studies (Darwish et al, 2001; Mayfield et al, 2001; Kwok et al, 2001).

The use of thesauri in IR has been extensively studied under a variety of names, such as keyword clustering (Spark Jones, 1971), co-ocurrence thesauri (Jing and Croft 1994; Schütze and Pedersen, 1994) and Latent Semantic Indexing (Deerwester et al, 1990). One difference from existing approaches is that our thesaurus was derived from a parallel corpus instead of a monolingual corpus. Another difference is that our thesaurus is probabilistically motivated. A similar idea was proposed by (Berger and Lafferty, 1999). But that work used artificially synthesized training data while ours used a parallel corpus. The use of parallel corpora for Cross-lingual IR has been well studied (Sheridan and Ballerini, 1996; Nie et al, 1999; J. McCarley, 1999). The use of probabilistic generative models for IR appeared in a number of studies (Ponte and Croft, 1998; Miller et al, 1999; Hiemstra and de Jong, 1999).

9 CONCLUSIONS AND FUTURE WORK

This work evaluated a number of search strategies for the retrieval of Arabic documents, using the TREC Arabic corpus as the test bed. For Arabic monolingual retrieval, spelling normalization significantly improved retrieval performance by 22%. Stemming is critical, improving retrieval performance by 40%. Two techniques were explored to further improve retrieval. Tri-grams from stems produced a modest improvement in retrieval, but the improvement is not statistically significant. In comparison, a sophisticated statistical thesaurus boosted retrieval performance by 18%. We also studied the impact of spelling normalization and stemming on Arabic CLIR. Retrieval results show that their impact on CLIR is very small.

There are a number of areas for future work. First, we would like to compare our stemming algorithms with other Arabic stemming algorithms. Second, we would like to compare our thesaurus-based technique with similar techniques such as Latent Semantic Indexing. Third, we would like to apply our thesaurus-based technique to other languages. For example, by simply swapping the roles of Arabic and English, we can induce an English thesaurus instead of an Arabic one from the UN parallel corpus. Fourth, we would like to classify the word pairs in our thesaurus and investigate which category (broken plurals or general synonyms) has the largest impact on retrieval. Fifth, we would like to validate our techniques on more test corpora when they are available. Such a validation is necessary since the problem of incomplete relevance judgments is potentially severe for the TREC 2001 Arabic corpus (Gey and Oard, 2001).

REFERENCES

[1] Abu-Salem, H., Al-Omari, M., and Evens, M. 1999. "Stemming Methodologies Over Individual Query Words for an Arabic Information Retrieval System". In *JASIS*, 50(6), May, 1999.

[2] Aljlayl, M., Beitzel, S., Jensen, E., Lee, M., Grossman, D., Frieder, O., Chowdhury, A. and Holmes, D. "IIT at TREC-10." In *TREC 2001 Proceedings*.

[3] Allan, J., Callan, J., Feng, F., and Malin, D. 2000. "INQUERY at TREC8." In *TREC8 Proceedings*, Special publication by NIST, 2000.

[4] Beesley, K. 1996. "Arabic Finite-State Morphological Analysis and Generation." *COLING*-96, 1996.

[5] Berger, A., and Lafferty, J. 1999. "Information retrieval as statistical translation." In *Proceedings of SIGIR,* 1999.

[6] Brown, P., Della Pietra, S., Della Pietra, V., Lafferty J., and Mercer., R. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". In *Computational Linguistics,* 19(2), 1993.

[7] Buckley, C., Singhal, A., Mitra M., and Salton, G. 1996. "New Retrieval Approaches using SMART". In *TREC 4 Proceedings*, NIST Special Publications.

[8] Buckwalter, T. 2001. Personal Communications.

[9] Darwish, K., Doermann, D., Jones, R., Oard D., and Rautiainen, M. "TREC-10 Experiments at Maryland: CLIR and Video." In *TREC* 2001 *Proceedings*.

[10] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R. 1990. "Indexing by Latent Semantic Analysis." *JASIS*, 41, 391-407, 1990.

[11] Gey, F., and Oard, D. 2001. "The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries." In *TREC 2001 Proceedings*, to be published.

[12] Harman, D. 1991. "How effective is suffixing?" *JASIS*, 42:7-15, 1991.

[13] 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.

[14] Hull, D. 1993. "Using statistical testing in the evaluation of retrieval experiments." In *Proceedings of SIGIR,* 1993, pages 329-328.

[15] Jing Y., and Croft, W. B. 1994. "An Association Thesaurus for Information Retrieval." In *Proceedings of RIAO,* 1994, pages 146-160.

[16] Khoja S., and Garside, R. 2001. "Stemming Arabic Text", Tech Report, Dept of CS, Lancaster Univ., U.K.

[17] Kwok, K., Grunfeld, L., Dinstl, N., Chan, M. 2001. "TREC 2001 Question Answering, Web, Cross-lingual Experiments using PIRCS." In *TREC 2001 Proceedings*.

[18] Mayfield, J., McNamee, P., Costello, C., Piatko, C., and Banerjee. A. 2001. "JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, Video and Web retrieval." In *TREC 2001 Proceedings*.

[19] McCarley, J. 1999. "Should we translate the documents or the queries in cross-language information retrieval." In *Proceedings of ACL*, pages 208-214, June 1999.

[20] Miller, D., Leek, T., and Schwartz, R. 1999. "A Hidden Markov Model Information Retrieval System." In *Proceedings of SIGIR*, 1999.

[21] Nie, J., Simard, M., Isabella, P., Durand, R. 1999. "Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Text from the Web." In *Proceedings of SIGIR*, 1999, pages 74-81.

[22] Noamany, M. 2001. Personal communications.

[23] Och, F., and Ney, H. 2000. "Improved Statistical Alignment Models." In *proceedings of ACL,* 2000.

[24] Ponte, J., and Croft, W. B. 1998. "A language modeling approach to information retrieval." In *Proceedings of SIGIR*, 1998, pages 275-281.

[25] Porter, M. 1980. "An algorithm for suffix stripping." *Program* 14, 3(1980), pages 130-137.

[26] Schütze, H., and Pedersen, J. 1994. "A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval." In *Proceedings of RIAO*, 1994, pages 266-274.

[27] Sheridan, P., and Ballerini, J. 1996. "Experiments in Multilingual Information Retrieval using the SPIDER System." In *Proceedings of SIGIR*, 1996, pages 58-65.

[28] Spark Jones, K., 1971. *Automatic Keyword Classification for Information Retrieval*. Butterworth, London.

[29] Voorhees, E. 2001. *TREC 2001 Proceedings*. To be published by NIST.

[30] Xu, J,. Weischedel, R,. and Nguyen, C. 2001. "Evaluating A Probabilistic Model for Cross-lingual Information retrieval." In *Proceedings of SIGIR,* 2001, pages 105-110.