Building an Arabic Stemmer for Information Retrieval

Aitao Chen
School of Information Management and Systems
University of California at Berkeley, CA 94720-4600, USA
aitao@sims.berkeley.edu

Fredric Gey
UC Data Archive & Technical Assistance (UC DATA)
University of California at Berkeley, CA 94720-5100, USA
gey@ucdata.berkeley.edu

1 Summary

In TREC 2002 the Berkeley group participated only in the English-Arabic cross-language retrieval (CLIR) track. One Arabic monolingual run and three English-Arabic cross-language runs were submitted. Our approach to the cross-language retrieval was to translate the English topics into Arabic using online English-Arabic machine translation systems. The four official runs are named as BKYMON, BKYCL1, BKYCL2, and BKYCL3. The BKYMON is the Arabic monolingual run, and the other three runs are English-to-Arabic cross-language runs. This paper reports on the construction of an Arabic stoplist and two Arabic stemmers, and the experiments on Arabic monolingual retrieval, English-to-Arabic cross-language retrieval.

2 Background

Arabic has much richer morphology than English. Arabic has two genders, *feminine* and *masculine*; three numbers, *singular*, *dual*, and *plural*; and three grammatical cases, *nominative*, *genitive*, and *accusative*. A noun has the nominative case when it is a subject; accusative when it is the object of a verb; and genitive when it is the object of a preposition. The form of an Arabic noun is determined by its gender, number, and grammatical case. The definitive nouns are formed by attaching the Arabic article ال to the immediate front of the nouns. As an example, the Arabic word الطالبة means *the student* (feminine). Sometimes a preposition, such as ب (by) and ل (to), is attached to the front of a noun, often in front of the definitive article. For example, the Arabic word للّطالبين means *to the students* (masculine). Besides prefixes, a noun can also carry a suffix which is often a possessive pronoun. For example, the Arabic word بطالبي (by my student) can be analyzed as ب + طالب + ي, with one prefix ب (by) and one pronoun suffix ي (my). In Arabic, the conjunction word و (and) is often attached to the following word. For example, the word وبطالبها means *and by her student* (masculine). Arabic has two kinds of plurals: *sound* plurals and *broken* plurals. The *sound* plurals are formed by adding plural suffixes to singular nouns. The plural suffix is ات for feminine nouns in all three grammatical cases, و ن for masculine nouns in nominative case, and ين for masculine nouns in genitive and accusative cases. For example, the word مدرّسون (teachers, masculine) is the plural form of مدرّس (teacher, masculine) in nominative case, and مدرّسين (teachers, masculine) is the plural form of مدرّس (teacher, masculine) in genitive or accusative case. The plural form of مدرّسة (teacher, feminine) is مدرّسات (teachers, feminine) in all three grammatical cases. The dual suffix is ات for the nominative case, and ين for the genitive or accusative. The word مدرّسان means *two teachers*. The formation of broken plurals is more complex and often irregular; it is, therefore, difficult to predict. Furthermore, broken plurals are very common in Arabic. For example, the plural form of the noun طفل (child) is

أطفال (children), which is formed by attaching the prefix أ and inserting the infix ا. The plural form of the noun كتاب (book) is كتب (books), which is formed by deleting the infix ا. The plural form of امرأة (woman) is نساء (women). The plural form and the singular form are almost completely different. The examples presented in this secion show that an Arabic noun could potentially have a large number of variants, and some of the variants can be complex because of the prefixes, suffixes, and infixes. As an example, the word ولأطفالها (and to her children) can be analyzed as ها + أطفال + ل + و. It has two prefixes and one suffix.

Like nouns, an Arabic adjective can also have many variants. When an adjective modifies a noun in a noun phrase, the adjective agrees with the noun in gender, number, case, and definiteness. An adjective has a masculine singular form such as جديد (new), a feminine singular form such as جديدة (new), a masculine plural form such as جدّد (new), and a feminine plural form such as جديدات (new). For example, المدرّس الجديد means *the new teacher* (masculine), and المدرّسون الجدّد means *the new teachers* (masculine). The adjective has the feminine singular form when the plural noun denotes something inanimate. As an example, the word جديدة (new) in الكتب الجديدة (the new books) is the feminine singular form.

Arabic verbs have two tenses: perfect and imperfect. Perfect tense denotes actions completed, while imperfect denotes incompleted actions. The imperfect tense has four mood: indicative, subjective, jussive, and imperative [4]. Arabic verbs in perfect tense consist of a stem and a subject marker. The subject marker indicates the person, gender, and number of the subject. The form of a verb in perfect tense can have subject marker and pronoun suffix. The form of a subject-marker is determined together by the person, gender, and number of the subject. Take درس (to study) as an example, the perfect tense is درست for the third person, feminine, singular subject, درسوا for the third person, masculine, plural subject. A verb with subject marker and pronoun suffix can be a complete sentence. For example, the word درسته has a third-person, feminine, singular subject-marker ت (she) and a pronoun suffix ه (him), it is also a complete sentence, meaning "she studied him." Often the subject-makers are suffixes, but sometimes a subject-marker can be a combination of a prefix and a suffix. For example, the word study in a negative sentence is تدرسي (did not study). For verbs in imperfect tense, in addition to the subject-marker, a verb can also have a mood-marker.

3 Test Collection

The document collection used in TREC 2002 cross-language track consists of 383,872 Arabic articles from the Agence France Press (AFP) Arabic Newswire during the period from 13 May, 1994 to 20 December, 2000. There are 50 English topics with Arabic translations. A topic has three tagged fields: *title*, *description*, and *narrative*. The newswire articles are encoded in Unicode (UTF-8) format, while the topics are encoded in ASMO 708.

4 Preprocessing

Because the texts in the documents and topics are encoded in different schemes, we converted both the documents and topics to Windows CP-1256 encoding. The set of valid characters include the Arabic letters and the English letters in both lower and upper cases. The Arabic punctuation marks, ،, ؛, and ؟, were considered as delimiters. A consecutive sequence of valid characters was recognized as a word in the tokenization process. The words that are stopwords were removed during documents and topics indexing. We say a word is *minimally* normalized when أً, أٌ, إٍ, أَ, أُ, إِ, and آ are changed to ا. A word is *lightly* normalized when additionally the Shadda character (the character above ل in لّل) is deleted, and the characters آ, أ, and إ are changed to ا, the final ى is changed to ي, and the final ه is changed to ة. In the Arabic document collection, the word امرأة (woman) is sometimes spelled as أمرأة or إمرأة. The Arabic shadda character is sometimes dropped in spelling. For example, for the word مدرّس (teacher) is sometimes spelled as مدرس.

5 Construction of stopword list

At TREC 2001, we created an Arabic stopword list consisting of Arabic pronouns, prepositions, and the like that are found in an elementary Arabic textbook [4] and the Arabic words translated from an English stopword list. For TREC 2002, we first collected all the Arabic words found in the Arabic document collection. The number of unique Arabic words found in the collection after minimal normalization is 541,681. We then translated the Arabic words, word-by-word, into English using the *Ajeeb* online English-Arabic machine translation system available at http://www.ajeeb.com. From this Arabic-English bilingual wordlist, we created an Arabic stopword list consisting of the Arabic words whose translations consists of only English stopwords. The Arabic stopword list has 3,447 words after minimal normalization, containing stopwords such as لكنكم (you), فية (in him), بينهم (between them), and بعدما (after). The English stopword list has 360 words. There are a couple of reasons why the Arabic stopword list automatically generated is much larger than the English stopword list. First, pronouns can have more than one form. For example, the Arabic word for *these* has four forms: هاتان (feminine, nominative), هاتين (feminine, genitive/accusative), هذان (masculine, nominative), and هذين (masculine, genitive/accusative). Second, pronouns and prepositions are sometimes joined together.

6 Construction of stemmers

At TREC 2001, we built a rather simple Arabic stemmer to remove from words the definite article prefix ال, the plural suffixes ان, ون, and ات, and the suffix ة. At TREC 2002, we created two Arabic stemmers, a *MT-based stemmer* and a *light stemmer*.

6.1 MT-based stemmer

We built a MT-based Arabic stemmer from the Arabic words found in the Arabic documents and their English translations using the online *Ajeeb* machine translation system. We partitioned the Arabic words into clusters based on the English translations of the Arabic words. The Arabic words whose English translations, after removing English stopwords, are conflated to the same English stem form one cluster. And all the Arabic words in the same cluster are conflated to the same Arabic word, the shortest Arabic word in the cluster. For example, an English stemmer usually changes plural nouns into singular, so *children* is changed to *child*. In order to change the variants of the Arabic word for *child* or *children* to the same Arabic stem, we first grouped all the Arabic words whose English translations contain the headword *child* or *children*. Then in stemming, all the Arabic words in this group are changed to the shortest Arabic word in the group. The Arabic adjectives and verbs were stemmed in the same way. For English, we used a morphological analyzer [2] to map plural nouns into singular form, verbs into the infinitive form, and adjectives into the positive form. This stemmer changes the *broken* plural forms of an Arabic word into its singular form. The broken plural forms are common and irregular, so it is generally difficult to write a stemmer to change the broken plural forms to singular forms. For example, Table 1 presents part of the Arabic words whose English translations contain the headword *child* or *children*. All the Arabic words shown in table 1 belong to the same cluster since, after removing the English stopwords, the English translations consist of either the word *child* or *children*, both being conflated to the same word by the English morphological analyzer. In stemming, the Arabic words shown in table 1 are conflated into the same word طفل. The English translations were produced using the online Ajeeb machine translation system. One can also create an Arabic stemmer from English/Arabic parallel texts or bilingual dictionaries. With a large English/Arabic parallel corpus available, one can first align the texts at the sentence level, then use a statistical machine translation toolkit such as GIZA++ to create an Arabic-to-English translation table. If we keep only the most likely English translation for an Arabic word, then we have a bilingual wordlist. Using this bilingual wordlist, we can translate all the Arabic words found in the Arabic document collection into English. We can create an Arabic stemmer by partitioning the Arabic words into clusters, each consisting of the Arabic words whose English translations are conflated to the same word by the English morphological analyzer. Stemmers for other languages can also be automatically generated using this method as long as some translingual resources, such as MT, parallel texts, or bilingual dictionaries, are available.

Arabic word	English translation	Arabic word	English translation	Arabic word	English translation	Arabic word	English translation
أطفال	children	اطفالهن	their children	بطفل	by child	فالطفلة	then the child
أطفالا	children	اطفالي	my children	بطفلة	by child	فطفل	then child
أطفالنا	our children	الأطفال	children	بطفلتنا	by our child	كأطفال	as children
أطفاله	and his children	الاطفال	children	بطفلته	by his child	كالطفل	as the child
أطفاله	his children	الطفل	the child	بطفله	by his child	لأطفال	to children
أطفالها	her children	الطفلان	the children	بطفلها	by her child	لطفلها	to her child
أطفالهم	their children	الطفلة	the child	بطفلهما	by their child	للّطفلة	to the child
أطفالهن	their children	الطفلتان	the children	بطفلين	by children	وأطفالنا	and our children
أطفالي	my children	الطفلتين	the children	بطفليها	by her children	والأطفال	and the children
اطفال	children	الطفله	the child	طفل	child	وبطفل	and by child
اطفالا	children	الطفلين	the children	طفلا	child	وبطفلين	and by children
اطفالك	your children	بأطفال	by children	طفلان	children	وطفلة	and child
اطفالكم	your children	بأطفاله	by his children	طفلاها	her children	وطفلتان	and children
اطفالكن	your children	بأطفالها	by her children	طفلة	child	وطفلنا	and our child
اطفالنا	our children	بالأطفال	by the children	طفلت	child	وطفلها	and her child
اطفاله	his children	بالطفل	by the child	طفلتان	children	وطفليه	and his children
اطفالها	her children	بالطفلة	by the child	طفلتة	his child	وطفليها	and her children
اطفالهم	their children	بالطفلتين	by the children	طفلتنا	our child	ولأطفالها	and to her children
اطفالهما	their children	بالطفلين	by the children	طفلته	his child	وللّطفل	and to the child

Table 1: Arabic words whose English translations contain the headword *child* or *children*.

6.2 Light stemmer

We developed a second Arabic stemmer called *light stemmer* that removes only prefixes and suffixes. We identified one set of prefixes and one set of suffixes that should be removed based on the grammatical functions of the affixes, their occurrence frequencies among the Arabic words found in the Arabic document collection, the English translations of the affixes, and empirical evaluation using the test collection of the previous CLIR track. We generated three lists consisting of the initial, the first two, or the first three characters, respectively, of the Arabic words in the document collection, and three lists consisting of the final, the last two, or the last three characters, respectively, of the Arabic words. We then sorted the six lists of suffixes or prefixes in descending order by the number of unique words in which a prefix or suffix occurs. Table 2 presents the most frequent one-, two-, and three-character prefixes among the unique Arabic words found in the document collection. The frequency shown in the table is the number of unique Arabic words that begins with a specific prefix. Table 3 shows the most frequent one-, two-, and three-character suffixes among the unique Arabic words. The frequency count for a given suffix is the number of unique Arabic words that end with that suffix. We identified 9 three-character, 14 two-character, and 3 one-character prefixes that should be removed in stemming, and 18 two-character, and 4 one-character suffixes that should be removed in stemming. The 9 three-character prefixes are وال (and the), بال (by the), فال (then the), كال (as the), وللّ (and to the), مال, اال, سال, لال. The 14 two-character prefixes to be removed are the most frequent ones as shown in table 2. Our light stemmer shares many of the prefixes and suffixes that should be removed with the light stemmer developed by Larkey et al. [5] and the light stemmer developed by Darwish[3].

The stemmer non-recursively removes the prefixes in the pre-defined set of prefixes, and recursively removes the suffixes in the pre-defined set of suffixes in the following sequence.

1. If the word is at least five-character long, remove the first three characters if they are one of the following: وال,

Rank	Initial character	Frequency	Initial two characters	Frequency	Initial three characters	Frequency
1	و	117324	ال	55364	وال	19411
2	ا	94043	وا	32787	الم	12711
3	ب	49319	با	16789	بال	9079
4	ل	48862	للّ	10912	الا	6666
5	م	33776	وم	10124	الت	3907
6	ت	25649	وت	9196	الب	2813
7	س	23385	وب	8865	وبا	2760
8	ف	21828	لا	7482	است	2559
9	ك	19794	سي	7447	وسي	2372
10	ي	19004	وس	7155	الس	2260
11	ن	10905	وي	6772	فال	2213
12	د	8445	ول	6527	الع	1973
13	ر	8345	كا	6083	للّم	1919
14	ش	7058	فا	5648	الك	1915
15	غ	6680	او	4933	الف	1783
16	ه	6435	ما	4877	كال	1751
17	ج	6383	لي	4749	للّا	1736
18	أ	5394	بو	4702	الن	1665
19	ع	5207	كو	4583	واس	1613
20	ح	4450	ان	4415	الح	1610
28					وللّ	1391
168					مال	412
203					اال	365
262					سال	312
268					لال	306

Table 2: Most frequent initial character strings.

بال, فال, كال, وللّ, مال, اال, سال, لال.

2. If the word is at least four-character long, remove the first two characters if they are one of the following: وا, ال, با, للّ, وم, وت, وب, لا, سي, وس, وي, ول, كا, فا.
3. If the word is at least four-character long and begins with و, remove the initial letter و.
4. If the word is at least four-character long and begins with either ب or ل, remove ب or ل only if, after removing the initial character, the resultant word is present in the Arabic document collection.
5. Recursively strips the following two-character suffixes in the order of presentation if the word is at least four-character long before removing a suffix: ها, ية, هم, نا, ما, وا, يا, ني, يا, هن, كم, كن, تم, تن, ين, ان, ات, ون.
6. Recursively strips the following one-character suffixes in the order of presentation if the character is at least three-character long before removing a suffix: ة, ه, ي, ت.

Rank	Final character	Frequency	Last two characters	Frequency	Last three characters	Frequency
1	ا	91571	ها	26412	تها	6544
2	ن	69574	ين	24601	هما	6286
3	ي	52418	ان	19089	تين	4591
4	ة	44683	ات	17612	ليّن	4262
5	ه	34288	ون	15724	تهم	3836
6	ت	33351	ية	13877	يان	2960
7	م	27346	هم	13570	يتش	2747
8	ر	25748	نا	11794	تان	2722
9	و	21123	ما	8811	اني	2534
10	ل	18531	وا	8276	يون	2443
11	س	14668	يا	7702	رات	2250
12	د	13352	ني	7553	مان	2056
13	ى	12037	ته	7379	اته	2050
14	ف	11265	ري	5187	تنا	1953
15	ك	9278	لي	5090	رين	1918
16	ب	8863	ار	5027	نية	1833
17	ز	6973	ير	4869	رها	1805
18	ش	6777	يه	4611	نها	1801
19	ش	6777	تي	4377	ينا	1775
20	ع	5987	يل	4268	لها	1759

Table 3: Most frequent last character strings.

In our implementation, the suffix ني is removed only if the word is at least four-character long and the resultant word after removing the suffix is present in the Arabic document collection. The prefix وبال is often the combination of three prefixes و (and), ب (by), and ال (the), and should be removed. The light stemmer we used for the TREC 2002 experiments did not remove this prefix combination. We decided to remove the initial letter WAW (و) since it the most frequent initial letter and often is the conjunction word attached to the following word. The other two initial letters that were removed are BEH (ب) and LAM (ل). The prefix ب is sometimes a preposition prefix, meaning *by*, and the prefix ل is also sometimes a preposition prefix, meaning *to*. Our light stemmer removes ب and ل only when, after removing the prefix, the resultant stem is also a word in the collection.

Among the two-letter suffixes to be removed, six are pronoun suffixes (ها, هم, نا, هن, كم, كن); four are plural suffixes (ين, ان, ات, ون); and three are subject markers (وا, تم, تن). The suffix ية is a nisba ending. The single-letter suffix ة is the feminine ending, ه a pronoun suffix, ي a pronoun suffix, and ت a subject marker. Sometimes the suffix ة is inseparable since, if removed, the resultant word is completely a different word. As an example, the word الملكة means *the queen*, after removing the suffix ة, the resultant word الملك means *the king*.

7 Experimental Results

7.1 Retrieval system

The retrieval system we used for the experiments is an implementation of the retrieval algorithm presented in [1]. For term selection, we assume the top-ranked m documents in the initial search are relevant, and the rest of the documents

in the collection are irrelevant. For the terms in the documents that are presumed relevant, we compute term relevance weighting [6] as follows:

$$w_t = log\frac{m_t(n - n_t - m + m_t)}{(m - m_t)(n_t - m_t)} \quad (1)$$

where n is the number of documents in the collection, m the number of top-ranked documents after the initial search that are presumed relevant, m_t the number of documents among the m top-ranked documents that contain the term t, and n_t the number of documents in the collection that contain the term t. Then all the terms found in the top-ranked m documents are ranked in decreasing order by relevance weight w_t. The top-ranked k terms are weighted and then merged with the initial query terms to create a new query. Some of the selected terms may be in the initial query. For the selected top-ranked terms that are not in the initial query, the weight is set to 0.5. For those top-ranked terms that are in the initial query, the weight is set to 0.5*t_i, where t_i is the occurrence frequency of term t in the initial query. The selected terms are merged with the initial query to formulate an expanded query. When a selected term is one of the query terms in the initial query, its weight in the expanded query is the sum of its weight in the initial query and its weight assigned in the term selection process. For a selected term that is not in the initial query, its weight in the final query is the same as the weight assigned in the term selection process, which is 0.5. The weights for the initial query terms that are not in the list of selected terms remain unchanged.

A query, like a document, is normally represented in our retrieval system by a set of unique words in the query with within-query term frequency. For the experiments reported in this paper, a word occurring n times in a query is represented by n occurrences of the same word with within-query frequency of one.

7.2 Monolingual Retrieval Results

The BKYMON run is our only official Arabic monolingual run in which only the *title* and *desc* fields in the topics were indexed. After removing stopwords from both documents and topics, the remaining words were stemmed using Berkeley light stemmer as described in section 6.2. The stopword list used in this run was the one created from the translations of Arabic document words using the online Ajeeb machine translation. The development of the Arabic stoplist was described in section 5. The stopword list has 2,942 words after light normalization. Table 4 presents the evaluation results for additional retrieval runs.

The monolingual run *mon0* was produced without stemming. The words were lightly normalized and stopwords removed. Two runs were performed using overlapping trigram indexing, one without word boundary crossing (mon1) and the other with word boundary crossing (mon2) . For example, without word boundary crossing, the following trigrams are produced from the phrase بلوغا صيانة: صيا, يان, انة, بلو, لوغ, وغا. But with word boundary crossing, two additional trigrams, نة ب and ةبل, are produced. The words were lightly normalized and the stopwords were removed before trigrams were generated from the normalized words.

The monolingual run *mon3* used the light stemmer named *Al-Stem*, developed by Darwish [3]. The numeric digits from '0' to '9' are treated as part of a token in Darwish's stemmer which also reduces 616 unnormalized words found in the Arabic documents to empty string, effectively treating them as stopwords. The stemmer also normalizes words. For the run *mon3*, words were aggressively normalized within the stemmer. For all other runs, the numeric digits were treated as word delimiters, and the words were normalized using our own light normalizer.

For the run *mon4*, the words were stemmed using the automatically generated MT-based stemmer. The words were first normalized and then the stopwords removed.

For the runs, mon0, mon3, mon4, and BKYMON, 20 words were selected from the top-ranked 10 documents for query expansion; and for the runs, mon1 and mon2, 40 trigrams were selected from the top-ranked 10 documents for query expansion.

The increase in performance without query expansion is substantial, however, the difference remains small after query expansion.

7.3 Cross-language Retrieval Results

Our approach to cross-language retrieval was to translate the English topics into Arabic, and then search the translated Arabic topics against the Arabic documents. The source English topics were translated into Arabic using two online English-Arabic machine translation systems: *Ajeeb* and *Almisbar*, available at http://www.almisbar.com/.

			without expansion		with expansion	
run id	stemmer	index unit	recall	precision	recall	precision
mon0	NONE	word	4035	0.2365	4583	0.2872
mon1	NONE	trigram (without crossing)	3914	0.2398	4632	0.3239
mon2	NONE	trigram (with crossing)	4018	0.2479	4681	0.3178
mon3	Al-Stem stemmer	word	4500	0.2858	4864	0.3482
mon4	MT-based stemmer	word	4402	0.2948	4885	0.3348
BKYMON	Berkeley light stemmer	word	4543	0.3099	4952	0.3666

Table 4: Monolingual retrieval performances. The number of relevant documents for all 50 topics is 5909. Only the *title* and *description* fields were indexed.

We submitted three official cross-language runs: BKYCL1, BKYCL2, and BKYCL3. The BKYCL1 run was produced by merging the results of two English-to-Arabic retrieval runs: cl1 and cl2. The first run used the *Ajeeb* English-to-Arabic translations, and the second run used the *Almisbar* English-to-Arabic translations. For both intermediate runs, the words were stemmed using Berkeley's light stemmer after removing stopwords. For query expansion, 20 terms were selected from the top-ranked 10 documents. When two runs were merged topic by topic, the estimated probabilities of relevance were summed for the same documents. The merged list of documents was sorted by the combined estimated score of relevance, and the top-ranked 1000 documents per topic were kept to produce the official run BKYCL1. Only the *title* and *desc* fields in the topics were used to produce the BKYCL1 run. The average precision for run cl2 is 0.2782 with overall recall of 4823/5909. The average precision for run cl1 is 0.2962 with overall recall of 4441/5909.

The BKYCL2 run was produced by merging the results of three English-to-Arabic retrieval runs. The first two intermediate runs, cl1 and cl2, were the same two runs that were merged to produce BKYCL1 run. The third intermediate run, named cl3, was produced using the English-to-Arabic bilingual dictionary created from the U.N. English/Arabic parallel texts. The bilingual dictionary was provided as part of the standard translation resources for the cross-language track. Readers are referred to [7] for details on the construction of the bilingual dictionary. The English texts of the parallel corpus was stemmed using Porter stemmer, while the Arabic texts was stemmed using the Al-Stem stemmer which is part of the standard resources created for the cross-language track. Each entry in the English-to-Arabic bilingual dictionary consists of one stemmed English word and a list of stemmed Arabic words with the probabilities of translating the English word into the Arabic words. We translated the English topics into Arabic by looking up each English word after stemming using the same English porter stemmer in the English-to-Arabic bilingual dictionary, and keeping the two Arabic words of the highest translation probabilities. That is, the two most likely Arabic translations for each English word. Since only two Arabic translations were retained, the sum of their translation probabilities is at most one. In the case where the sum is less than one, the word translation probabilities were normalized so that the sum of the translation probabilities of the retained two Arabic words is one. The within-query term frequency of an English word is distributed to the retained Arabic words proportionally according their translation probabilities. For the cl3 run, we indexed the Arabic documents using the Al-Stem stemmer. The intermediate run cl3 was produced using the bilingual dictionary-translated topics. The average precision for run cl3 is 0.3072 with overall recall of 4826/5909. The official run BKYCL2 was produced by merging cl1, cl2, and cl3 runs. The estimated probabilities of relevance were summed during merging.

The official run BKYCL3 was produced again by merging two intermediate runs, cl3 and cl4. The cl3 run was described in the previous paragraph. The intermediate run cl4 was produced using the Ajeeb-translated topics like the cl1 run. The only difference is that the standard light stemmer, Al-Stem, was used in cl4. The average precision for run cl4 is 0.2710 with overall recall of 4350/5909.

The unofficial run, bkycl4, was produced like the official run BKYCL1 except that the MT-based stemmer was used here. The run bkycl4 was produced by merging cl5 and cl6. The cl5 run used the Ajeeb topic translations, while the cl6 run used the Almisbar topic translations. For both runs, the MT-based stemmer automatically constructed from Ajeeb-translated words was used. The average precision for run cl5 is 0.2733 with overall recall of 4118/5909, and the average precision for run cl6 is 0.2751 with overall recall of 4735/5909.

Table 5 shows the overall precision for the five runs. There are a total of 5,909 relevant documents for all 50 topics. The run BKYCL3 used standard resources only. Like the monolingual run, all cross-language runs were produced with query expansion in which 20 terms were selected from the top-ranked 10 documents after the initial search. Our best

Run ID	Type	Topic Fields	Recall	Precision	% of MONO
BKYMON	MONO	T,D	4952	0.3666	
BKYCL1	CLIR	T,D	4614	0.3000	81.83%
BKYCL2	CLIR	T,D	4874	0.3224	87.94%
BKYCL3	CLIR	T,D	4856	0.3089	84.26%
brkcl4	CLIR	T,D	4553	0.2857	77.93%

Table 5: Performances of the CLIR runs.

cross-language performance is 87.94% of the monolingual performance.

8 Conclusions

In summary, we performed one Arabic monolingual run and three English-Arabic cross-language retrieval runs, all being automatic. We took the approach of translating queries into document language using two machine translation systems. Our best cross-language retrieval run achieved 87.94% of the monolingual retrieval performance. We developed one MT-based Arabic stemmer and one light Arabic stemmer. The Berkeley light stemmer worked better than the automatically created MT-based stemmer. The experimental results show query expansion substantially improved the retrieval performance.

9 Acknowledgements

This research was supported by research grant number N66001-00-1-8911 (Mar 2000-Feb 2003) from the Defense Advanced Research Projects Agency (DARPA) Translingual Information Detection Extraction and Summarization (TIDES) program within the DARPA Information Technology Office.

References

[1] W. S. Cooper, A. Chen, and F. C. Gey. Full text retrieval based on probabilistic equations with coefficie nts fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.

[2] M. Zaidel D. Karp, Y. Schabes and D. Egedi. A freely available wide coverage morphological analyzer for english. In *Proceedings of COLING*, 1992.

[3] K. Darwish. *http://www.glue.umd.edu/∼kareem/research/*.

[4] Peter F. Abboud [et al.], editor. *Elementary modern standard Arabic*. Cambridge University Press, 1983.

[5] L. Larkey, L. Ballesteros, and M.E. Connell. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In *SIGIR'02, August 11-15, 2002, Tampere, Finland*, pages 275–282, 2002.

[6] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, pages 129–146, May–June 1976.

[7] Jinxi Xu, Alexander Fraser, and Ralph Weischedel. Trec 2001 cross-lingual retrieval at bbn. In E.M. Voorhees and D.K. Harman, editors, *The Tenth Text Retrieval Conference (TREC 2001)*, pages 68–77, May 2002.