Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval

Aitao Chen,* and Fredric Gey†
*School of Information Management and Systems
†UC Data Archive & Technical Assistance (UC DATA)
University of California at Berkeley, CA 94720, USA
aitao@sims.berkeley.edu, gey@ucdata.berkeley.edu

1 Introduction

In TREC-10 the Berkeley group participated only in the English-Arabic cross-language retrieval (CLIR) track. One Arabic monolingual run and four 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 bilingual dictionaries and machine translation software. The five official runs are named as BKYAAA1, BKYEAA1, BKYEAA2, BKYEAA3, and BKYEAA4. The BKYAAA1 is the Arabic monolingual run, and the rest are English-to-Arabic cross-language runs. The same logistic regression based document ranking algorithm without pseudo relevance feedback was applied in all five runs. We refer the readers to the paper in [1] for details.

2 Test Collection

The document collection used in TREC-10 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 25 English topics with Arabic and French translations. A topic has three tagged fields, *title*, *description*, and *narrative*. The newswire articles are encoded in UTF-8 format, while the topics are encoded in ASMO 708. The cross-language retrieval task is to search the English topics against the Arabic documents and present the retrieved documents in ranked order.

3 Preprocessing

Because the texts in the documents and topics are encoded in different schemes, we converted the documents and topics to Windows 1256 code. We created a stoplist of 1,131 words using two sources. First, we translated our English stopword list to Arabic using the *Ajeeb* online English-Arabic dictionary. Second, we garnered some of the stopwords from the Arabic-English glossary published in *Elementary Modern Standard Arabic*.

A consecutive sequence of Arabic letters, except for the punctuation marks, was recognized as a word. The words that are stopwords were removed when the documents and topics were indexed. The tokens were normalized by removing the initial letter و, the final letter ة, and the initial letters ال. In addition, the letters أ and إ were changed to the letter ا. The marks above or underneath the letter ا in أ, أ, إ, أ, أ, إ, ■, آ, if present, were also removed.

Arabic has a definite article, but no indefinite articles. The definite article 'al-' is sometimes attached to a word as a prefix. In addition to the singular and plural forms, Arabic also has a form called *dual* which is formed by adding the suffix -$\bar{a}n$. The plurals have regular (also called *sound*) and irregular (also called *broken*) forms. However, the irregular forms are very common, and it is difficult to predict execept that there exist several commonly occurring patterns. The regular plural is formed by adding the suffix $-\bar{u}n$ for the masculine and $-\bar{a}t$ for the feminine form. In Arabic, the adjectives modifying plural nouns also have to be in plural form. Arabic has only two genders, masculine and feminine. The feminine is formed from masculine nouns and adjectives by adding the suffix $-a$.

Since neither of the authors really knows Arabic, it is difficult to write a linguistically motivated Arabic stemmer. One of us learned a little Arabic during the course of participating in this English-Arabic cross-language track and wrote a simple stemmer to remove the definite article $al-$ from the definite nouns, the suffix $\bar{a}n$ from nouns in *dual* form, $-\bar{u}n$ from masculine plural nouns, $-\bar{a}t$ from feminine plural nouns, and suffix $-a$ from feminine noun. Here we assumed that the categories (i.e. part of speech) of words are known. Unfortunately we do not have the part of speech for each word in the collection, nor do we have a part of speech tagger to tag the words. So we cannot simply apply the rules described here. We took a data-driven (i.e, corpus-based) approach to stemming. First we collected all the words in their original form from the document collection. Then we applied each of the rules to the list of Arabic words. For example, to remove the suffix $-\bar{u}n$ from masculine plural nouns, we remove the suffix $-\bar{u}n$ from a word if both the word with the suffix $-\bar{u}n$ and the word without the suffix $-\bar{u}n$ occur in the document collection. Because a word ends with the letters $-\bar{u}n$ is not necessary a masculine plural noun, it is possible to remove the suffix $-\bar{u}n$ from a word incorrectly. The same mistake may also be committed in applying other stemming rules. Our stemming, despite being simple and imperfect, brought an improvement of 9.4% in overall precision for the Arabic monolingual retrieval over the baseline run without stemming.

4 Query Translation

Our approach to cross-language retrieval is to translate the English topics into Arabic, and then search the translated Arabic topics against the Arabic documents.

4.1 Translation Resources

Two online English-Arabic bilingual dictionaries and one online machine translation system were utilized in translating the English topics into Arabic in our cross-language retrieval experiments. The first online English-Arabic dictionary is publicly accessible at http://dictionary.ajeeb.com/en.htm. We will refer to this dictionary as the *Ajeeb* dictionary. The English-Arabic machine translation system is also available from http://dictionary.ajeeb.com/en.htm. The second one is the *Ectaco* dictionary publicly available at http://www.get-together.net/.

4.2 Translation Term Selection

Each word in the English topics was submitted to both English-Arabic online dictionaries. The translations from both dictionaries were merged to form the translation for the English word. To use the Ectaco Arabic-English dictionary, one has to enter nouns in the singular form, verbs in the infinitive form, and adjectives in their positive form. Before we submitted each word as a query to the Ectaco online dictionary, we normalized the English words using an English morphological analyzer [2]. The Ajeeb Arabic-English dictionary can take un-normalized words as input. All the Arabic translations for an English word were sorted and ranked by their occurrence frequency in the Arabic document collection. The top-ranked Arabic translations, but not more than five, an English word were retained as the translation of the English word.

4.3 Translation Term Weighting

After term selection, the term frequency of a source English word in the original query was distributed among the Arabic translations of the English word according to their occurrence frequency in the Arabic collection. The weight assigned to an Arabic translation is proportional to its occurrence frequency in the document collection. That is,

$$qtf_{ai} = qtf_e * \frac{ctf_i}{\sum_{j=1}^{n} ctf_j} \quad (1)$$

where qtf_e is the within-query term frequency of the English word e, ctf_i is the within-collection term frequency of the ith Arabic translation, qtf_{ai} is the weight assigned to the ith Arabic translation, and n is the number of translations retained for the source English word. For the word *education*, the five translations

	Arabic Translation	Frequency in Collection	Translation Weight
1	بحث	15,183	0.35
2	دراس	11,185	0.25
3	ثقاف	6,484	0.15
4	خبر	5,527	0.13
5	تعليم	5,500	0.13

Table 1. The top-ranked five Arabic translations for *education*.

that occur most frequently in the document collection are shown in the second column in table 1. Column 3 in the table shows the number of times each Arabic translation is found in the Arabic collection, and the last column the weight assigned to each of the Arabic translations of *education*, assuming *education* occurs only once in the original English topics. Otherwise, the translation weight is multiplied by the term frequency of *education* in the original query.

5 Experimental Results

The official runs we submitted are summarized in table 2. The BKYAAA1 is our only Arabic monolingual run in which all three topic fields were indexed, stopwords removed from both topics and documents, and remaining words stemmed. The BKYEAA2 run used only the machine translation to translate the English topics to Arabic, while the BKYEAA3 used the online dictionaries only to translate the English topics into Arabic. For the other two runs, BKYEAA1 and BKYEAA4, the English topics were separately translated into Arabic using the machine translation system and the bilingual dictionaries first, then their translations were merged before being searched against the Arabic document collection. The only difference between BKYEAA4 and BKYEAA1 is that the former indexed only the title and description fields, where as the latter indexed all three topic fields.

Table 3 shows the overall precision for the five runs. There are a total of 4,122 relevant documents for all 25 topics. As mentioned above, all five runs were performed without pseudo relevance feedback. Our best cross-language performance is 85.68% of the monolingual performance. The queries translated from the combined online dictionaries substantially outperformed those translated from the machine translation system. We believe that the superior performance of the combined dictionaries could be attributed in part to the fact that up to five translation terms from the online dictionaries were retained for the source words while the machine translation system retained only one translation for each source word.

Run ID	Type	Topic Fields	Translation Resources
BKYAAA1	Arabic Monolingual	Title,Description,Narrative	
BKYEAA1	English-to-Arabic	Title,Description,Narrative	Dictionaries and MT
BKYEAA2	English-to-Arabic	Title,Description,Narrative	MT
BKYEAA3	English-to-Arabic	Title,Description,Narrative	Dictionaries
BKYEAA4	English-to-Arabic	Title,Description	Dictionaries and MT

Table 2. Summary of official runs.

recall level	BRKAAA1 (MONO)	BRKEAA1 (CLIR)	BRKEAA2 (CLIR)	BRKEAA3 (CLIR)	BKYEAA4 (CLIR)
at 0.0	0.8432	0.7803	0.7133	0.7052	0.7372
at 0.1	0.6174	0.5250	0.4374	0.5119	0.4901
at 0.2	0.4582	0.3970	0.3229	0.4418	0.3807
at 0.3	0.3716	0.3241	0.2752	0.3463	0.2967
at 0.4	0.3021	0.2627	0.2265	0.2870	0.2493
at 0.5	0.2487	0.1967	0.1780	0.2257	0.2026
at 0.6	0.1959	0.1309	0.1290	0.1490	0.1437
at 0.7	0.1604	0.0945	0.0861	0.1206	0.1134
at 0.8	0.1200	0.0620	0.0588	0.0915	0.0874
at 0.9	0.0701	0.0121	0.0170	0.0240	0.0200
at 1.0	0.0141	0.0014	0.0015	0.0141	0.0200
average precision	0.2877	0.2337	0.2006	0.2465	0.2316
relevant retrieved	2,393	2,579	2,485	2,490	2,300
% of mono		81.23%	69.73%	85.68%	80.50%

Table 3. Evaluation results for one Arabic monolingual run and three English to Arabic cross-language retrieval runs.

A number of additional experimental runs were performed and evaluated locally to show the effect of various aspect of preprocessing on the retrieval performance. We broke down the preprocessing of the texts into three steps: stopwords removal, word normalization, and word stemming. Table 4 presents the overall precision and recall by incrementally adding more features into the preprocessing of the Arabic texts. The overall precision was .1581 when no preprocessing was performed at all. That is, no words were removed from indexing, words were not normalized and stemmed. When stopwords were removed from indexing, the overall precision increased to .2046, and when words were normalized as described above the overall precision was substantially improved. Further improvement was shown by stemming the words even though our stemmer was rather simple. Many more possible word form changes were not considered at all in our stemmer. The very simple normalization of words brought 28.54% improvement in overall precision over the run without word normalization. The results presented in table 4 leads us to believe that further gain in overall precision could be achieved by using a more sophisticated Arabic stemmer or morphological analyzer. All three topic fields were indexed in the runs shown in table 4. Our official monolingual run, BKYAAA1, included all three steps in preprocessing. The overall recall for our official monolingual run

was only 58.05%. Besides applying a more sophisticated Arabic stemmer, we believe that pseudo relevance feedback should also improve both overall recall and overall precision.

recall	stoplist	normalization	stemming	precision	recall
baseline	-	-	-	0.1581	1594/4122
mono1	+	-	-	0.2046	1930/4122
mono2	+	+	-	0.2630	2333/4122
BKYAAA1	+	+	+	0.2877	2393/4122

Table 4. Arabic monolingual retrieval performance.

For the runs, BKYEAA1 and BKYEAA4, the separately translated topics using online dictionaries and online machine translation system were merged before being searched against the Arabic collection. We also experimented with linearly combining the ranked lists produced in searching the translated topics separately against the Arabic documents. That is, we first ran the dictionary-translated topics against the Arabic documents, and the machine translation system-translated topics against the Arabic documents. Then we merged the two ranked lists by averaging the probabilities of relevance. The overall precision for the long queries increased from .2337 of BKYEAA1 to .2552, a 9.20% improvement.

6 Conclusions

In summary, we performed four English-Arabic cross-language retrieval runs and one Arabic monolingual run, all being automatic. We took the approach of translating queries into document language using two online dictionaries and one machine translation system. Our best cross-language retrieval run achieved 85.68% of the monolingual run. Furthermore, our cross-language run using online bilingual dictionaries substantially outperformed the run using an online machine translation system. All of our runs had low overall recall, which we believe could be in part attributed to our failure to conflate the various forms of the words to their stems. Even though the preprocessing was quite simple, it substantially improved the overall precision and recall over the baseline run without any preprocessing at all. We believe that further improvement could be achieved by applying a more sophisticated Arabic stemmer and pseudo relevance feedback.

7 Acknowledgements

This work was supported by DARPA (Department of Defense Advanced Research Projects Agency) under research contract N66001-97-C-8541, AO-F477.

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.