On Arabic-English Cross-Language Information Retrieval: *A Machine Translation Approach*

Mohammed Aljlayl, Ophir Frieder, & David Grossman
Information Retrieval Laboratory
Illinois Institute of Technology
{aljlayl, ophir, dagr}@ir.iit.edu

Abstract

A Machine Translation (MT) system is an automatic process that translates from one human language to another language by using context information. We evaluate the use of an MT-based approach for query translation in an Arabic-English Cross-Language Information Retrieval (CLIR) system. We empirically evaluate the use of an MT-based approach for query translation in an Arabic-English CLIR system using the TREC-7 and TREC-9 topics and collections. The effect of query length on the performance of the machine translation is also investigated to explore how much context is actually required for successful MT processing.

1. Introduction

Cross-Language Information Retrieval (CLIR) is the retrieval of relevant documents based on queries expressed by a human in a given natural language against a collection on which the documents are expressed in another language. Arabic-English CLIR, therefore, focuses on the retrieval of documents based on queries formulated by a user in the Arabic language and the documents are in the English language. There are three mainstream general approaches to CLIR: machine translation (MT), comparable or parallel corpus, and machine-readable dictionary. We focus on Arabic-English machine translation, and in particular, on the evaluation of the ALKAFI Arabic-English MT system for Arabic-English CLIR. We investigated strictly the effectiveness (accuracy) and not its efficiency (speed of processing) of the MT-based Arabic-English CLIR. High speed processing is meaningful only when high accuracy is obtained. Currently, the state of the art does not support highly accurate Arabic-English CLIR. The experiments are evaluated using short, medium, and long queries of TREC-7 and TREC-9 topics and collections.

Arabic language is one of the six official languages of the United Nations. According to Egyptian Demographic Center, it is the mother tongue of about 300 million people [8]. The orientation of writing is from right-to-left, and the Arabic alphabet consists of 28 letters. As discussed in [15], the Arabic alphabet can be extended to ninety elements by writing additional shapes, marks, and vowels. Most Arabic words are morphologically derived from a list of roots; it can be tri, quad, or pent-literal. Most of these roots are three constants.

Arabic words are classified into three main parts of speech, nouns (adjectives, and adverbs), verbs, and particles. In formal writing, Arabic sentences are delimited by commas and periods as in English, for instance.

The remainder of this paper is organized as follows. Initially, we review the prior related art, namely the work on Arabic information retrieval and on CLIR. We continue by presenting our experimental framework and our experimental findings. Our conclusions are summarized in the final section.

2. Prior work

For brevity of presentation, we assume that the reader is familiar with the general domain of information retrieval. For additional detail, see [9].

El-Dessouki, et al. [7] developed an expert system to recognize Arabic sentences. The authors used learning by example mechanism to implement the syntactic analyzer. Beesley [5] developed a finite-state morphological analyzer of written standard Arabic.

In CLIR, either the documents or the queries are translated. Since the document translation is computationally expensive [10], most efforts focus on accurate translation of the query. There are three main approaches to CLIR: machine translation, bilingual dictionary, and parallel or comparable corpora methods.

Dictionary-based methods [1,2,3,4] perform query translation by looking up terms on a bilingual dictionary

and generating a target language query by considering some or all of the translations. In investigating Spanish-English CLIR, Ballesteros and Croft [2] introduced the notion of pre-translation, post-translation, and combined approaches and yielded improvement over transnational dictionary-based approaches. They later investigated the effect of phrasal translation in improving the effectiveness [3]. A co-occurrence method was used to resolve the ambiguity. An approach to reduce ambiguity of phrasal and term translation was eventually developed [4].

In corpus-based methods [12, 14], queries are translated on the basis of multilingual terms extracted from parallel or comparable document collections. In parallel corpora, the pair or set of documents are identical but in different languages. A comparable corpus contains similar documents in different languages, i.e., the pair documents are conceptually similar [14]. A corpus-based fully automatic method was proposed in [12]. It is known as Cross-Language Latent Semantic Indexing (CL-LSI). CL-LSI is a method for CLIR in which no query translation is required.

The ultimate goal of CLIR machine translation (MT) systems is to translate queries from one language to another by using a context. Many factors contribute to the difficulties of machine translation, including words with multiple meanings, sentences with multiple grammatical structures, uncertainty about what a pronoun refers to, and other problems of grammar. The hope of CLIR machine translation researchers is to take the advantage of the extensive research on MT and the availability of the commercial products to support retrieval.

Many researchers criticize MT-based CLIR approach. The reasons behind their criticisms mostly stem from the fact that the current translation quality of MT is poor. In particular, typical search terms lack the context necessary for MT systems to correctly perform proper syntactic and semantic analysis of the source text. Another reason is that MT systems are expensive to develop and their application degrades the retrieval efficiency (run time performance) due to the lengthy processing times associated with the linguistic analysis.

Hull and Grefenstette [10] stated that current MT systems, in the setting of general language translation, are less than satisfactory for CLIR. A study [13] compared the retrieval effectiveness of the French-English CLIR using SYSTRAN machine translation system with the effectiveness of their EMIR dictionary-based query translation. They determined that EMIR was more effective than their MT-based query translation technique.

Other researchers, in contrast, showed that machine translation approaches could achieve reasonable effectiveness. Jones, et al. [11], showed that full disambiguation by MT system outperforms dictionary methods that include many terms as candidates in the query. Also, many participants in the TREC-8 CLIR track [6] concluded that MT-based CLIR is an effective strategy.

Our other CLIR efforts [1] focus on Arabic-English CLIR using predominantly a machine-readable dictionary approach. Using a two-pass method, we developed a retrieval strategy that statistically improved over prior one-pass dictionary approaches. Only cursory evaluation of a machine translation approach was conducted. Given the uncertainty of finding regarding the accuracy of MT-based CLIR approaches, here, we evaluate, in detail, a machine translation based approach for Arabic-English CLIR.

3. Experimental approach

Presently, no benchmark data are available for Arabic-English CLIR. To provide a means to compare our efforts with future Arabic-English CLIR efforts, we used a readily available English benchmark document collections and provide our Arabic queries, a translation of the National Institute of Science and Technology, Text Retrieval Conference (TREC) queries on our web site at www.ir.iit.edu. Briefly, TREC has three distinct parts: the documents, the topics, and the relevance judgments. We used two collections. The first, the 2.1 GB TIPSTER Disks 4 and 5, and the second, the 10 GB TREC-9 Web collection, consist of roughly 500,000 and 1,700,000 documents, respectively. For queries, we manually translated the TREC-7 (topics 351-400) and TREC-9 (topics 451-500) queries to Arabic. We used these 100 translated versions as our original Arabic queries issued against the TREC English collection. The Arabic queries were translated back to English using the ALKAFI MT system. Indexing is done using the Porter and K-stem algorithms after eliminating the stop-words. Similarity, querying is done after stemming and eliminating the stop-words of the translated target English queries.

The ALKAFI Arabic-English MT system is a commercial system developed by CIMOS Corporation and is the first Arabic to English machine translation system. The system attempts to analyze terms in context and then builds semantic links. Then, the English text is generated by the transfer technique according to English language grammar.

The TREC queries (or topics in the TREC vernacular) consist of three fields: title, description, and narrative. The title is considered short; it consists of one, two or three concept terms. In Table 1, we illustrate an example of the original Arabic title and its translation. The description field is of medium length; it consists of one or two sentences. In Table 2, we provide an example of the description field and its translation. The longest part is the narrative field; in Table 3, we show an example of the narrative field and its translation using the ALKAFI MT system. To measure the effectiveness of an MT system for

CLIR, we experimented using all three-query types to determine the effects of query length on the performance of the MT-based method for CLIR.

Arabic query	Translated English Query
مُعَدَّات التشفير التصديريّة	The export equipments of the encryption

Table 1. The title of the original Arabic and the translated English query

Arabic query	Translated English Query
عرِّف الوثائق التي تَناقش اهتمامات الولايات المتّحدة بشأن المصدَّر من مُعَدَّات التشفير.	Define the documentations which the debate of the United States concerns regarding exported from equipments of the encryption.

Table 2. The description of the original Arabic and the translated English query

Arabic query	Translated English Query
الوثائق التي تذكر اسم الشركة أو المجموعة التي تُنتج مُعَدَّات التشفير فقط ، و لكن لا تذكر بالمصدَّر و/ أو الاستثمار التجاريّ لمُعَدَّات التشفير ليست ذات صلة. الوثائق التي تشير إلى الوصول الحكوميّ لنُظم التشفير لغايات ضدّ نشاطات الاستخبارات أو نشاطات مكافحة الجريمة ، تكون ذات صلة.	The documentations which she remembers the name of the company or the group which produces encryption equipments of the encryption only , but not you remember by exported and / or the commercial investment equipments of the encryption have no relevancy. The documentations which she points out the governmental arrival for organisms of the encryption the purposes of a briskness opposite have the secret services or the briskness of the crime struggle , you are a relevancy.

Table 3. The narrative of the original Arabic and the translated English query

4. Results

We use three performance measures. The first uses the recall-precision scores at 11 standard points. In CLIR systems, given the expenses of the translation, a user is most likely to be interested in only the top few retrieved Web pages. Thus, we provide measures for the top *n* retrieved documents. We also provide the overall average of precision of each run.

We evaluate the effects of the MT system in Arabic-English CLIR. As described earlier, we used both the TREC-7 and TREC-9 topics and collections. For TREC-7, as shown in Table 4, the machine translation achieved 61.8%, 64.7%, and 60.2% for title, description, and narrative fields, respectively. The 11-point average recall-precision for TREC-7 topics is shown in Figures 1, 2, and 3 for the title, description, and narrative fields, respectively. As shown, the MT-based approach on description is more effective than title and narrative. In each figure, we also illustrate the "ideal" system score that is represented by the monolingual query. At the higher precision-lower recall levels, the difference is even more noticeable. Since it is unrealistic to expect the user to read many retrieved documents that are expressed in a language other than the user's native language, the higher precision region is of greater interest.

The degraded effectiveness of the machine translation on title queries is due to the fact that the ALKAFI machine translation system is designed to perform best on well-formed sentences or at least on a sequence of words that form a context. However, the titles of topics 351-400 are all three words or less; thus, no substantive context is formed.

For the narrative run results shown in Figure 3, the MT system is unable to preserve its accuracy when extra, potentially noise, terms are presented in the source query. The greater the number of source query terms, except for, of course, keywords or words of high query disambiguation content, the greater is the performance degradation of a CLIR system. These additional, potentially noise, terms do not provide a strong basis of the source query. The ALKAFI MT system, however, is still capable of maintaining 60.2% of the monolingual retrieval. At the higher precision-lower recall levels, the narrative run is more effective than the title. At the higher recall level (up to 0.8), the title run is more effective than the narrative run. As measured by average precision, there is a slight difference between the narrative and the title runs. It is not surprising that the narrative run is strictly worse in accuracy then the descriptive run since the MT system achieves its best performance on the fewest sequence of words that still provides a full context.

	Original	MT	% Monolingual
Title	0.1733	0.1071	61.8
Description	0.1838	0.1190	64.7
Narrative	0.1522	0.0917	60.2

Table 4. Average precision of the title, description and narrative fields of topics 351-400

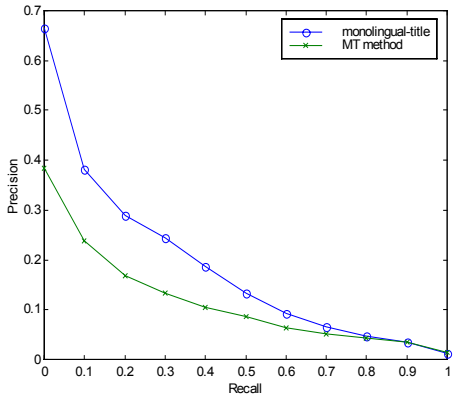

Figure 1. Average precision and recall of original Arabic query titles of topics 351-400 and MT method

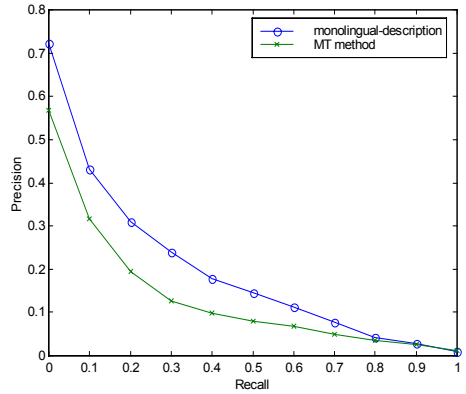

Figure 2. Average precision and recall of the descriptions of the original Arabic query of topics 351-400 and MT method

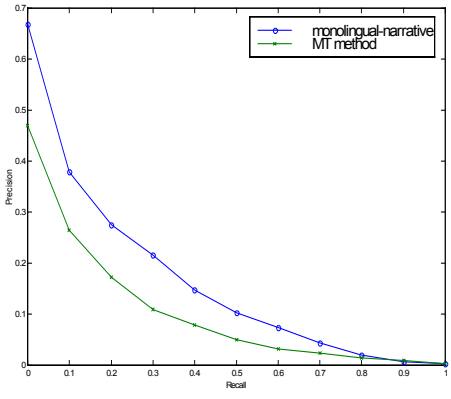

Figure 3. Average precision and recall of the narratives of the original Arabic query of topics 351-400 and MT method

Precision at	Original _ Title	MT _Title	Original _ Desc	MT _ Desc	Original _Nar	MT_ Nar
5	0.4240	0.2200	0.4880	0.3560	0.4360	0.3000
10	0.3800	0.1960	0.4160	0.2920	0.3780	0.2620
15	0.3413	0.1907	0.3787	0.2573	0.3347	0.2413
20	0.3170	0.1940	0.3420	0.2340	0.3130	0.2180
30	0.2700	0.1667	0.3020	0.2040	0.2700	0.1793
100	0.1746	0.1162	0.1780	0.1206	0.1656	0.1066
200	0.1226	0.0825	0.1245	0.0833	0.1124	0.0742
500	0.0731	0.0497	0.0721	0.0494	0.0624	0.0432
1000	0.0459	0.0305	0.0455	0.0315	0.0389	0.0270

Table 5. Precision at 1000 documents retrieved of topics 351-400

In Table 5, we illustrate the results up to 1000 documents retrieved for TREC-7 queries 351-400. As shown, the description run consistently outperforms both the title and the narrative runs.

In Table 6, we illustrate the average precision of TREC-9 topics. Our CLIR approach using the ALKAFI MT system achieves 58.4%, 57.1%, and 53.4% for title, description, and narrative fields, respectively. The 11-point average recall-precision for TREC-9 topics is shown in Figures 4, 5, and 6 for the title, description, and narrative fields, respectively. Again, the "ideal" monolingual run is likewise illustrated in each figure.

	Original	MT	% Monolingual
Title	0.1305	0.0763	58.4
Description	0.1857	0.1061	57.1
Narrative	0.1678	0.0897	53.4

Table 6. Average precision of the title, description and narrative fields of queries topics 451-500

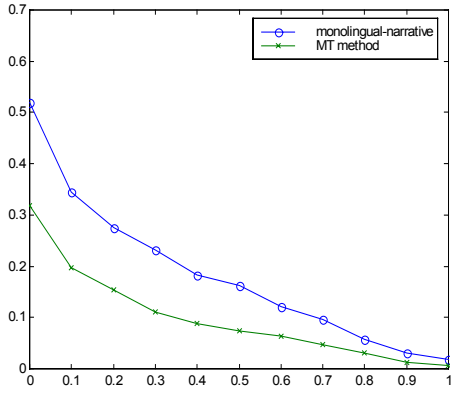

Figure 4. Average precision and recall of the titles of the original Arabic query of topics 451-500 and MT method

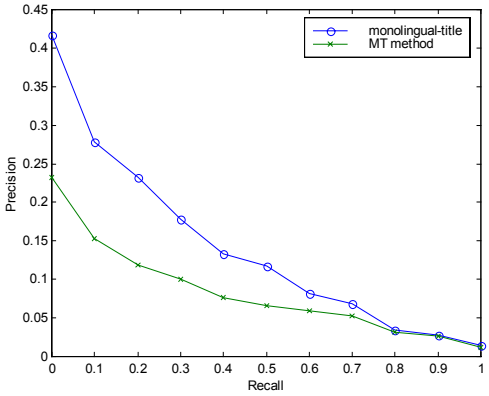

Figure 5. Average precision and recall of the descriptions of the original Arabic query of topics 451-500 and MT method

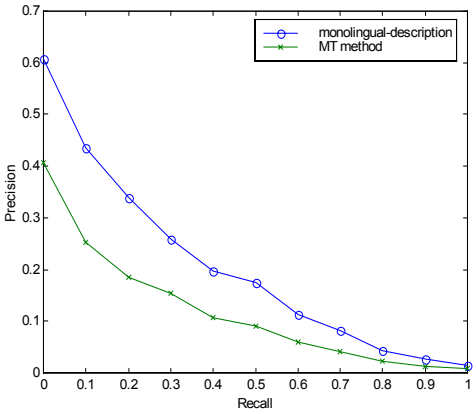

Figure 6. Average precision and recall of the narratives of the original Arabic query of topics 451-500 and MT method

In Tables 7, we illustrate the results up to 1000 documents retrieved for TREC-9. As shown, again, the description run consistently outperforms both the title and narrative runs. However, as shown in Table 6, the percentage of degradation of the title run from the "ideal" monolingual title run is less that that of the descriptive run. This result is seemingly inconsistent with the results

Precision at	Original_Title	MT_Title	Original_Desc	MT_Desc	Original_Nar	MT_Nar
5	0.2227	0.1222	0.3560	0.2360	0.2600	0.1880
10	0.1886	0.1111	0.2740	0.1980	0.2460	0.1580
15	0.1712	0.1037	0.2627	0.1800	0.2200	0.1373
20	0.1545	0.0944	0.2330	0.1690	0.1990	0.1240
30	0.1348	0.0917	0.2167	0.1447	0.1700	0.1053
100	0.0834	0.0633	0.1260	0.0898	0.1064	0.0642
200	0.0581	0.0457	0.0894	0.0664	0.0713	0.0433
500	0.0316	0.0276	0.0514	0.0373	0.0384	0.0253
1000	0.0184	0.0165	0.0314	0.0235	0.0229	0.0156

Table 7. Precision at 1000 documents retrieved of topics 451-500

obtained for the machine translation on titles run for queries 351-500 as presented in Table 4.

The reason behind this seeming contradiction in accuracy performance is that the titles of query 451-500 are actually quite long. The average title query length for queries 351-400 is 2.72 word per query while the average length for queries 451-500 is 3.46 words. This 27% difference in query length was sufficient to provide our MT system with the possibility to form a proper context for many more queries in the TREC-9 query set as compared to the TREC-7 query set. This is especially so considering that the TREC-9 query set had 16 queries with 4 or more words as compared to the only 6 queries of similar length in the TREC-7 query set. For example, the title of the query number 482 is:

أين يُمكِن أن أجِدْ معدّلات النّموّ لشجرة الصنوبر ؟

The translated query using ALKAFI MT system is:

"Where is he possible that I find the rates of the growth for the tree of the pine? _

This query provides a full context allowing the ALKAFI machine translation to produce the most accurate translation. Adding more contexts to that query does not help the MT system to provide better translation accuracy. Finally, for completeness, we provide a brief overview of efficiency results. In Table 8, we summarize the efficiency (run time performance) of the ALKAFI MT system to translate the titles, descriptions and narratives fields of topics TREC-7 and TREC-9.

	Title	Description	Narrative
TREC-7	6	18	51
TREC-9	7	17	40

Table 8. The running time of the MT system measured in seconds

In Table 9, we summarize the efficiency (running time performance) of AIRE search engine to the run the translated titles, descriptions and narratives fields of topics TREC-7 and TREC-9.

	Title	Description	Narrative
TREC-7	445.655	1972.256	6143.799
TREC-9	3752.002	12630.42	28708.65

Table 9. Total time to run queries measured in seconds

The narrative fields as described in Tables 4 and 6, which represent the long queries, are not effective compared to the description fields, which represents the medium length queries. According to theses findings, the less terms provided in the original query that form a context to obtain unambiguous representation, the better running time as well as the better retrieval effectiveness.

As presented in Tables 8 and 9, the total running time for the description and narrative runs of TREC-7 is 6194.799 and 1990.256 seconds, respectively. The difference is 4204.543 seconds. In fact, the difference of the running time degrades the performance of our CLIR system without any improvement on the effectiveness. These findings are consistent with TREC-9 topics and collection as presented in Tables 8 and 9.

The difference between the title and description runs of TREC-7 is 1538.601 seconds. Accordingly, the achieved performance of the description run is more effective than the title run. Thus, choosing few terms that form a full context achieves better accuracy at the expense of efficiency, a trade-off whose merits are application dependent. Similar findings exist for the TREC-9 queries.

5. Conclusions

We evaluated the effectiveness of an MT-based Arabic-English CLIR by using the ALKAFI system and two standard TREC collections and topics. To explore the effects of the context to the quality of translation, we experimented with various query lengths. The experimental results indicate that the less source terms that are needed to form a context, the better is the retrieval accuracy and efficiency. However, the problem of semantics is perennial. Without some level of semantic representation, MT systems are unable to achieve high quality translation, because they cannot differentiate between cases that are lexically and syntactically ambiguous. A possible extension to our work is to expand the original source query using pseudo-relevance feedback to emphasize the context of the source query. Accordingly, a well-formed source query makes the MT system able to provide its best accuracy.

6. References

[1] Aljlayl, M. and Frieder, O., "Effective Arabic-English Cross-Language Information Retrieval via Machine Readable Dictionaries and Machine Translation", *ACM Tenth Conference on Information and Knowledge Management*, Atlanta, Georgia, November 2001.

[2] Ballesteros, L., and Croft, B., "Dictionary Methods for Cross-Lingual Information Retrieval", *In the Proceeding of the 7th International DEXA Conference on Database and Expert Systems Applications*, Pages 791-801. 1996

[3] Ballesteros, L., and Croft, B., "Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval", *SIGIR* 1997, 84-91.

[4] Ballesteros, L., and Croft, B.,"Resolving Ambiguity for Cross-Language Retrieval", *SIGIR*, 1998, Pages 64-71.

[5] Beesley, K., "Arabic Morphological Analysis on the Internet", *Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing, Cambridge*, 1998.

[6] Braschler, M., Peters, C. and Schuable, P.,"Cross-Language Information Retrieval (CLIR) Track Overview", *TREC-8* Proceedings. 1998.

[7] El-Dessouki, A., El-Dessouki, O., Nazif, A. and Ahmad, M., "An ATN Approach for Understanding Arab Sentences", *In Proceedings of the 11th National Computer Conference and Exhibition*, Dhahran, Saudi Arabia, 1989, Pages 762-773

[8] Egyptian Demographic Center, 2000.http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm

[9] Grossman, D. and Frieder, O., Information Retrieval: Algorithms and Heuristics, Kluwer Academic Publishers, ISBN 0-7923-8271-4, 1998.

[10] Hull, D. and Grefenstette, G., "Querying across languages. A dictionary-based approach to multilingual information retrieval*", In proceedings of the 19th Annual international ACM SIGIR*, 1996, Zurich, Switzerland, 49-57.

[11] Jones, G., Sakai, T., Collier, N., Kumano, K., Sumita, K., "A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval", *SIGIR 1999*: 269-270.

[12] Landauer, T. K., and Littman, M. L.,"Full automatic cross-language document retrieval using latent semantic indexing", *In Proceeding of the 6th conference of UW center for New OED and Text Research*, 1990, Pages 31-38.

[13] Radwan, K., Fluhr, C.,"Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval", *In Fourth Annual Symposium on document analysis and information retrieval*, 1995, Pages 121-136.

[14] Sheridan, P. and Ballerini, J.P., "Experiments in multilingual information retrieval using the SPIDER system", *In Proceedings of the 19th Annual International ACM SIGIR*. 1996, 58-65.

[15] Tayli, M., and Al-Salamah, A., "Building Bilingual Microcomputer Systems", *In Communications of the ACM*, Vol. 33, No.5,1990, Pages 495-505.