The Affect of Machine Translation on the Performance of Arabic-English QA System

Azzah Al-Maskari
Dept. of Information Studies,
University of Sheffield,
Sheffield, S10 2TN, UK
lip05aaa@shef.ac.uk

Mark Sanderson
Dept. of Information Studies,
University of Sheffield,
Sheffield, S10 2TN, UK
m.sanderson@shef.ac.uk

Abstract

The aim of this paper is to investigate how much the effectiveness of a Question Answering (QA) system was affected by the performance of Machine Translation (MT) based question translation. Nearly 200 questions were selected from TREC QA tracks and ran through a question answering system. It was able to answer 42.6% of the questions correctly in a monolingual run. These questions were then translated manually from English into Arabic and back into English using an MT system, and then re-applied to the QA system. The system was able to answer 10.2% of the translated questions. An analysis of what sort of translation error affected which questions was conducted, concluding that factoid type questions are less prone to translation error than others.

1 Introduction

Increased availability of on-line text in languages other than English and increased multi-national collaboration have motivated research in Cross-Language Information Retrieval (CLIR). The goal of CLIR is to help searchers find relevant documents when their query terms are chosen from a language different from the language in which the documents are written. Multilinguality has been recognized as an important issue for the future of QA (Burger et al. 2001). The multilingual QA task was introduced for the first time in the Cross-Language Evaluation Forum CLEF-2003.

According to the Global Reach web site (2004), shown in Figure 1, it could be estimated that an English speaker has access to around 23 times more digital documents than an Arabic speaker. One can conclude from the given information shown in the Figure that cross-language is potentially very useful when the required information is not available in the users' language.

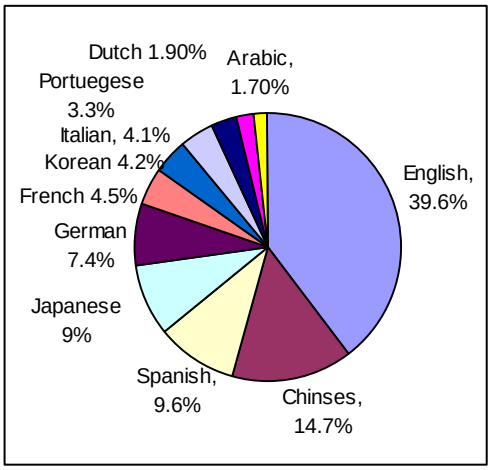

Figure 1: Online language Population (March, 2004)

The goal of a QA system is to find answers to questions in a large collection of documents. The overall accuracy of a QA system is directly affected by its ability to correctly analyze the questions it receives as input, a Cross Language Question Answering (CLQA) system will be sensitive to any errors introduced during question translation. Many researchers criticize the MT-based CLIR approach. The reason for their criticism mostly stem from the fact that the current translation quality of MT is poor, in addition MT system are expensive to develop and their application degrades the retrieval efficiency due to the cost of the linguistic analysis.

This paper investigates the extent to which MT error affects QA accuracy. It is divided as follows: in section 2 relevant previous work on cross-language retrieval is described, section 3 explains the experimental approach which includes the procedure and systems employed, it also discuss the results obtained, section 4 draws conclusions and future research on what improvements need to be done for MT systems.

2 Related Research

CLIR is an active area, extensive research on CLIR and the effects of MT on QA systems' retrieval effectiveness has been conducted. Lin and Mitamua (2004) point out that the quality of translation is fully dependent upon the MT system employed.

Perret (2004) proposed a question answering system designed to search French documents in response to French queries. He used automatic translation resources to translate the original queries from (Dutch, German, Italian, Portuguese, Spanish, English and Bulgarian) and reports the performance level in the monolingual task was 24.5% dropping to 17% in the bilingual task. A similar experiment was conducted by Plamondon and Foster (2003) on TREC questions and measured a drop of 44%, and in another experiment using Babelfish, the performance dropped even more, 53%. They believe that CLEF questions were easier to process because they did not include definition questions, which are harder to translate. Furthermore, Plamondon and Foster (2004) compare the cross language version of their Quantum QA system with the monolingual English version on CLEF questions and note the performance of a cross-language system (French questions and English documents) was 28% lower than the monolingual system using IBM1 translation.

Tanev et al. (2004) note that DIOGENE system, which relies on the Multi-WordNet, performs 15% better in the monolingual (Italian-Italian) than cross-language task (Italian-English). In Magnini et al.'s (2004) report for the year 2003, the average performance of correct answers on monolingual tasks was 41% and 25% in the bilingual tasks. In addition in the year 2004, the average accuracy in the monolingual tasks was 23.7% and 14.7% in bilingual tasks.

As elucidated above, much research has been conducted to evaluate the effectiveness of QA systems in a cross language platform by employing MT systems to translate the queries from the source language to the target language. However, most of them are focused on European language pairs. To our knowledge, only one past example of research has investigated the performance of a cross-language Arabic-English QA system Rosso et al (2005). The QA system used by Rosso et al (2005) is based on a system reported in Del Castillo (2004). Their experiment was carried out using the question corpus of the CLEF-2003 competition. They used questions in English and compared the answers with those obtained after the translation back into English from an Arabic question corpus which was manually translated. For the Arabic-English translation process, an automatic machine translator, the TARJIM Arabic-English machine translation system, was used. Rosso el al reported a decrease of QA accuracy by more than 30% which was caused by the translation process.

Work in the Rosso paper was limited to a single QA and MT system and also did not analyze types of errors or how those errors affected different types of QA questions. Therefore, it was decided to conduct further research on MT systems and its affect on the performance in QA systems. This paper presents an extension on the previous mentioned study, but with more diverse ranges of TREC data set using different QA system and different MT system.

3 Experimental Approach

To run this experiment, 199 questions were randomly compiled from the TREC QA track, namely from TREC-8, TREC-9, TREC-11, TREC-2003 and TREC-2004, to be run through AnswerFinder, the results of which are discussed in section 3.1. The selected 199 English TREC questions were translated into Arabic by one of the authors (who is an Arabic speaker), and then fed into Systran to translate them into English. The analysis of translation is discussed in detail in section 3.2.

3.1 Performance of AnswerFinder

The 199 questions were run over AnswerFinder; divided as follows: 92 factoid questions, 51 definition questions and 56 list questions. The answers were manually assessed following an assessment scheme similar to the answer categories in iCLEF 2004:

- *Correct:* if the answer string is valid and supported by the snippets

- *Non-exact:* if the answer string is missing some information, but the full answer is found in the snippets.
- *Wrong:* if the answer string and the snippets are missing important information or both the answer string and the snippets are wrong compared with the answer key.
- *No answer:* if the system does not return any answer at all.

Table 1 provides an overall view, the system correctly answered 42.6% of these questions, whereas 25.8% wrongly, 23.9% no answer and 8.1% non-exactly. Table 2 illustrates AnswerFinder' abilities to answer each type of these questions separately.

Answer Type	
Correct	42.6%
Non exact	8.1%
Wrong	25.8%
No Answer	23.9%

Table 1. Overall view of AnswerFinder Monolingual Accuracy

	Factoid	Definition	List
Correct	63	6	15
Not exact	1	6	9
Wrong	22	15	13
No answer	6	23	18

Table 2. Detail analysis of AnswerFinder Performance-monolingual run

To measure the performance of AnswerFinder, recall (ratio of relevant items retrieved to all relevant items in a collection) and precision (the ratio of relevant items retrieved to all retrieved items) were calculated. Thus, recall and precision and F-measure for AnswerFinder are, 0.51 and 0.76, 0.6 respectively.

3.2 Systran Translation

Most of the errors noticed during the translation process were of the following types: wrong transliteration, wrong word sense, wrong word order, and wrong pronoun translations. Table 3 lists Systran's translation errors to provide correct transliteration 45.7%, wrong word senses (key word) 31%, wrong word order 25%, and wrong translation of pronoun 13.5%.

Below is a discussion of Systran' translation accuracy and the problems that occurred during translation of the TREC QA track questions.

Type of Translation Error	Percentage
Wrong Transliteration	45.7%
Wrong Sense	31%
Wrong Word Order	25%
Wrong Pronoun	13.5%

Table 3. Types of Translation Errors

Wrong Transliteration

Wrong transliteration is the most common error that encountered during translation. Transliteration is the process of replacing words in the source language with their phonetic equivalent in the target language. Al-Onaizan and Knight (2002) state that transliterating names from Arabic into English is a non-trivial task due to the differences in their sound and writing system. Also, there is no one-to-one correspondence between Arabic sounds and English sounds. For example P and B are both mapped to the single Arabic letter "ب"; Arabic "ح" and "ه "are mapped into English H.

Original text	Who is Aga[1] Khan[2]?
Arabic version	من يكون[1] اجا[2] خان؟
Translation (wrong)	From [EEjaa] [1] betrayed[2]?

Table 4. Incorrect use of translation when transliteration should have been used

Original text	Who is Hasan[1] Rohani[2]
Arabic version	من يكون[1] حسن[2] روحاني؟
Translation (wrong)	From spiritual[2] goodness[1] is?

Table 5. Wrong translation of person's name

Transliteration mainly deals with proper names errors when MT doesn't recognize them as a proper name and translates them instead of transliterating them. Systran produced 91 questions (45.7%) with wrong transliteration. It translated some names literally, especially those with a descriptive meaning. Table 4 provides an example of such cases where "Aga" was wrongly transliterated; and "khan" was translated to "betray" where it should have been transliterated. This can also be seen in table 5; "Hassan Rohani" was translated literally as "Spiritual goodness".

Wrong Word Sense

Wrong translation of words can occur when a single word can have different senses according to the context in which the word is used. Word sense problems are commoner in Arabic than in a language like English as Arabic vowels (written as diacritics) are largely unwritten in most texts.

Systran translated 63 questions (31.2%) with at least one wrong word sense, 25% of them could have been resolved by adding diacritics. Table 6 illustrates an error resulting from Systran's failure to translate words correctly even after diacritics have been added; the term "عِلم النَّفس" (psychology) was wrongly translated as "flag of breath". The Arabic form is a compound phrase; however Systran translated each word individually even after diacritics were added.

Original text	Who was the father of psychology?
Arabic version	مَن أب عِلم النَّفس
Translation (wrong)	From father flag of the breath?

Table 6. Example of incorrect translation due to incorrect sense choice

These errors can occur when a single word can have different senses according to the con-text in which the word is used. They also occur due to the diacritization in Arabic language. Arabic writing involves diacritization (vowel), which is largely ignored in modern texts. Ali (2003) gives a good example that can make an English speaker grasp the complexity caused by dropping Arabic diacritization. Suppose that vowels are dropped from an English word and the result is 'sm'. The possibilities of the original word are: some, same, sum, and semi.

Systran translated 63 questions out of 199 (31.2%) with wrong word sense, 25% of them can be resolved by adding diacritization.

Wrong Word Order

Word order errors occurred when the translated words were in order that made no sense and this problem produced grammatically ill formed sentences that the QA system was unable to process. Systran translated 25% of the questions with wrong word orders which lead to the construction of ungrammatical questions. Table 7 shows an example of wrong word order.

Original text	What are the names of the space shuttles?
Arabic version	ما أسماء المكوك الفضائي؟
Translation (wrong)	What space name of the shuttle?

Table 7. Wrong word order

Wrong Pronoun

Systran frequently translated the pronoun "وه" to "air" in place of "him" or "it" as shown in table 8. Table 9 shows an example pronoun problems; the pronoun "اه" is translated into "her", instead of "it", which refers to "building" in this case.

Original text	Who is Colin Powell?
Arabic version	مَن هُوَ كولن باول؟
Translation (wrong)	From Collin Powell air?

Table 8. Wrong translation of "*who*"

Original text	What buildings had Frank designed?
Arabic version	ما المباني التي صممها فرانك؟
Translation (wrong)	What the buildings which designed her Frank?

Table 9. wrong pro-drop of translation of "it"

It has been observed that Systran translation errors exhibited some clear regularities for certain questions as might be expected from a rule-based system. As shown in tables 2,3,4,7 the term "مَنْ" was translated to "from" instead of "who". This problem is related to the recognition of diacritization.

Arabic Word	English Translation Returned	Expected
حجم	Big	size
أطول	Tall, big	long
الأرض	Ground	earth
يوجد	Create	locate
دول	Nation	country
أبعد	Far	distance
أكثر	Many	most, much

Table 10. List of wrong key word returned by Systran

The evaluator observed that Systran' propensity to produce some common wrong sense translations which lead to change the meaning of the questions, table 10 shows some of these common sense translation.

3.3 The Effectiveness of AnswerFinder combined with Systran' Translation

After translating the 199 questions using Systran, they were passed to AnswerFinder. Figure 2 illustrates the system's abilities to answer the original and the translated questions; AnswerFinder was initially able to answer 42.6% of the questions while after translation, its accuracy to return correct answers dropped to 10.2%. Its failure to return any answer increased by 35.5% (from 23.9% to 59.4%); in addition, non-exact answers decreased by 6.6% while wrong answers increased by 3.6% (from 25.4% to 28.9%).

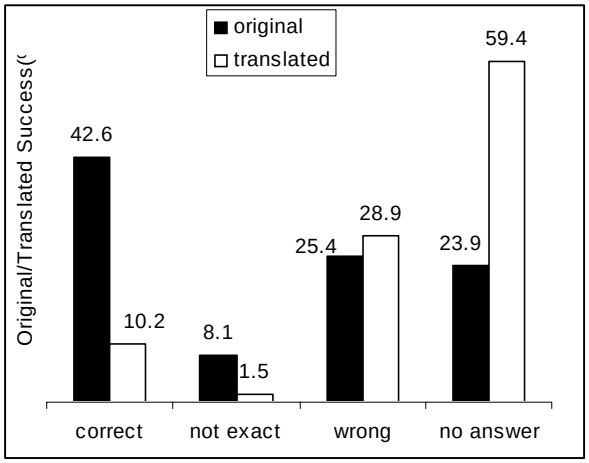

Figure 2. Answering Original/Translated Questions (%)

AnswerFinder was able to answer 23 translated questions out of 199. Out of these 23 questions, 12 were correctly translated and 11 exhibited some translation errors. Looking closely at the 12 corrected translated question (shown in table 11), 9 of them are of the factoid type, one of definition type and two of the list type. And by investigating the other 11 questions that exhibited some translation errors (shown in table 12), it can be noticed that 9 of them are factoid and 2 are list types. Our assumption for Systran' ability to translate factoid questions more than definition and list questions is that they exhibited less pronouns, in contrast to definition and list questions where many proper names were included.

In total, Systran has significantly reduced AnswerFinder's ability to return correct answers by 32.4%. Table 13 shows recall, precision and F-measure before and after translation, the value of recall before translation is 0.51 and after translation has dropped down to 0.12. Similarly, the precision value has gone down from 0.76 to 0.41, accordingly the F-measure also dropped down from 0.6 to 0.2. Altogether, in multilingual retrieval task precision and recall are 40.6% and 30%, respectively, below the monolingual retrieval task.

Question Type	Correctly translated	Correctly answered after translation
Factoid	(9/92)	(19/92)
Defn.	(1/51)	(0/51)
List	(2/56)	(3/56)
Total	12	13

Table 11. Types of questions translated and answered correctly in the bilingual run

Translation Error Type	Question Type
Word Sense	4 factoid
Word Order	2 factoid, 2 list
Diacritics	2 factoid
Transliteration	1 factoid
Total	11 questions

Table 12. Classification of wrongly translated questions but correctly answered

Effectiveness measure	Before Translation	After Translation
Recall	0.51	0.12
Precision	0.76	0.41
F-measure	0.6	0.2

Table 13. Effectiveness measures of AnswerFinder

4 Conclusions

Systran was used to translate 199 TREC questions from Arabic into English. We have scrutinized the quality of Systran's translation through out this paper. Several translation errors appeared during translations which are of the type: wrong transliteration, wrong word sense, wrong word order and wrong pronoun. The translated questions were fed into AnswerFinder, which had a big impact on its accuracy in returning correct answers. AnswerFinder was seriously affected by the relatively poor output of Systran; its effectiveness was degraded by 32.4%. This conclusion confirms Rosso et al (2005) findings using different QA system, different test sets and different machine translation system. Our results validate their results which they concluded that translation of the queries from Arabic into English has reduced the accuracy of QA system by more than 30%.

We recommend using multiple MT to give a wider range of translation to choose from, hence, correct translation is more likely to appear in

multiple MT systems than in a single MT system. However, it is essential to note that in some cases MT systems may all disagree with one another in providing correct translation or they may agree on the wrong translation.

It should also be borne in mind that some keywords are naturally more important than others, so in a question-answering setting it is more important to translate them correctly. Some keywords may not be as important, and some keywords due to the incorrect analysis of the English question sentence by the Question Analysis module, may even degrade the translation and question-answering performance.

We believe there are ways to avoid the MT errors that discussed previously (i.e. wrong transliteration, wrong word senses, wrong word order, and wrong translation of pronoun). Below are some suggestions to overcome such problems:

- One solution is to make some adjustments (Pre or Post- processing) to the question translation process to minimize the effects of translation by automatically correcting some regular errors using a regular written expression.
- Another possible solution is to try building an interactive MT system by providing users more than one translation to pick the most accurate one, we believe this will offer a great help in resolving word sense problem. This is more suit-able for expert users of a language.

In this paper, we have presented the errors associated with machine translation which indicates that the current state of MT is not very reliable for cross language QA. Much work has been done in the area of machine translation for CLIR; however, the evaluation often focuses on retrieval effectiveness rather than translation correctness.

Acknowledgement

We would like to thank EU FP6 project BRICKS (IST-2002-2.3.1.12) and Ministry of Manpower, Oman, for partly funding this study. Thanks are also due to Mark Greenwood for helping us with access to his AnswerFinder system.

References

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin,C., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E., Weishedel, R., (2001) Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). *Technical report, National Institute of Standards and Technology (NIST)*

Del Castillo, M., Montes y Gómez, M., and Villaseñor, L. (2004) QA on the web: A preliminary study for Spanish language. *Proceedings of the 5th Mexican International Conference on Computer Science (ENC04)*, Colima, Mexico.

Hull, D. A., and Grefenstette, G. (1996) Querying across languages: a dictionary-based approach to multilingual information retrieval. *Research and Development in Information Retrieval*, pp46-57.

Lin, F., & Mitamura, T. (2004) Keyword Translation from English to Chinese for Multilingual QA. In: *The 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*

Magnini, B., Vallin, A., Ayache C., Erbach, G., Peñas, A., Rijke, M. Rocha, P., Simov, K., Sutcliffe, R. (2004) Overview of the CLEF 2004 Multi-lingual Question Answering Track. In: *Proceedings of CLEF 2004*

Perrot, L. (2004). Question Answering system for the French Language. In: *Proceedings of CLEF 2004*

Plamondon, L. and Foster, G. (2003) Quantum, a French/English Cross-Language Question Answering System. *Proceedings of CLEF 2003*

Tanev, H., Negri, M., Magnini, B., Kouylekov, M. (2004) The Diogenes Question Answering System at CLEF-2004. *Proceedings of CLEF 2004*

Yaser Al-Onaizan and Kevin Knight. (2002) Translating Named Entities Using Monolingual and Bilingual Resources. In: *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*