Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 12 (13): 1371-1377, 2012
ISSN 1812-5654 / DOI: 10.3923/jas.2012.1371.1377

Arabic Malay Machine Translation for a Dialogue System

[1]Hamida Ali Almeshrky and [1,2]Mohd Juzaiddin Ab Aziz
[1]Department of Computer Science, Faculty of Qualifying Teachers,
Al-Mergeb University, Alkomes, Libya
[2]School of Computer Science, Faculty of Information Technology, University Kebangsaan Malaysia,
43600 Bangi, Selangor, Malaysia

Abstract: The purpose of any machine translation is set the translation by develop an automated translator sufficient in translating to achieve similar result with human translates as in case the translating from Arabic into Malay language. However, there are some challenges for translate from Arabic into Malay language where the translation depended on the purpose for which it meant according to context. In addition, Arabic and Malay languages are constructed in different structures such as free-word-order where Malay words order in target language is not the same order as the Arabic source language, pro-drop subject when it is attached as pronouns in word. To address such challenges in the system, the main goal of this research is to translate Arabic-Malay language dialogue system based on the transfer approach. Our system named AMmtDS has been created and developed. The notion of transfer approach involved three phases: analysis, transfer and generation phase. Considering the system evaluation has tested on a set of dialogues included the university campus dialogues from various Arabic dialogue books. The accuracy was 89.4% as result of comparing between human translation and our system the AMmtDS system. Based on the achieved results, the accuracy showed that the transfer based approach was available to translate a dialogue between Arabic and Malay languages.

Key words: Machine translation, dialogue system, transfer based approach, Arabic language, Malay language

INTRODUCTION

When a computer translates an entire document automatically and after that presents it to a human, the process is called Machine Translation (MT). MT sometimes called Automatic Translation (AT) (Abu-Shquier, 2009). MT is formally defined as the use of a computer to translate a message, typically text or speech, from one natural language (human language) into another natural language (Salem and Nolan, 2009). These definitions involve accounting some process for the grammatical structure of each language and using rules and grammar to transfer the grammatical of Source Language (SL) into Target Language(TL). MT attempts to automate part or all of the process of translating from one human language to another such as English and Arabic with or without human assistance (Arnold *et al*., 1994; Zughol and Abu-Alshaar, 2005). Depending on challenging which the work in field of Machine Translation (MT) faced it, whereas requiring accuracy and consistency at the same time. Among these challenges in MT which involve ambiguity multiword units such as collocations, idioms and structural differences units between languages (Nalluri and Kommaluri, 2011) for example word orderings that differ between languages. Another structural difference problem is tense generation. Tenses may exist in one language but in another may not. The lexical gaps also present challenges for translation where the target language has to represent a word in the source language. As well, the translation from Arabic to Malay is very challenged due to different structures in both sources and target language. Because of this, the translation accuracy requires understanding of context, besides understanding of the structure and rules of both languages.

To shift from previous part, researchers has recently access to many systems, both commercial and research of varying levels of performance.

In case of Malay language, there has been some work on Malay to or from a few languages like English and Japanese. Ogura *et al.* (1999) introduced a prototype Japanese-to-Malay translation system based on a semantic transfer method. This work address The different structure, where Japanese and Malay have quite free word orders but the orders are very different, where Japanese sentences are typically SOV, with the subject followed by

Corresponding Author: Mohd Juzaiddin Ab Aziz, School of Computer Science, Faculty of Information Technology, University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

the object and the verb, whereas Malay is SVO, like English. In addition, the noun phrase, in Japanese, modifiers come before the head noun (like in English), whereas in Malay they typically follow the head. Abdual Rahman and Abdul Aziz (2004) develop Example-Based MT system for English-Malay translation. The system depends on example translations kept in a Bilingual Knowledge .The work addressing the problem of "many English words to be represented in one Malay word", which then improves the linguistic quality of translation by English to Malay EBMT.

In case of Arabic language, much research has been conducted, during the last decade, in MT where the Arabic language was the focus. Most of them are on English -Arabic MT. Also, there are many researches on Arabic MT to or from other languages such as Chines. Shaalan (2000) addressed the translation of the Arabic interrogative sentence into English for automating the translation of user interfaces. This includes also the (imperative) form of the verbal sentence (Shquier and Sembok, 2008). This work presents English to Arabic approach for translating well-structured English sentences into well-structured Arabic sentences, using a Grammar-based technique to handle the problems of ordering and agreement. Shirko *et al*. (2010) this system present dealing with word order which Arabic exhibits. This poses a significant challenge to MT. The order in this study focus on noun phrase only, usually adjective order. Salem (2009) developed rule based approach using role and reference grammar based on Interlingua approach to translate from Arabic to English; they used the representation of the logical structure of an Arabic sentence in the proposed system, their desire to show how characteristics of Arabic language will affect the progress of MT tool. Newly works in Arabic (Habash and Hu, 2009). In this study, they explore different ways of pivoting through English to translate Arabic to Chinese. They use a standard phrase-based MT approach that is in the same spirit of most statistical MT to handle different linguistic phenomena in which Arabic, English and Chinese such as Subject-verb order, pro drop subject and nominal modification. The word order of Prepositional Phrases (PP) which differ between Arabic and chines.

Mat (2010) mention that the translation between Arabic and Malay was begun early in the fifteenth century. In fact, there was a book which had been translated on common belief (aqidah), entitled "Mother of the Evidence" in 1575. The translation then was named "The Beginning of Guidance" (Bidayah al-Hidayah). In spite of this, there is no one work on Arabic -Malay machine translation so far. So this study aim to combine Arabic and Malay in Machine translation for dialogue by using transfer based approach at first time.

DIALOGUE REPRESENTATION

It is very important to classify the language text to separate templates, then, develop machine translation of which of them. There are some templates according (Ahmed, 1999):

- Letter templates, the structure of this template as the following: letter date, letter header, letter, connectors, letter tail and sender name. This includes many special forms which need special grammar
- Traditional template: This template used to satisfy the needs of writing in so many fields like industry, sports, agriculture, economics and software
- The dialogues expression template is text that represents the dialogues between two people or more. The dialogue is not traditional sentence but it represents an interactive sentence like question phrase and incomplete or complete sentence, such as "what is your name?" and "My name is Ahmed" which represent question phrase and complete sentence, respectively

Sentences types of dialogue are likely to be one of the following types:

- Statement
- Interrogative (question)
- Command (imperative)
- Fragment (word or phrase)

Statement type: This type may be a noun phrase as example 1, 2, 3 and 4 in Table 1, or verb phrase as 5,6 and 7 in Table 1.

Table 1: Type of sentences in dialogue

Phrase	Arabic	Malay
Noun	صباح الخير	Selamat pagi
	بطاقة جديدة	Kad matrik baru
	بطاقتي الجديدة	Kad matrik baru saya
	المكتبة بعيدة جدا	Perpustakaan sangat terlalu
Verb	ذهبنا الى ماليزيا	Kami pergi ke Malaysia
	زوجي يعمل في الجامعة	Suami saya bekerja di Universiti
	يعمل زوجي في الجامعة	Suami saya bekerja di Universiti

Arabic is morpho-syntactically complex with many differences from Malay. We describe here the

prominent syntactic issues in which Arabic, Malay vary widely: order verb phrase and noun phrase.

First, as for the word order of noun phrases (example 1, 2, 3 and 4) in (Table 1), Arabic and Malay are similar in that modifiers always precede the modified such as example 2 in Table 1 "بطاقة"جديدة" kad matrik baru"(noun +Adjective). However, if the noun attached with possessive pronoun in Arabic structure, that lead to different order in the structure of Malay sentence such as example 3 in Table 1 "بطـــاقـــــــــي الجـــــــــديدة" (noun+pronoun+adjective) where the possessive pronoun "ي" connected with noun "بطاقة", in Malay "kad matrik baru saya" the possessive pronoun follow the adjective.

In Malay, if it has the structure (noun+adjective+ Adverb). The adverb comes in front of adjective (Karim, 1995). In contrast, the adverb in Arabic followed by the adjective (noun+adjective+adverb) such as example 4 " المكتبة بعيدة جدا" which translate to Malay "Perpustakaan sangat terlalu".

Secondly, Arabic verb subjects may be:

- Pro dropped (verb conjugated) as example 5 in Table 1 "ذهبنا الى ماليزيا" the subject verb (نا) is already connected to the verb (اتى) which indicates to a third person (نحن) (kami) the whole word meaning is "kami pergi "
- SVO the Similarities between Malay and Arabic languages both of them have been treated as SVO language for example 6 or " الجامعة زوجي يعمل في"
- VSO the default word order in Arabic is to begin VSO. By contrast, Malay language generally subject-verb languages as example 7 in Table 1 "يعمل زوجي في الجامعة" the subject (زوجي) appears after initial verb in Arabic sentence but at the beginning of the sentence in Malay language ""Suami saya bekerja di universiti". The morphology of the Arabic verb varies in the three cases. By contrast Malay is generally subject-verb language

Interrogative (question) type: Interrogative sentences used in the dialogue is to ask a question, or search for information using intonation and words between questioning and explanation of the type of head start in two different sentences interrogative, which are:

- **With interrogative pronoun:** There are some interrogative pronouns in both languages Arabic and Malay for example " اسمك ما", "apa nama awak?", where the interrogative "ما" gives the question meaning for the sentence. The following Table 2 realized the other interrogative pronoun that uses in dialogue

Table 2: Interrogative with pronoun

Arabic	Malay
ما،ماذا	Apa
متى	Bila
اين	Mana
من	Siapa
لماذا	Kenapa
كيف	Bagaimana

Table 3: Interrogative with prepositional phrase

Arabic	Malay
من اين	Dari mana
الى اين	Ke mana
مع من	Dengan siapa
منذ متى	Subah berapa lama

Table 4: Imperative type

Arabic	Malay
اعطني الكتاب	Beri saya buku
من فضلك املا النموذج	Sila isi borang

In Table 2 the interrogative "ما" has two meanings "apa and siapa", for "siapa" it used when asking about person name only, the word apa is rendered by siapa. It is impolite to use interrogative "apa"when ask about name for person (Sulaiman, 2000). In contrast, when asking about location name we use the word "apa".

- **Interrogative with prepositional phrase:** The second type in interrogative sentence can also formed with interrogative prepositional phrase as example "الى اين تريد ان تذهب " where the interrogative "الى اين " consists of a preposition "الى " and a interrogative "اين" so called interrogative with prepositional Phrase. Among the interrogative prepositional phrase are as following Table 3

Imperative type: An imperative sentence gives an order or makes a request. To identify the imperative sentence by using imperative verb type at beginning of sentence. as example 1 in table 4 or it may contain of the function word "من فضلك" , "sila" as second example in Table 4.

Fragment type: Fragment type is an incomplete sentences that are either a word as "لماذا", or phrases such as "في البيت" .

DESIGN ARCHITECTURE AND ALGORITHM

Figure 1 represents the whole process. Various methods of analysis and transformation will be used before obtaining the final result. The methods which are chosen and the emphasis depends largely on the design of the system, however, the system includes the following stages.

After input the any type of dialogue sentences, the tokenization process is tokenizing each sentence of

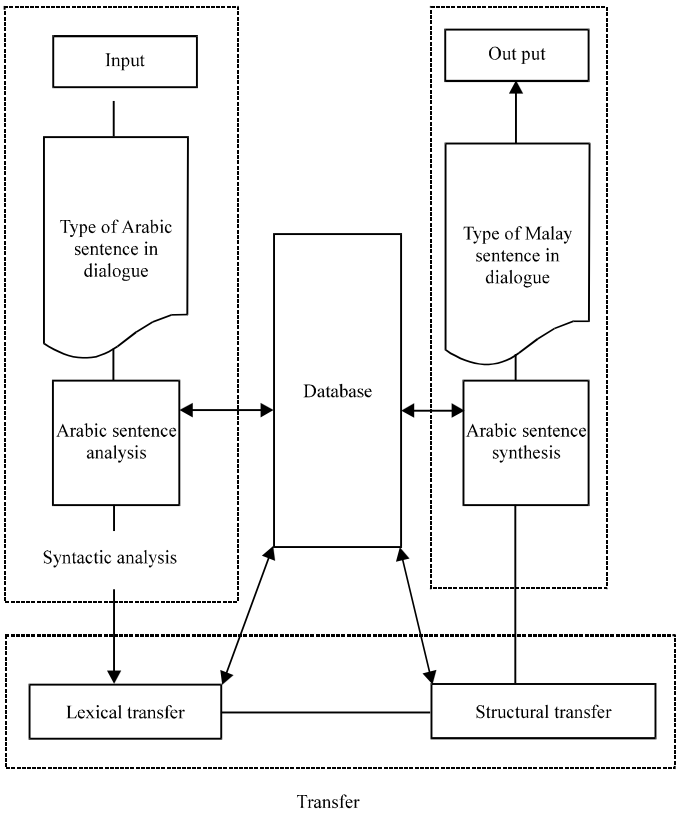


Fig. 1: AMmtDS architecture

dialogue for many tokens (words). The processing translation is based on the transfer approach with three main components: an analyzer, a transfer component and a generation component as follows:

Analysis components: The analysis step comprises Arabic morphological analyzer and syntactic analysis. The most important steps of this phase as follows:

Morphological analysis of Arabic sentence: Morphological analysis is the identification of a stem-form from a full (inflected) word-form (and sometimes also the identification of the syntactic category of the stem) (Shaalan, 2000). Arabic language is considering rich inflectional language, affixes can come anywhere in the word it can come at the first of the word (prefixes) or at the end of the word (suffixes).

The process of Morphological analysis as follow:

- Scan each enters word to break it down into beginning additions (prefix and suffix). What remain form breaking words is stem. Each word will be passing through the following multi-level
- **First level:** The input words without prefixes and suffixes, which means by making comparison the input word with the word stored in the database, then tokenizing it. If the word is not in the database then go to the second level
- **Second level:** The input word without prefixes and with suffix, which means to break down all the possible suffixes and then go to the database. If the word is not in the database then go to the third level.
- **Third level:** The input words without suffixes and with prefix, which mean to isolate all the possible prefixes and then go to the database. If the word is not in the database then we go to the forth level
- **Fourth level:** The input words with prefixes and suffixes, which means that we have to isolate all the possible prefixes and suffixes and then compare it with word stored in database. If the word is not in the database then we consider it no word

After these process of morphological analysis the assumed example "ذهبنا الى ماليزيا" it gives the result as the Table 5, where the each word analysis to prefix , stem and suffix.

The word "ذهبنا" has morphological analysis the stem "ذهب" and suffix "نا". The other words "الى" and "ماليزيا" the morphological has stem only because they do not attach with prefix or suffix.

Syntactic analysis: After morphological analysis the syntactical analysis is the process of analyzing a sequence of tokens (words) to determine its linguistic parts with respect to a given formal grammar. Syntactical analysis start with detecting the basic grammatical categories of V (Verb), N (Noun), Adj (Adjective), Adv (Adverb), Prep (Preposition), Conj (Conjunction), PN (Proper noun), Interrogative (Interr) Pronoun (Pron), function word (funcw) and their features sg (singular), pl (plural).

In Syntactical analysis, will be divided the sentence into smaller grouping according to their syntactic function. For example in Table 5 can be representation this sentence in next Table 6, where the word "ذهبنا" has grammatical categories verb and the grammatical categories of suffix is pronoun and has features plural (pl), where the analysis of word "الى" has grammatical categories as preposition (Prep) and Malaysia is proper noun (pn).

The transfer components: This second stage in translation, the transfer component is a system of rules that relate words and structures in one language to words and structures of another language (Arnold *et al*., 1994). Transfer starts with the output of the analysis phase and ends where the phase of generation starts. The transfer process is carried out by the help of a database. There two process of transfer: lexical transfer, structure transfer.

Lexical transfer: Lexical transfer applied at the word level. This maps Arabic lexical units and morphological features to their Malay equivalent.

Structural transfer: Structural transfer is applied transfer at the sentence level. This maps the Arabic syntactic structure result from lexical transfer to the equivalent Malay syntactic structure. The transfer of structural rules assigns restructuring of the linguistic parts of the source language and reordering the words according to linguistic parts of the target language. It is provide the rules for converting source language into equivalent target language (Trujillo, 1999).

For the example in Fig. 2 it's carry statement type in dialogue and has the pattern: V PRON PERP N which is VSO should be transferred to the pattern: PRON V PERP N, to SVO when translated to Malay should be reorder to equivalent Malay language structure.

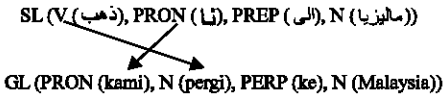


Table 5: Morphological analysis

Arabic	Prefix	Stem	Suffix
ذهبنا	-	ذهب	نا
الى	-	الى	
ماليزيا	-	ماليزيا	

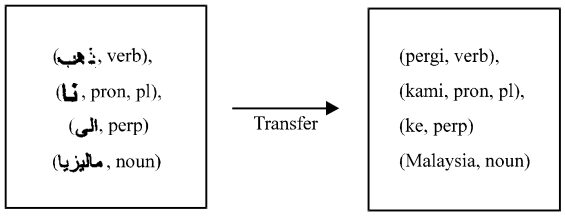
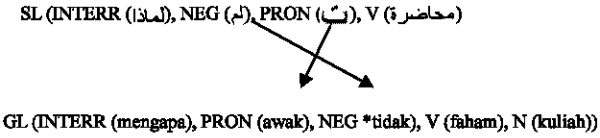
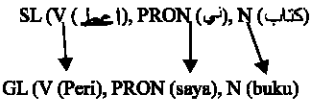


Fig. 2: Lexical transfer

For the example "لماذا لم تفهم المحاضرة؟" its carry interrogative type in dialogue and has the pattern: INTERR NEG PRON V N should be transferred to the pattern: INTERR PRON NEG V N ? Translated to Malay should be reorder to equivalent Malay language.

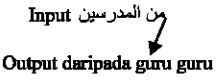


Another example of transferring rule, for imperative sentence its verb carries command meaning and it has the pattern: V PRON N should be transferred into the same pattern: V PRON N. This procedure illustrated in the sentence "اعطني الكتاب".



Generation components: The last step in translation process is generation of the target language input. The morphological generation module is responsible for synthesizing the target words in its right form. The input to this module is the origin of the target words which is passed by transfer phase with some information about each word such as number, tense, of noun.

This module synthesizes gives the rights form tense of verbs and the right number of nouns. But the tense of verb not marked in Malay language. However, to plural of noun in Malay by reduplicated the noun.



RESULTS AND DISCUSSION

In general, the aim of this experiment is to investigate whether AMmtDS machine translation system, sufficiently robust to be translated from Arabic dialogue to Malay. The experiment was based on the compare the output AMmtDS system with human translation. The experiment was applied on the set of test data include 56 sentences from university campus dialogues.

There are problems that appeared in machine translation from Arabic to Malay, implementation by AMmtDS system then classify the problems and assigns suitable scores for them. The score is given by human expert in translation and it measures the differences between the human translations depending on amounts of the magnitude of error mismatch in structure or meaning according effect of problem in sentence which expressed in a hypothetical translation:

10 = Match All
9-7 = Match Most
6-5 = Match Much
4-3 = Match Little
1-0 = Match
2-0 = None

The Error mismatch: There are some mismatches of error test examples which arise some problems in the output target sentence. The following specifies the problems that appeared in the generation target sentence:

- Synonyms problem. This problem shown because different synonyms of words are involved, such as the verb "يقي" could be translated to datang, or berasal
- Ambiguity. This problem appeared because the translation some words has more than one meaning. For example the meaning of subject pronoun انت. Refer to subject pronoun saya or refer to pronoun awak. The structure of sentence is correct grammatically but it isn't give correct meaning in dialogue according who speaker and listener. Also interrogative maybe translate to siapa or as preposition dari or daripada
- Addition and deletion. This problem appeared because the original translation contains extra words that have no corresponding words in
- Target language translation or target language has no corresponding words in original translation
- Conjunction with "و" and "ا". This problem appeared because in some situations an exception is made to use the "و" to separate two or more nouns. But in Malay used between the last noun and noun which preceded it only
- **Order of demonstrative:** This problem appeared because the order of demonstrative has different order in target language according using it in dialogue as pronoun demonstrative or as adjective demonstrative such as هذا الكتاب it translate to buku ini or ini buku.

Experiential results: Table 7 shows the sample for sentences of dialogue. In addition, it shows the output of sentences in AMmtDS system compared with original translation with number of problem that caused to mismatch the translation, then the score for each sentence.

We show the first row in AMmtDS output is match the original translation so it have score 10. In the fifth row the output is most match the original translation because according the problem 2 so it have score 8.

The result of translation evaluation as showing in Table 8 has performed by counting out the total score of sentences. The total score of whole sentences which is 483 divided by the whole number of sentences which (56*10) and finding their overall percentage which is 89.4.

The Table 8 showed that AMmtDS has scored the highest percentage. The significant improvement is attributed to the use of specific rules of dialogue sentences. We have already developed transfer-based framework for machine translation from Arabic dialogue to Malay. The translation form Arabic to Malay was newly developed well in this study. Various rules were discovered and developed in order to be able to cover more problems for Arabic dialogue sentences to Malay.

Table 6: Syntactic analysis

Arabic work	Prefix	Stem	Suffix
ذهبنا	-	ذهب, verb	نا, pron, pl
الى	-	الى, Prep	
ماليزيا	-	ماليزيا, noun	

Table 7: Test suite

Sentence	AMmtDS	Original translation	Problem No.	Score
صباح الخير	Selamat pagi	Selamat pagi		10
اهلا و سهلا	Selamt datang	Selamat datang		
ما اسمك؟	Siapa nama awak?	Siapa nama awak?		10
اسمي سليم ؟	Nama saya sleem.	Nama saya sleem		10
من اين اتيت؟	Dari mana saya datan?	dari mana awak datan?	2	8
اتيت من ليبيا	saya datang dari Libya	saya datang dari Libya		
اين تدرس؟	Di mana awak belajar?	Di mana awak belajar?		10

Table 8: Result of AMmtDS system

Machine translation	AMmtDS
Total score	483
Overall percentage	89.4

CONCLUSION

This study has been concentrated on issues in the implementation of a transfer-based MT system, which translates the dialogue into Malay. The dialogue is not traditional language but it is interactive which contain interrogative, imperative, statement or fragment sentences. We showed that the transfer based approach is promising and used to automate the translation of set of dialogues of 56 sentences from university campus and that system get 89.4% of correct translation. In future works, solve the ambiguity and the meaning problems by create specialized lexicons, study the more structure and design that developed to execute more complex sentences format which used in dialogue because this study focused on the most structure that used but not all. These improvements will raise the correctness of the translations from 89.4-100%.

REFERENCES

Abdual Rahman, S. and N. Abdul Aziz, 2004. Improving word alignment in an English-Malay parallel corpus for machine translation. The Amazing Utility of Parallel and Comparable Corpora, pp: 22-25. http://www.mt-archive.info/LREC-2004-Rahman.pdf

Abu-Shquier, M.A., 2009. Word agreement and ordering in English-Arabic machine translation: A rule-based approach. Ph.D. Thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia.

Ahmed, A.F., 1999. Developing an Arabic parser in a mulitilngual machine translation system. M.Sc. Thesis, Faculty of Computers and Information, Cairo University, Egypt.

Arnold, D.J., L. Balkan, S. Meijer, H.R. Lee and L. Sadler, 1994. Machine Translation: An Introductory Guide. Blackwell Publisher, Manchester, UK.

Habash, N. and J. Hu, 2009. Improving Arabic-Chinese statistical machine translation using English as pivot language. Proceedings of the 4th EACL Workshop on Statistical Machine Translation, March 30-31, 2009, Athens, Greece, pp: 173-181.

Karim, S., 1995. Malay Grammar for Academics and Professionals. Ministry of Education, Kualalumpur, Malaysia, ISBN-13: 978-9836240705, Pages: 391.

Mat, A.C., 2010. Revisiting Arabic-Malay translation experience in Malaysia: A historical and contemporary account. Can. Center Sci. Educ., 2: 99-103.

Nalluri, A. and V. Kommaluri, 2011. Statistical machine translation using joshua: An approach to build enTel system. http:// www. languageinindia. com/ may 2011 /anitha.pdf

Ogura, K., F. Bond and Y. Ooyama, 1999. ALT-J/M: A prototype Japanese-to-Malay translation system. Mach. Transl. Summit, 7: 444-448.

Salem, Y. and B. Nolan, 2009. Designing an XML lexicon architecture for Arabic machine translation. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, April, 22-23, 2009, Cairo, Egypt, pp: 221-229.

Salem, Y., 2009. A generic framework for Arabic to English machine translation of simplex sentence using the role and reference grammar linguistic model. M.Sc. Thesis, Computing in the School of Information and Engineering, The Institute of Technology Blanchardstown, Dublin, Ireland.

Shaalan, K., 2000. Machine translation of Arabic interrogative sentence into English. Proceedings of the 8th International Conference on Artificial Intelligence Applications, December 3-6, 2000, Egyptain Computer Society, Egypt, pp: 473-483.

Shirko, O., N. Omar, H. Arshad and M. Albared, 2010. Machine translation of noun phrases from Arabic to English using transfer-based approach. J. Comput. Sci., 6: 350-356.

Shquier, M.A. and T.M. Sembok, 2008. Word agreement and ordering in English-Arabic machine translation. Proceeding of the International Symposium on Information Technology, August 26-28, 2008, IEEE Xplore Press, Kuala Lumpur, pp: 1-10.

Sulaiman, O., 2000. Malay for Everyone: Mastering Malay through English. Kelana Jaya, Malaysia.

Trujillo, A., 1999. Translation Engines: Techniques for Machine Translation. 1st Edn., Springer-Verlag, London, UK., ISBN-13: 978-1852330576, Pages: 315.

Zughol, M.R. and A.M. Abu-Alshaar, 2005. English/Arabic/English machine translation: A historical perspective. Meta Transl. J., 50: 1022-1041.