ARABIC-MALAY MACHINE TRANSLATION USING RULE-BASED APPROACH

Ahmed Jumaa Alsaket and Mohd Juzaiddin Ab Aziz

Faculty of Information Science And Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

ABSTRACT

Arabic machine translation has been taking place in machine translation projects in recent years. This study concentrates on the translation of Arabic text to its equivalent in Malay language. The problem of this research is the syntactic and morphological differences between Arabic and Malay adjective sentences. The main aim of this study is to design and develop Arabic-Malay machine translation model. First, we analyze the adjective role in the Arabic and Malay languages. Based on this analysis, we identify the transfer bilingual rules form source language to target language so that the translation of source language to target language can be performed by computers successfully. Then, we build and implement a machine translation prototype called AMTS to translate from Arabic to Malay based on rule based approach. The system is evaluated on set of simple Arabic sentences. The techniques used to evaluate the correctness of the system translation are the BLEU metric algorithm and the human judgment. The results of the BLEU algorithm show that the AMTS system performs better than Google in the translation of Arabic sentences into Malay. In addition, the average accuracy given by human judges is 92.3% for our system and 75.3% for Google.

Keywords: Arabic, Malay, Machine Translation, Rule-Based

1. INTRODUCTION

Machine Translation (MT) is officially defined as the use of computers to translate messages in the form of text or speech from one natural language (human language) into another language of nature (Salem *et al*., 2008). This definition involves several processes accounting for grammatical structure of each language and uses rules and grammar for grammatical transfer from Source Language (SL) into the Target Language (TL).

To successfully conduct the process of translation, human translators need to have four types of knowledge. The first knowledge of the source language (lexicon, morphology, syntax and semantics) in order to understand the meaning of the source text. Second type is the knowledge of the target language (lexicon, morphology, syntax and semantics) in order to produce a comprehensible, acceptable and well formed text. The third type is the knowledge of "the subject matter". This enables the translator to understand the specific and contextual usage of terminology. Finally, the knowledge of the relation

between source and target language in order to be able to transfer lexical items and syntactic structures of the source language to the nearest matches in the target language.

Only a few machine translation systems can translate between these Arabic and Malay languages (Almeshrky and Aziz, 2012). The output of these translation systems, when translating from Arabic to Malay, still of low quality as they do not deal with these two languages directly. They use an intermediate language (a pivot language) and double translation process. On the other hand, building a direct statistical machine translation system requires a large parallel corpus for model training which is not yet available (Brown *et al*., 1993).

Malay is a major language of the Melayu-Polynesian, Oceanic or Austronesia family. At the level of morphology, Malay is an agglutinative language. New words in Malay language are formed by three methods: Attaching affixes onto a root word (affixation), formation of a compound word (composition), or repetition of words or portions of words (reduplication). At the level of syntax, the default sentence structure in Malay language is

Corresponding Author: Ahmed Jumaa Alsaket, Faculty of Information Science And Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

Subject-Verbal-Object (SVO) (Winstedt *et al*., 1957). In Malay, verbal grammatical category includes trunk verbs, adjectives and possessive verb.

Arabic language is a Semitic language. At the level of morphology, Arabic is a templatic, inflectional and derivational language (Al-Amoudi *et al*., 2013; Albared *et al*., 2010; 2011a; Mohammed and Aziz, 2011). At the level of syntax, Arabic is a subject pro-drop language. It has relatively free world order, mainly, nominal Sentence (SVO) and Verbal Sentence (VSO). However, the default sentence structure is Subject-Verb-Object (SVO).

This study describes our attempt to design a Machine Translation system from Arabic to Malay. Machine Translation is not a trivial task by nature of translation process itself especially when it involves two unrelated languages; languages that are not from the same family. We identify similarities and differences in morphological and syntax aspects between the Arabic and the Malay language in order to develop the translation rules. These rules should capture structural information and a set of constraints that capture feature information.

1.1. Related Work

Only a few machine translation systems can translate between these Arabic and Malay languages (Almeshrky and Aziz, 2012). Has implemented a machine translation system for Arabic to Malay language for a dialogue system. They state that both Arabic and Malay languages are constructed in different structures such as free-word-order, pro-drop subject when it is attached as pronouns in word. They use the transfer approach which consists of three main components: Analysis, transfer and generation component. This study identifies the rules to translate the dialogue from Arabic language to Malay language and build the database that includes the suitable words sense of Arabic and Malay words used in dialogue.

Abodina (2012) has implemented an Arabic-Malay dialogue translation system based on the rules which focus on the different structure of interrogative sentence, verb conjugated ordering and the different of adjective and adverb order in a dialogue sentence. In fact, this study is an extension of the (Almeshrky and Aziz, 2012) as they deal with the translation of Arabic dialogue sentences into Malay. Unlike Almeshrky and Aziz (2012) they deal with medical domain dialogues and they handle different problems.

Abdalla (2012) has implemented a machine translation system to translate Malay sentence into Arabic using rule-based Approach. In their rule-based machine translation system, the original text (Malay sentence) is first analyzed morphologically and syntactically in order to obtain a syntactic representation. Then, the syntactic representation is refined to be in more abstract level putting emphasis on the parts relevant for translation and ignoring other types of information. The transfer process then converts this final representation (still in the original language) to a representation of the same level of abstraction in the target language.

2. MATERIALS AND METHODS

The main processes and activities of the translation Arabic Malay system is illustrated in **Fig. 1** based on rule based approach.

The following subsections give detail descriptions of the process of Arabic to Malay translation system.

2.1. The Pre-Processing Stage

In the pre-processing step, a collection of operations are applied on Arabic input text to make it processable by the translation system. In this phase of the Arabic Malay translation system, several activities include text normalization, tokenization and proper nouns translations are applied to the Arabic sentences to processes them and to make them ready for translation. The following presents these activities in more detail.

2.2. Normalization

Normalization is a preliminary step to Arabic tokenization to ensure that the text is steady and predictable (Albared *et al*., 2011b; Shirko *et al*., 2010). It is a basic task that researchers in Arabic NLP always apply with a common goal in mind: Reducing noise and sparsely in the data. The major reasons for this problem in Arabic can be attributed to the phonetic variety in Arabic, transliteration of proper names and words borrowed from foreign languages. In this module, the following processes are performed:

- Removing the diacritics "خرج الطالب إلى الحديقة" الحَديقَةِ إلى الطُّلابُ خَرَجَ"
- Adding deleted characters. In Arabic, sometimes, some characters of a noun or verb are deleted due to its position in a sentence or if it is preceded with a special particle "لم تر النور" "لم ترى النور"
- Removal of redundant and misspelled space
- Resolution of the orthographic ambiguity "أإآإأ" "ي ى" in Arabic
- Removing the stretching character "~ "
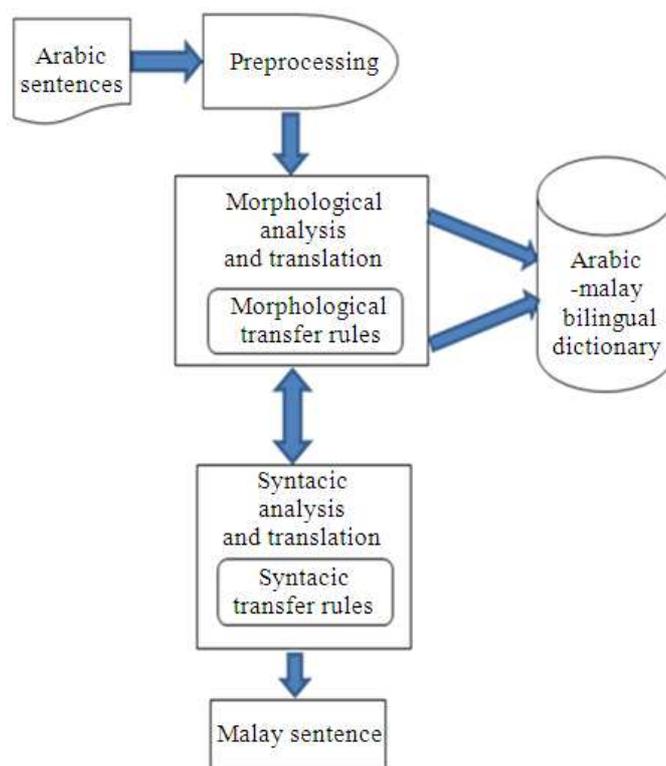
Fig. 1. Framework of AM-TS system

2.3. Tokenization

In this step, the system splits the sentence into words (tokens). The token can be a word, a part of a word (or a clitic), a multiword expression, or a punctuation mark (Attia, 2007). In fact, some of the Arabic researchers has identified this task as a part of the morphological analysis processes. The tokenization in our system extract clitics, the prefixes and the suffixes of each word in the input sentence.

- Clitics can be proclitics, which are precede the word (like a prefix) or enclitics which are follow the word (like a suffix). Enclitics for verbs in Arabic are object pronouns. Examples of clitics are Arabic Object Pronouns which are attached to verbs as their objects such as (verb+ني me/saya) and Arabic possessive Pronouns which are attached to nouns (noun+ي my/saya)
- A verb prefix: Verb prefixes are employed in general to specify the tense of a verb usually the present verb. May be a connected pronoun (subject pronoun), such as (ا، ت، ن، ي), or prefix that attached the present verb to indicates to the future tense such as (اس)
- A noun prefix: A noun prefix may a determiner (ال)
- A verb suffix: Verb suffixes are employed in general to specify the past tense when attached the verb such as (ان، ت)
- A noun suffix: On the other hand, noun suffixes are mainly concerned with determining the features of noun "person, number and gender" such as (ي،ني،ك،ت،ين).

2.4. Replacing Proper Nouns

Proper nouns such as personal names, days of month, days of week, country names, city names, bank names, organization names, ocean names, river names and university names from large percentage of unseen words. Instead of translating the proper nouns, the system identified them and transliterated them to their Malay equivalents. These words are stored into the proper noun database. Thus, to process this task, the system uses an Arabic proper noun database that has been built by other researchers (Benajiba, 2009). A sample of this database is shown in **Table 1**.

Table 1. Sample of the arabic named entities

هشام	الناصرة
سعد	افغانستان

2.5. Morphological Analysis and Translation

Arabic is a morphologically complex language. The morphological analysis of an Arabic word consists of determining the values of several morphological features, such part-of-speech, gender, number and so on. The analysis of words in a machine translation system is needed to determine their syntactic and semantic properties (Papineni *et al.*, 2002). In our system, we have designed our morphological analyzer using table lookup approach (dictionary based approach). An example of the output of the morphological analysis is shown in **Table 2**, given an input sentence " مدرسنا يملك دراجه نارية جديدة ".

2.6. The Morphological Generator

The main purpose of this sub phase is to produce the inflected Malay words in their correct forms. These Malay words may have passed from the previous sub phase (the morphological analysis) to this sub phase in their singular form with some features. Furthermore, the following discusses the generation rules that have been applied to generate the final Malay words.

2.7. Noun

- Removing the definite article: In general, if the Arabic word contains the definite article 'ال' then we remove it when translating to Malay. Such as "اليوم الثاني راح" that translated to "hari yang kedua panas". The following (**Fig. 2**) rule has been added
- Dual and plural forms: In case of translating dual forms in Arabic which usually end with 'ني' or 'ان', they are translated to Malay by adding the word 'dua' before the noun as shown in Example (1). The following (**Fig. 3**) rule has been added

In case of plural nouns, broken (irregular) plurals and sound (regular) plurals (Masculine sound plural nouns end in 'نو' or 'ني' and feminine sound plural nouns end in تا), are translated to Malay as in Example (1):

Example (1):

Number +N(plural form)	ست سيارات	Number+classifier +N(singular form)	Enam buah kereta
Number + N(plural form)	ست كراسي	Number+classifier +N(singular form)	Enam buah kerusi
Number +N(plural form)	أربعة فراشات	Number+ classifier +N-N	empat ekor rama-rama
[adverb + plural noun]	عدة دول	Number+classifier +N(singular form)	banyak Negara

Table 2. Example of the output of the morphological analysis

Arabic	Features	Malay
مدرس	Noun, S	Guru
نا	POSS_PR,1,P	Kami
يملك	Verb,M,S,3	Ada
دراجه	Noun,F,S	Basikal
جديدة	Adjective, F,S	Baru

[(defined) N: (defined)ADJ] Malay equivalent [N: ADJ]

اليوم الثاني: hari yang kedua

Fig. 2. Representation of delete definite article generation rule

[N (dual form) & ends with "ين" or "ان"] ⟺ 'dua' + N]

Example (1):

N(dual form)	كرسيان	Number +N	dua kerusi
N(dual form)	كرسيان	Number+classifier +N	Dua buah kerusi

Fig. 3. Representation of dual nouns generation rule

Translation of gender information: Arabic nouns are either masculine or feminine. Malay nouns are no directly inflected for gender. To translate the Arabic nouns with their gender information, First, these Arabic nouns are classified to two types (1) person nouns (2) animal nouns to adjust them with Malay system. Second, words laki-laki (male) and perempuan (female) are added to Malay sentence when Arabic noun refer to person or words jantan and betina as in Example (2):

Example (2):

pelajar lelaki	الطالب	الطالبة	pelajar perempuan
N N	N	N	N+N

Generation rules of possessive pronouns: These rules appear with nouns that contain possessive pronouns that are (ك/your, ي/my, هم/their, ان/our, ه/his, ها/her). On the other hand, in Malay the possessive pronouns are not a attached to noun, they are as one word that are (anda/your, saya/my, kami/our,mereka/their). The following rule (**Fig. 4**) has been added.

2.8. Syntactic Analysis and Generation

Syntactic analysis deals with the order and structure of a sentence (Abu Shquier, 2009). The syntactic analysis tries to handle a large difference of sentence constructions.
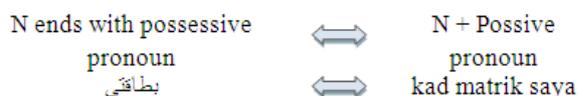
Fig. 4. Possessive pronoun generation rules

The syntactic analysis and generation of AM-TS system analyzes the phrasal structure and category of the Arabic sentence and uses the syntactic rules to transfer the Arabic sentence to the Malay sentence with right structure.

The following show some of these grammatical rules are produced from analysis of Arabic and Malay sentences.

Classifiers Transfer Rules: Another distinguishing feature of Malay is its use of measure words (penjodoh bilangan). In Malay language, classifiers (Penjodoh Bilangan) must be used when counting any object in a sentence. These classifiers are always followed by the nouns. The correct order is: Number + classifier + noun. The Arabic language does not use these types of classifiers, the order in Arabic Number + noun. To deals with this problem, we have added a special feature to classify Arabic nouns in the database, this feature have five values (type1, type2, type3, type4 or type5). After that we have use these rules to generate the corresponding Malay phrases.

Number+[N1 | NP1] ⟺ Number+Orang+[N | NP]
Number+[N2 | NP2] ⟺ Number+Ekor+[N | NP]
Number+[N3 | NP3] ⟺ Number+Batang+[N | NP]
Number+[N4 | NP4] ⟺ Number+ Buah +[N | NP]

Examples of applying these rules:

Tiga orang polis	ثالثة رجال شرطة
Tiga ekor ikan	ثالثة سمكات
Tiga ekor rama	ثالثة فراشات
Tiga batang pen	ثالثة أقلام
Tiga buah kerusi	ثالثة كراسي

Tense generation rules: Unlike Arabic verbs which are inflected for tenses, Malay Verbs are not inflected tense. The same form of verb can be used in all these situations. However, tense is instead denoted by time adverbs (such as "semalam") or by other tense indicators, such as sudah "already" and belum "not yet".

To translate Arabic sentences, we impose the following rules.

If the verb contains imperfect verb begins with any imperfect characters (أ، ن، ي، ت) (حرف مضارعة) and the sentence do not contain any imperfect adverbs such as "لازال" and "الآن", then the sentence translate straightforward as shown in the following Example (3):

Example (3):

هو يكتب الرسالة	dia menulis surat

If the verb contains imperfect verb begins with any imperfect characters (أ، ن، ي، ت) (حرف مضارعة) and the sentence contains any imperfect adverbs such as "لازال" and "الآن", then the sentence is translated to Malay and a time adverb or a tense indicators (sedang, masih, telah) is added as shown in Example (4):

Example (4):

هو يكتب الرسالة الآن	Dia sedang menulis surat sekarang
هو لازال يكتب الرسالة	Dia masih menulis surat
هو قد كتب الرسالة	Dia telah menulis surat

The same procedures are applied to sentences with perfect verbs.

In the above sentences we can note that the pronoun is explicitly written" separated pronouns", so we translate them directly. In the other case where the pronouns are not explicitly written "connected pronouns", we always check the verb prefix and the verb suffix to get the number, the gender and the tense of the sentence, Example (5):

Example (5)

يكتب الرسالة	Dia menulis surat
هو يكتب الرسالة	Dia menulis surat
هما يكتبان الرسالة	Mereka menulis surat
يكتبان الرسالة	Mereka menulis surat

Superlative adjectives generation rules: To translate the superlative adjective in both Arabic to Malay: We classified Arabic superlative adjectives into three categories (SADJ1, SADJ2 and SADJ3), this classification is based on the number of ways their Malay Superlative adjectives equivalent are formed, Example (6):

Example (6):

الأفضل	SADJ1	Yang baik sekali
الأقوى	SADJ2	Yang paling kuat
الأكبر	SADJ3	Terbasar

In addition, when inflected Arabic adjectives are translated, they are first stemmed to remove inflection and then we look in the lexicon for the direct translation of these stems.

3. RESULTS AND DISCUSSION

There are many methodologies for evaluating the performance of Machine Translation system. Most of

these strategies are based on computing some kind of similarity score between the output of an MT system and one or more reference translations. In this research, we have used have used two methodologies to evaluate the performance of the AM-TS. The first experiment we evaluate our system IBLEU metric (Papineni *et al.*, 2002). In the second experiment human judgment methodology is used for evaluation.

3.1. The BLUE Evaluation Methodology

In this experiment we have evaluated a sample of our system and Google translation output using the iBLEU system which is online implementation of BLUE algorithm. First, the evaluation procedure is done sentence by sentence from the test case. We compute BLUE scores (1-gram, 2-grams and 3-grams) for all sentences in a MT outputs. After that we compute the overall average of each n-gram BLUE scores. **Table 3** presents BLUE score of Google and our system for 1-gram, 2-gram and 3-gram.

According to results of the iBLEU evaluation, we can assert that the AM-TS system performs better than Google in the translation of simple Arabic sentences into Malay. As shown in **Table 3** the average score of 1-gram, 2-gram and 3-gram for Google is 0.61, 0.44 and 0.55 respectively. In fact the Google translation of Arabic into Malay is not direct, it uses a pivot language. First it translates Arabic to English then from English to Malay. The use of the pivot language technique always leads to the loss in translation quality due to the process of double translation. **Table 3** also shows that the average score of 1-gram, 2-gram and 3-gram for AM-TS system is 0.98, 0.93 and 0.92 respectively. So based to results of 1-gram, 2-gram and 3-gram AM-TS system is able to generate a better translation than Google when it comes to the translation of simple Arabic sentences into Malay.

3.2. The Human Evaluation Methodology

Human judgment methodology is the traditional method used to evaluate the quality of machine translation. The following steps describe this methodology:

- Run and test AM-TS system on the selected test case
- Compare the human translation with the system output
- Assign a suitable score for each problem. A range of score between 0 and 10 While 0 indicates absolutely incorrect translation, 10 indicate absolutely matched translation between 0 to 10 amounts of the

magnitude of error in structure or meaning which expressed in a hypothetical translation:

10 = Match All
9-7 = Match Most
6-5 = Match Much
3-4 = Match little
0-2 = Match none

- Determine the correctness of the test case by computing the percentage of the total scores

The final average score given by this method are shown in **Table 4**.

As presented in **Table 4**, the average score of AM-TS based on the human evaluations are: 91.2% and the average score of Google is: 78.0%. Based on these results, it is obvious that the performance of AM-TS is better than Google's which indicates AM-TS can produce a better translation when it comes to the translation of simple Arabic sentences into Malay.

Table 3. Experiment 1 results: The BLEU Score for google and AM-TS

No.	Google			AM-TS		
	n1	n2	n3	n1	n2	n3
S1	0.6667	0.5000	0.5000	1.0000	1.0000	1.0000
S2	0.3333	0.2500	0.2500	1.0000	1.0000	1.0000
S3	0.3333	0.2500	0.2500	1.0000	1.0000	1.0000
S4	0.3333	0.2500	0.2500	1.0000	1.0000	1.0000
S5	0.3333	0.2500	0.2500	1.0000	1.0000	1.0000
S6	1.0000	1.0000	1.0000	0.6667	0.2500	0.2500
S7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
S8	0.6667	0.2500	0.2500	1.0000	1.0000	1.0000
S9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
S10	1.0000	0.5000	1.0000	1.0000	1.0000	1.0000
S11	1.0000	0.5000	1.0000	1.0000	1.0000	1.0000
S12	0.4000	0.1250	0.0833	1.0000	0.8333	0.6000
S13	0.5000	0.5000	1.0000	1.0000	1.0000	1.0000
S14	0.1667	0.1250	0.1250	1.0000	1.0000	1.0000
S15	0.6000	0.1250	0.0833	1.0000	0.5000	0.5000
S16	0.6667	0.2500	0.2500	1.0000	1.0000	1.0000
S17	0.5000	0.5000	1.0000	1.0000	1.0000	1.0000
S18	0.5000	0.5000	1.0000	1.0000	1.0000	1.0000
S19	0.7500	0.6667	0.5000	1.0000	1.0000	1.0000
S20	0.5000	0.3333	0.2500	1.0000	1.0000	1.0000
AVG	0.6125	0.4437	0.5520	0.9833	0.9291	0.9175

Table 4. The average score of human evaluation on google and AM-TS

Machine Translation (MT)	Average score
Google	78.5
AM-TS	91.2

4. CONCLUSION

In this study, we have demonstrated the application of morphological and syntactic translation rules approach for Arabic to Malay machine translation system. Our system (AM-TS) consists of three main phases, the pre-processing phase, morphological analysis and translation phase and the syntactic analysis and generation phase. Two evaluation methodologies have been used to evaluate AM-TS system: IBLEU metric (Papineni *et al.*, 2002) and Human judgment. Based on the results, it is obvious that the performance of AM-TS is better than Google's which indicates AM-TS can produce a better translation when it comes to the translation of Arabic sentences into Malay.

5. REFERENCES

Abdalla, H., 2012. Malay To Arabic Rules-Based Machine Translation Based on Word Ordering And Agreement. MSc Thesis, Universiti kebangsaan Malaysia, Bangi.

Abodina, A.A., 2012. Arabic to malay medical dialgue translation system based on the grammatical rules. MSc Thesis, Universiti kebangsaan Malaysia, Bangi.

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.

Al-Amoudi, A., H. Al-Mazrua, H. Al-Moaiqel, N. AlOmar and S. Al-Koblan, 2013. An exploratory study of arabic language support in software project management tools. Int. J. Comput. Sci., 10: 56-63.

Albared, M., N. Omar and A.M.J. Aziz, 2011a. Developing a competitive HMM arabic POS tagger using small training corpora. Proceedings of the 3rd International Conference on Intelligent Information and Database Systems, Apr. 20-22, Springer Berlin Heidelberg, Daegu, Korea, pp: 288-296. DOI: 10.1007/978-3-642-20039-7_29

Albared, M., N. Omar, A.M.J. Aziz and M.Z.A. Nazri, 2010. Automatic part of speech tagging for arabic: An experiment using bigram hidden markov model. Proceedings of the 5th International Conference, Rough Set and Knowledge Technology, Oct. 15-17, Beijing, China, pp: 361-370. DOI: 10.1007/978-3-642-16248-0_52.

Albared, M., N., Omar and A.M.J. Aziz, 2011b. Improving arabic part-of-speech tagging through morphological analysis. Proceedings of the Intelligent Information and Database Systems, Apr. 20-22, Daegu, Korea, pp: 317-326. DOI: 10.1007/978-3-642-20039-7_32

Almeshrky, H.A. and M.J.A. Aziz, 2012. Arabic malay machine translation for a dialogue system. J. Applied Sci., 12: 1371-1377. DOI: 10.3923/jas.2012.1371.1377

Attia, M.A., 2007. Arabic tokenization system. Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, (CIR' 07), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 65-72.

Benajiba, Y., 2009. Arabic named entity recognition. Ph.D. Thesis, University of Valencia, Spain.

Brown, P.F., V.J.D. Pietra, S.A.D. Pietra and R.L. Mercer, 1993. The mathematics of statistical machine translation: Parameter estimation. Comput. Linguist., 19: 263-311.

Mohammed, E.A. and A.M.J. Aziz, 2011. English to arabic machine translation based on reordring algorithm. J. Comput. Sci., 7: 120-120. DOI: 10.3844/jcssp.2011.120.128

Papineni, K., S. Roukos, T. Ward and W.J. Zhu, 2002. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, (ACL' 02), Stroudsburg, pp: 311-318. DOI: 10.3115/1073083.1073135

Salem, Y., A. Hensman and B. Nolan, 2008. Implementing Arabic-to-English machine translation using the role and reference grammar linguistic model. Proceedings of the English Annual International Conference on Information Technology and Telecommunication, (ITT' 08), Galway, Ireland, pp: 104-110.

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: 350-350. DOI: 10.3844/jcssp.2010.350.356

Winstedt, R.O., S.O. Winstedt and R.J. Wilkinson, 1957. Malay Grammar: By RO Winstedt. 2nd Edn., Clarendon Press, pp: 205.