Journal of Computer Science 6 (3): 350-356, 2010
ISSN 1549-3636

Machine Translation of Noun Phrases from Arabic to English Using Transfer-Based Approach

Omar Shirko, Nazlia Omar, Haslina Arshad and Mohammed Albared
Department of Computer Science, Faculty of Information Technology,
University Kebangsaan Malaysia, Bangi, 43600, Selangor, Malaysia

Abstract: Problem statement: Any Arabic to English Machine Translation (MT) system should be capable of dealing with word order which Arabic exhibits. This poses a significant challenge to MT due to the vast number of ways to express the same sentence in Arabic. The ordering features are very important and should be carefully applied to ensure the generation of sentence in the target language. Because they apply to the target language, it should fulfill the specific requirement of this language. Mistakes in the MT output can be either the result of analysis problems at the source language level, or due to generation problem at target language level. Word order rules are crucial for the generation of sentences in the target language. They also serve as rules for the ordering of sentence constituents. These rules draw their information from the syntactic knowledge. The word order problem becomes more obvious when making machine translation between languages that have rich morphological variations. **Approach:** The main objective of this research is to develop a machine translation that translates Arabic noun phrases into English by using transfer-based approach. A system called Npae-Rbmt has been developed in this research. Transfer-based machine translation is one instance of rule-based machine-translation approaches and is currently one of the most widely used methods of machine translation. The idea of transfer-based machine translation it is necessary to have an intermediate representation that captures the “meaning” of the original sentence in order to generate the correct translation. Using advantages of transfer-based machine translation such as analysis step, the Transfer-based becomes simpler as linguistic analysis goes deeper-as the representation of analysis step becomes more abstract. In fact, a major goal of MT research is to define a level of analysis which is so deep in which transfer-based machine translation is able to do. **Results:** The method was tested on 88 thesis titles and journals from the computer science domain. The accuracy of the result was 94.6%. These results proved the viability of this approach for distant languages. **Conclusion:** Based on the achieved results, we have managed to perform the syntactic reordering within an Arabic noun phrases to English translation task by using transfer-based machine translation and also achieved reasonable improvements in translation quality over related approach.

Key words: Npae-Rbmt, machine translation, transfer-based approach, noun phrases, Arabic language processing

INTRODUCTION

Machine Translation (MT) is formally defined as the use of a computer to translate a message, typically text or speech, from one natural language to another (Salem and Nolan, 2009). Machine translation system develops by using four approaches depending on their difficulty and complexity. These approaches are: rule-based, knowledge-based, corpus-based and hybrid MT. Rule-based machine translation approaches can be classified into the following categories: direct machine translation, interlingua machine translation and transfer-based machine translation (Abu Shquier and Sembok, 2008). This study adopts the transfer-based machine translation.

Transfer-based machine translation: One of the main features of transfer based machine translation systems is a phase that “transfers” an intermediate representation of the text in the original language to an intermediate representation of text in the target language (Shaalan *et al*., 2004). This can work at one of two levels of linguistic analysis, or somewhere in between. The levels are: Superficial transfer (or

Corresponding Author: Omar Shirko, Department of Computer Science, Faculty of Information Technology, University Kebangsaan Malaysia, Bangi, 43600, Selangor, Malaysia

syntactic). This level is characterized by transferring “syntactic structures” between the source and target languages. It is suitable for languages in the same family or of the same type, for example in the Romance languages between Spanish, Catalan, French, and Italian. Deep transfer (or semantic). This level constructs a semantic representation that is dependent on the source language. This representation can consist of a series of structures which represent the meaning. In these transfer systems predicates are typically produced. The translation also typically requires structural transfer. This level is used to translate between more distantly related languages (e.g., Spanish-English or Spanish-Basque). There are four advantages of the transfer-based architecture that make it appealing for many researchers: First, is applicability. While it is difficult to reach the level of abstractness required in interlingual systems, the level of analysis in transfer models is attainable. Second, portions of transfer modules can be shared when closely related languages are involved. For example, an English-Portuguese module may share several transformations with an English-Spanish module. Third, ease of implementation. Developing a transfer MT system require less time and effort than Interlingua. Four, it is easy to acquire linguistic knowledge, and it is easy to augment the grammar rules with heuristic rules (Shaalan, 2005).This is why many operational transfer systems have appeared in the market.

As a natural language, Arabic has much in common with other languages like English. However, it also is unique in terms of its history, internal structure, inseparable link with Islam, and the Arabic culture and identity (Farghaly and Shaalan, 2009). Over the last few years, Arabic Natural Language Processing (ANLP) has gained increasing importance, and several state of the art systems have been developed for a wide range of applications including such as machine translation (Farghaly and Shaalan, 2009). Arabic natural language processing in general is still underdeveloped (Monem *et al*., 2008). Moreover, tools used for other languages are not easily adaptable to Arabic due to the language complexity at both the morphological and syntactic levels (Monem *et al*., 2008). Arabic linguistic is usually unclear and the parts of speech are difficult to define (Salem and Nolan, 2009). That is why most researchers in Arabic machine translation are concentrating more in English to Arabic translations such as, Shaalan *et al*. (2004) the reported their attempt in automating the translation of English Noun Phrase (NP) into Arabic. The system is implemented in Prolog and the parser is written in DCG formalism. Abu Shquier and Sembok (2008) presented the word agreement and ordering in English-Arabic machine translation by using rule-based approach. On the contrary, little work has been done in developing Arabic-English MT systems. Shaalan *et al*. (2004) described a tool for translating the Arabic interrogative sentence into English. Salem and Nolan (2009) developed a system, which translates from Arabic to English using the Role and Reference Grammar Linguistic Model. At present there is not much work on Arabic noun phrases to English machine translation.

The research discusses a system based on NP, called Npae-Rbmt system. The Npae-Rbmt translates Arabic noun phrases to English by using Transfer-based bridge. The Npae-Rbmt systems understand the part of speech (pos) of a word, number, gender and the word type. The motivation for this study is to develop an automated translator sufficient in translating from Arabic noun phrases into English.

MATERIALS AND METHODS

The architecture of Npae-Rbmt system: The architecture of Npae-Rbmt system is given in Fig. 1. In this Fig. 1, the arrow shows the flow of information. Egg-shaped blocks represent the essential modules of the system. Rectangle blocks symbolize the linguistic knowledge. The Npae-Rbmt system is based on the transfer-based architecture with three major stages: Analysis stage, a transfer stage and a generation stage.

The following summarizes the translation process (Shaalan *et al*., 2004):

- In the first step of transfer-based architecture is the morphological analysis, where this step provides inflectional features as well as the stem form of an inflected Arabic word

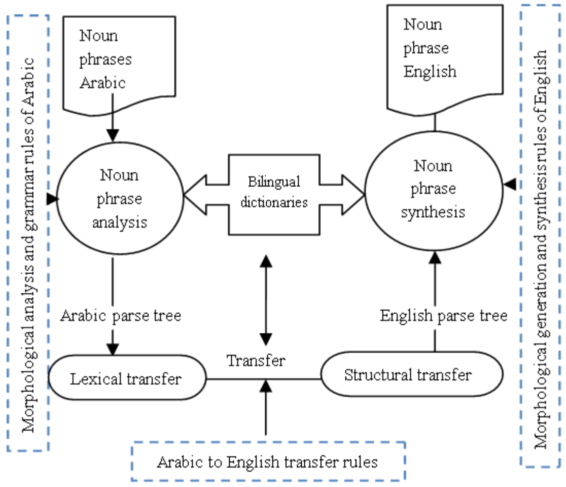


Fig. 1: Overall architecture of Npae-Rbmt system

- After the morphological analysis is performed, a syntactic parser will build the syntactic dependency tree. Syntactic dependency tree is considered important in the language translation because it provides the linguistic relationship between units of a noun phrases
- Lexical transfer will chart Arabic lexical units to their English lexical units. It will also chart Arabic morphological features to the equivalent set of English features
- Structure transfer will chart the Arabic dependence tree to the corresponding English syntactic arrangement
- The morphological generator will synthesize the inflected English word in its right form based on the morphological features
- A syntactic generator is accountable for tuning up and constructs the surface structure of these phrases. This step will be followed by navigating the final tree to generate the translation output

Syntax analysis: Arabic language has very numerous and complex morphological rules (Shaalan *et al*., 2007). Arabic morphological analysis has gained the focus of Arabic natural language processing research for a long time in order to achieve the automated understanding of Arabic (Shaalan *et al*., 2006). Arabic is based on the Semitic root-and-pattern scheme of forming word roots, as well as the concatenation of root and affixes (Shaalan, 2005). So to do the analysis in Arabic language need sophisticated morphological analysis (Shaalan, 2005). There are two main techniques in dealing with morphological analysis in MT systems. First technique uses a database to store all full-form words. Using full-form words means that the root and irregular forms are stored in a database (Abu Shquier and Sembok, 2008). Each of the items (تنفيذ و منفذ و يتنفذ) will all be entered explicitly into the database to be identified as relating to the same root (نفذ). When the system uses a full-form database it will not have to bother about irregular forms, as all words are treated in the same way and entered explicitly. Second technique uses a morphological analysis component to analyze words and identify them as roots and affixes or prefix. A morphological analysis component is a rule-based module which is able to analyze a word and relate it to its root form (Abu Shquier and Sembok, 2008).Its advantages and disadvantages are the opposite of the full-form technique. It can capture morphological generality and identify newly-formed words. The cost for updating and maintaining the system is minimal as modification is made in a single module, which is then applied to all morphological rules, may be higher than that of the first technique. In this study we have used the both techniques morphological analysis and database, as we stored the full-form words (the root and all its derivatives).

There are two steps of process in parser development, the rules of Arabic noun phrase should be obtained first and thus will give a precise account of what it is for a noun phrase to be grammatically correct (Shaalan *et al*., 2004). The analysis indicates that noun phrases can occur only on two levels, a simple form and complex form combination of two or more simple NPs. Those combinations are also separated by connector, preposition and separator and also special words such as symbol ":", colon or any word shows the beginning of new NP. Secondly, the parser which shows grammatical structure should be applied as NP input. To do this, a morphological analysis should be done on the structure of inflected Arabic words. The analyzer will reverse the parser of words in its stem form.

We have applied a rule-based module which is capable of analyzing a word and relate it to its root form and interpret meaning chunks conveyed by the affixes and suffixes attached to the word.

Bilingual dictionary: Dictionaries are the most important tools in a machine translation system (Shaalan *et al*., 2004). The process of translations is more helpful and become a significant advantage when a dictionary is referred to get more understanding of overview structure of particular vocabulary. The entries of bilingual dictionary content a particular vocabulary, these used to mention all the exacting area and terms in bilingual dictionary. In our transfer process, these vocabularies are needed in order to satisfy condition understanding of dictionary construction.

Our bilingual dictionary contains all full-form words (the root and all its derivatives) for both Arabic and English languages with all its features and Part Of Speech (POS).The following describe the entries (word categories) that included in our proposed bilingual dictionary:

- Noun: A content word which has four features the stem-form, the number (single, plural), definition and gender. This features for Arabic and English
- Adjective: In Arabic language adjective like nouns has the same properties. The English has stem-form features
- Adverb: Constant word which has the stem-form. This features for Arabic and English
- Quantifier: Constant word which has the stem-form. This features for Arabic and English

- Separator: A function word, which has a stem-form feature. This includes connectives, prepositions and special words that are used as a separator of a compound noun phrase

Syntax transfer MT: Many scholars such as Eynde (1993) have argued that in order to analyze the translation representation properly, two main steps need to be taken. Firstly, detect the constitute structure of the source language. Secondly, resolve the lexical and syntactic ambiguities. When the translation is in the transfer stage; all aspects of lexical or structural differences between the source and destination language will be captured. Transfer starts with the output of the analysis phase and ends where the phase of generation starts (Abu Shquier and Sembok, 2008). When the translation is in the transfer stage; all aspects of lexical or structural differences between the source and destination language will be captured. Transfer starts with the output of the analysis phase and ends where the phase of generation starts (Abu Shquier and Sembok, 2008). In the extent of this research; the translation actually occurs in the transfer phase. There are two types of transfer: First, lexical transfer. Translation experts Hutchins and Somers (1992) suggest that in an ideal hypothetical world of lexical transfer, each source language has only one equivalent language target word. The monolingual grammar of either target or source language can be seen in the rules of bilingual dictionary, for example: (N knowledge). The parsers use these rules to analyze each sentence for both Arabic and English language which demonstrate their fundamental structure and by generators to produce output sentences from such representation (Shaalan *et al.*, 2004). Every source sentence representation must relate into the target language representation, a representation which will form the basis for generating a target language translation (Shaalan *et al.*, 2004). The following is an example of a lexical transfer from Arabic to English:

معالجه	⟷	processing
اللغات	⟷	language
الطبيعية	⟷	natural

Features in translation system consist of numbers, definition, attribute or value such as singular or plural. Any rules of noun phrase are a simple and straightforward translation of source structure to target structure: {number = sg}_{number = sg} (Shaalan *et al.*, 2004). These dictionary rules can be seen as relating leaves (the word nodes) on the Arabic parse tree to leaves on the target English tree (Fig. 2).

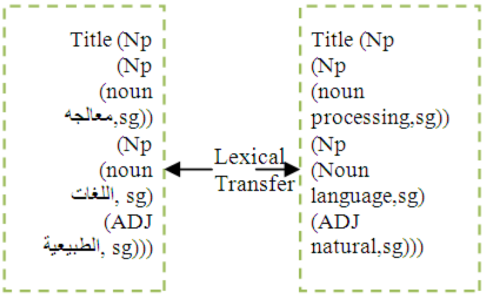


Fig. 2: Lexical transfer

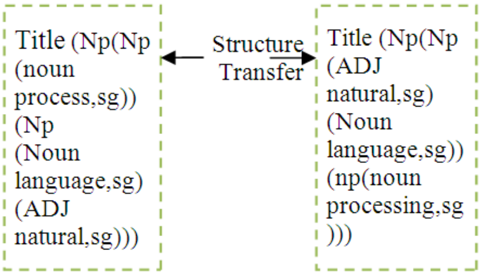


Fig. 3: Structure transfer

Second, structural transfer: The structural transfer provides the rules for converting source language parse trees into equivalent target language trees (Trujillo, 1999). Restructuring the parse tree and reordering the words are also performed in the transfer of structural according to the target linguistic. This process need for comparative grammar as it contains some structural rules and these rules are related to each other with the nodes of the two trees (Shaalan *et al.*, 2004). In Arabic NP, there is a strong connection among the adjacent lexical units. The order of positioning lexical units in the NP is different to English, which the nouns come before adjective, as an example in this phrase: رجل جيد. Contrary to English language, the adjective precedes its noun, as in the phrase: Good man. To do the restructuring, the Arabic parse tree is still needed to get the grammatically correct translation of the target English. The transfer rules described here deals with the restructuring of the parse tree and reordering of words (Fig. 3).

The main indicator of difference between Arabic and English can be seen through the parse tree representations, which in all the words are in opposite order. Here are some list of rules of transfer and its explanation with example to get better understanding. In this rule LHS describes the Arabic structure, while RHS describes the English structure. The LHS and RHS are considered standardized acronym for structural

Rule and $1, $2, $k are variables interpreting as standing for pieces of Arabic structure on one side and for their translations on the other side. A relatively simple straightforward example where a more complex example is called for involves the translation of " الشبكات تقيم اداء" into "networks performance evaluation" which shows the switching of words. Such a rule might look as follows:

(LHS) Arabic title $[w_i: \$1, w_{i+1}: \$2, \ldots, w_k: \$k]$ (1 <i < k)

⟷

(RHS) English equivalent $[w_k: \$k, w_{k-1}: \$k\text{-}1, \ldots, w_i: \$i]$ (1 <i < k)

This rule says that the translation of the word at level i is switched with the word at level k-i+1. Where k is the number of noun phrases equivalent to maximum (sub) tree level. This rule is used when we encounter sentence or a part of sentence that is completely consist of nouns or nouns and adjectives. In general, almost all noun phrases are in a compound form. The translation rule of a compound noun phrase looks as follows:

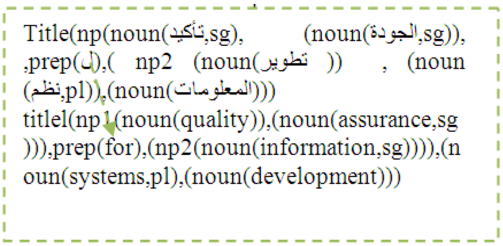
Arabic title [NP: $1, prep: $2, NP: $3]

⟷

English equivalent [NP: $1, prep: $2, NP: $3]

As an example, consider the translation " تأكيد الجودة لتطوير نظم المعلومات" into "Quality assurance for information system development". The rule associates $1 with the sub tree of " تأكيد الجودة", $2 with the node for ""ل and $3 with sub tree for ""تطوير نظم المعلومات. Translating each of these then becomes a separate task for transfer. It operates on these sub trees in the same way as in the original tree attempting to find rules which deal with these sorts of structure. In the Fig. 4 we will reverse the structure of the sentences in Arabic simply clear to the reader. The variables, which appears in Fig. 4 shows that the features that bring with every word like (s-single, n-noun, pl-plural).

Syntax generation: In the final step, the reversed target language of parsing rules is used to produce a sentence and it creates some target-language of words in sequence in which the meaning can be understand through the translated parse tree (Salem and Nolan, 2009). All the inflected English word-form is being integrated based on morphological features by the English generator component. It should traverse the syntactic tree in producing the outline of English noun phrase synthesizer. The generation step comprises of an English morphological synthesizer and an English noun phrase synthesizer.



Title(np(noun(تأكيد,sg), (noun(الجودة,sg)), ,prep(ل),(np2 (noun(تطوير)) , (noun (نظم,pl)),(noun(المعلومات)))

titlel(np1(noun(quality)),(noun(assurance,sg))),prep(for),(np2(noun(information,sg)))),(noun(systems,pl),(noun(development)))

Fig. 4: Compound transfer

The English morphological synthesizer is very important to produce the inflected English word through its accord of relationship between certain words. This relationship is between words in certain context such that a word in one position follows the word in a Corresponding position in some aspects: Such as number (single, plural) (Shaalan *et al.*, 2004). It reminded us earlier that the morphological analysis will analyze each word and keep all its features until it gain access to morphological synthesizer. Since in this study dealing with highly inflected language (Arabic language) (Salem and Nolan, 2009). Through this reason we have built sophisticated Arabic morphological analysis to make sure that the features access to English morphological synthesizer.

The most important role in constructing English Noun Phrase is to represent the translated noun phrase. This done through English noun phrase synthesizer. The construction depends on the syntactic category meanwhile the word is being synthesized by morphological synthesizer. This is the last phase of translation process, which it is responsible for improving, polishing and producing the English noun phrases in its right form. Finally, the parse tree is traversed into a depth-first manner to produce a list of English noun phrase.

There are three phases in noun phrase synthesis (Shaalan *et al.*, 2004). The first phase is choosing the right nouns according to its numbers and features. The second phase, the agreement of relationship between descriptive adjective and nouns should be certain for its feature. The last phase is to traverse the transformed tree to ensure the final output is well produced.

RESULTS AND DISCUSSION

In general, the aim of this experiment is to investigate whether machine translation system, namely, Google, Systran and Npae-Rbmt are sufficiently robust to be translated from Arabic noun phrases to English. The method was tested on 88 thesis titles and journals from the computer science domain.

The experiment gives the following results as shown in Table 1.

The percentage of the total score for each system has been found by dividing the total score by 880; as we have 88 test set and each is evaluated between 0-10. The score is given by human expert in translation and it measures the differences between the human translation and. Google, Systran and Npae-Rbmt systems.

Table 2 shows part of the result produced by this experiment. As seen in Table 2, the first example “ تحليل مناهج تنقيب البيانات”, the weakness of Systran system was in problems (1, 6, 2) and due to the system obtained score 7 out of 10 depending on the effect of the problem in the phrases. The weakness of Google system was in problem (1) and the system scored 9 out of 10. Npae-Rbmt system obtained score 10 out of 10 because there is no weakness translation in this example.

The following classifies the problems that appeared in machine translation from Arabic noun phrases to English:

- Synonyms of a noun: Nouns have many synonyms in both Arabic and English language. The problem that occurs here is how to match the correct words together, because some words in Arabic language may give different meaning in English and vice versa
- Order of simple noun phrases: In some cases, when translating simple phrases from Arabic to English the sentences become very weak and not understandable. This inconsistency may occur due to the fact that in English language, simple phrases are connected through “Separators”
- Successive words form an expression: This problem appeared because the successive words that form an expression are translated separately
- Translation of a preposition. In both Arabic and English language, sentences contains “prepositions”. These “prepositions” gives different meaning from sentence to sentence, depending on their position in the sentence. Therefore, when translating from Arabic to English, an inconsistency occurs resulting in weak sentences
- Conjunction with “و”: In Arabic language generally, the conjunction “و” is used to connect two noun phrases. However, in some cases an exception is made to connect two nouns. When translating from Arabic to English; an inconsistency occurs resulting in weak sentences
- Multiple word expression: Expressions are lexically, syntactically and morphologically rigid. This problem appears because the expression of this type should appear like a single word that happens to contain spaces, such as ‘الشرق الاوسط’ ‘the Middle East’ and ‘بيت لحم’ ‘Bethlehem’
- Order of the adjective: In both Arabic and English languages, there is a part of a sentence called “adjective”. In both languages the adjectives order is different. When translating from Arabic to English, an inconsistency occurs from positioning the adjective in the wrong position in the sentence, resulting in weak and not understandable sentences

Table 3 represents all type of errors returned by each of the examined system, namely, Google, Systran and Npae-Rbmt and their frequencies. For the Synonyms of noun will find that this type of problem frequented 54 times with Google, 78 times with Systran and only 9 times with Npae-Rbmt. This type of problem frequented 141times within all the system.

Table 1: Experiment results

Machine Translation (MT)	Google	Systran	Npae-Rbmt
Total score	712	530	833
Overall percentage	80.9	60.2	94.6

Table 2: Test suite

Title (SL)	Translation	English (MT)	Human translation (TL)	Problem (No)	Score
تحليل مناهج تنقيب البيانات	Analysis methods of prospection of the statements	Systran	Analysis of data mining methodologies	162	7
	Analysis of data mining method	Google		1	9
	Analysis of data mining methodologies	Npae-Rbmt			10
دراسة شاملة لتطبيقات الشبكات العصبية في البرمجة الرياضية	Complete study for nervous applications the nets in the athletic programming	Systran	Comprehensive study on neural networks applications in mathematical programming	147	6
	Comprehensive study of the applications of neural networks in the mathematical programming	Google		42	9
	Comprehensive study for neural networks applications in mathematical programming	Npae-Rbmt		4	9

Table 3: Type of problem frequencies with Arabic noun phrases to English

Problem No.	Type of problem	Total frequency	Google	Systran	Npae-Rbmt
1	Synonyms of a noun	141	54	78	9
2	Order of simple noun	37	15	13	10
3	Successive words form an expression	17	5	12	0
4	Translation of a preposition	56	21	28	7
5	Conjunction with “و”	12	5	4	3
6	Multiple word expression	13	3	10	0
7	Order of the adjective	60	19	33	8
	Total frequencies of problem	336	121	178	37

Table 1 showed that Npae-Rbmt has scored the highest percentage among all the systems. This proves that overall result is better than the benchmark system Google and Systran. The significant improvement is attributed to the use of specific rules of noun phrases. We have already developed transfer-based framework for machine translation from Arabic noun phrases to English. The patterns form Arabic to English noun phrases was newly developed well in this study. Various rules were discovered and developed in order to be able to cover more problems for Arabic noun phrases to English.

CONCLUSION

The improvement to the translation can be done only by formalizing our linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon (Abu Shquier and Sembok, 2008). However, machine translation has not been able to deliver fully automated high-quality translations. Yet there is a lot that we can do to improve the quality of MT output and increase its usefulness. In this study we presented the necessity to handle the problems of translation Arabic noun phrases into English. We proposed a rule-based approach to solve these problems. However this covers only a restricted domain to clarify the approach and in the same way the system can be completed to cover all patterns of the language. Validation rules have been applied in both the database design and the programming code in order to ensure the integrity of data. In order to get best translation this study should be merged with a comprehensive MT system that handles the ambiguity, abbreviations and the meaning problems.

REFERENCES

Abu Shquier, M. and T. Sembok, 2008. Word Agreement and ordering in english-arabic machine translation Proceeding of the International Symposium on Information Technology, Aug. 2008, IEEE Xplore Press, USA., pp: 1-10. DOI: 10.1109/ITSIM.2008.4631625

Eynde, V.F., 1993. Linguistic Issues in Machine Translation. Pinter Publishers, London, ISBN: 1-85567-024-0, pp: 239.

Farghaly, A. and K. Shaalan, 2009. Arabic natural language processing: challenges and solutions. ACM Trans. Asian Language Inform. Process. Assoc. Comput. Mach., 8: 1-22. DOI: 10.1145/1644879.1644881

Hutchins, W.J. and H.L. Somers, 1992. An Introduction to Machine Translation. Academic Press, London, ISBN: 0-12-362830-x, pp: 362.

Monem, A.A., K. Shaalan, A. Rafea and H. Baraka, 2008. Generating Arabic text in multilingual speech-to-speech machine translation framework. Mach. Trans., 22: 205-258. DOI: 10.1007/s10590-009-9054-9

Salem, Y. and B. Nolan, 2009. Designing an XML lexicon architecture for Arabic machine translation based on role and reference grammar. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Apr. 22-23, pp: 221-229. http://www.mt-archive.info/MEDAR-2009-Salem.pdf

Shaalan, K., A. Rafea, A. Mmonem, and H. Baraka, 2004. Machine translation of English noun phrases into Arabic. Int. J. Comput. Process. Orient. Languages, 17: 121-134. DOI: 10.1142/S021942790400105X

Shaalan, K., 2005. An intelligent computer assisted language learning system for Arabic learners. Comput. Assist. Language Learn., 18: 81-109. DOI: 10.1080/09588220500132399

Shaalan, K., A. Abdel Monem and A. Rafea, 2006. Arabic morphological generation from interlingua. Proceeding of the Intelligent Information Processing III, IFIP TC12 International Conference on Intelligent Information Processing, Sept. 20-23, Springer, Boston, pp: 441-451. DOI: 10.1007/978-0-387-44641-7_46

Shaalan, K., H. Talhami and I. Kamel, 2007. Automatic morphological generation for the indexing of Arabic speech recordings. Proc. Int. J. Comput. Process. Languages, 20: 1-14. DOI: 10.1142/S0219427907001561

Trujillo, A., 1999. Translation Engines Techniques for Machine Translation. Springer-Verlag, Heidelberg, ISBN: 1-85233-057-0111, pp: 220-222.