Transfer-based Arabic to English Noun Sentence Translation Using Shallow Segmentation

Namiq Abdullah
Department of Electrical and Computer Engineering, University of Duhok, Iraq

Abstract: *The quality of machine translation systems decreases considerably when dealing with long sentences. In this paper, a transfer-based system is developed for translating long Arabic noun sentences into English. A simple method used for dividing a long sentence into phrases based on conjunctions, prepositions, and quantifier particles. These particles divide a sentence into phrases. The phrases of a source sentence are translated individually. In the end of translation process, target sentence is constructed by connecting the translated phrases. The system was tested on 100 thesis long titles from the management and economy domain. The results show that the method is very efficient with most of the tested sentences.*

Keywords: *Machine translation, transfer-based approach, noun phrases, sentence partitioning*.

Received May 15, 2015; accepted March 24, 2016

1. Introduction

Machine Translation (MT) is a field of computational linguistics that aims to translate one natural language into another natural language. Having input sentence in a language (source), the MT system generates a sentence in another language (target) equivalent to the source sentence in meaning. There are many obstacles on developing a MT that conveys the complete meaning from source language to target language because of the high complexity of natural languages [3]. Nevertheless, the advancement in technology provides efficient tools for enhancing the accuracy of MT systems [4].

The major techniques used in MT systems are rule-based, statistical, and example-based techniques [7] [19]. Rule-based MT is based on linguistic information that includes morphological, syntactic, and semantic of both the source and target languages.

The statistical and example-based techniques need parallel corpora for translation [18]. A hybrid method that combines more than one technique provides better quality for the translation system [14].

There are three approaches being used for developing rule-based translation systems: direct translation, transfer-based translation, and interlingua-based translation. While the direct approach uses word-to-word translation, the transfer-based approaches apply linguistic rules for creating a transitional representation from which the target language is generated. With interlingua approaches, the source language is mapped to an abstract intermediate representation from which the target language is generated [16].

In the transfer-based approach, the translation process of an input sentence passes through three steps. First step is the syntactic analysis that produces an abstract representation of the input sentence. In second step, the abstract representation of input language is transferred to abstract representation of target language. In this step, grammatical rules of both languages are used for relating every input representation to some corresponding target representation. Last step is the generation of output sentence in target language.

Most of researchers who are interested in translation between Arabic and English concentrated on transfer-based technique for implementing their systems. Furthermore, most of these systems concentrated on noun phrases[10, 17] and verb phrases [2] rather than long sentences. Some translators are implemented for sentences on distinct domains of knowledge such as statistical [1] and interrogative [15] fields. The achieved researches on the translation of long sentences in other languages are much more than that achieved on translation to and from Arabic language [12, 13].

The objective of this paper is to implement a transfer-based machine translation system for long Arabic noun sentence into English. The source sentence is segmented into phrases by considering the prepositions, conjunctions, and other particles as shallow separators, which means that the phrases before and after a separator are syntactically and semantically separated. Investigating the structure of Arabic long noun sentences, and the role of the particles in connecting the parts of a sentence, is based on analyzing real titles of 100 M.Sc. theses in the field of management and economy. The system is tested on all the 100 titles.

2. Arabic Simple Noun-Phrase

2.1. Arabic Noun

Most nouns [8] in Arabic are derived from three-consonantal root. There are a number of affixes added to the simple nouns to indicate their definiteness, case, and number. There are two genders in Arabic: masculine and feminine. The plural in Arabic takes two forms, broken plural which has different patterns and sound plural. Sound plural uses different suffixes for masculine sound plural and feminine sound plural. The masculine sound plural is marked for nominative with ‘ون’ (مدرّسون, *modarriso:n*, teachers), and for genitive and accusative with ‘ين’ (مدرّسين, *modarrisi:n*, teachers). The suffix ‘ات’ is used for feminine sound plural.

Apart from a plural, Arabic also has a dual. This is formed with the suffix ‘ان’ for nominative nouns, whether masculine (مدرّسان, *modarrisa:n*, two teachers) or feminine (مدرّستان, *modarrisata:n*, two teachers) and ‘ين’ is used for genitive and accusative (مدرّسَين, *modarrisayn*, two teachers) for masculine and(مدرّسَتين, *modarrisatayn*, two teachers) for feminine.

2.2. Noun Phrase

A Noun Phrase (NP) is a phrase whose head is a noun or a pronoun, optionally accompanied by a set of modifiers. Arabic nouns can be modified in different ways, by demonstratives and adjectives. Arabic has two demonstratives, (هذا) and (ذلك). A demonstrative is placed before the noun it modifies. A noun modified by a demonstrative also takes the definite article (هذا الكتاب, ‘this book’).

Adjectives always have a masculine and a feminine form. Adjectives agree with the noun in gender, case, number, and definiteness:

حصانٌ جميلٌ	a beautiful horse
الحصانُ الجميلُ	the beautiful horse

Note that when the noun is definite and the adjective indefinite, the phrase is interpreted as a sentence (الحصانُ جميلٌ, the horse is beautiful).

2.3. Possessive Structure

A noun can be modified by another noun. The two nouns, head noun and modifier, form a rigid structure. The order is always the head noun followed by a modifier. The head noun is not marked for definiteness, while the modifier must be marked for definiteness:

كتابُ الرجلِ	the man’s book

Dual and plural nouns can also be modified by a genitive noun. In this case, masculine sound plurals and duals lose their normal suffix ‘ن’ (e.g., معلمو المدرسةِ, ‘teachers of the school’).

The possessor that modifies a noun can be a pronoun. In Arabic, this genitive pronoun takes the form of a suffix on the noun. The full paradigm is given in Table 1.

Table 1. Possessive pronouns (كتاب, 'book').

Person	Gender	Singular	dual	Plural
3	M	كتابهُ	كتابهما	كتابهم
	F	كتابها	كتابهما	كتابهن
2	M	كتابكَ	كتابكما	كتابكم
	F	كتابكِ	كتابكما	كتابكن
1	M, F	كتابي	كتابنا	كتابنا

A possessive structure can be modified by demonstratives and adjectives. If a demonstrative is added to the genitive modifier, it is placed before this noun:

كتابُ هذا الرجلِ	this man’s book

Adding an adjective to the genitive modifier is straight-forward, the adjective follows the genitive noun and agrees with it in the usual manner:

صاحبُ البيتِ الكبيرِ	owner of the big house

In this example, there are two nouns followed by an adjective. The adjective modifies second noun, hence the similarity between them in case and definiteness. In such phrases, diacritics are crucial in determining the modified noun. Ambiguities can still exist with using diacritics if the modified noun and inner noun are of the same gender, number, and case [6]. The following phrase can be translated to “in the yard of wide school” or “in a wide yard of the school”

في ساحةِ المدرسةِ الواسعةِ

Noun phrase can grow to include more inner nouns that may intervene between the modified noun and the modifier.

3. Connecting Particles

Arabic noun sentence can occur in two types; a single phrase and a combination of more than one phrase connected by some particles. The particles in Arabic include prepositions, conjunctions, interjections, and sometimes adverbs. Prepositions and conjunctions occur frequently in Arabic text. All prepositions in Arabic are added before nouns. Some of these are usually attached to the beginning of a word while the others are written separately [11].

Arabic uses small set of conjunctions, basically ‘و’, ‘ف’, and ‘ثم’. Although these conjunctions can be translated to English word ‘and’, each has different function that indicate the semantic relations between sentence parts. Hence, translating Arabic conjunctions into English is not an easy task. However, modern standard Arabic reduces the meanings and the functions of conjunctions as well as it concentrates on using the conjunction ‘و’ much more than the others use.

To find the most used particles and the number of their occurrences in noun sentences, titles of 100 M.Sc.

theses in the field of management and economy are investigated. The results in Table 2 show that only few of prepositions 'في', 'ل', 'من', 'على', one quantifier 'بعض', and one conjunction particle 'و' are used in the selected text. The "others" column includes prepositions 'عن', 'مع', 'بين', and 'ب' which are used very rare. The Table also shows the ratio of each particle to the total number of these particles, which are 309 particles.

Table 2. Times and ratios of particles used in the text.

في	و	ل	من	على	بعض	Others
127	58	62	33	14	7	8
41%	18.8%	20%	10.7%	4.5%	2.3%	2.6%

The investigated titles vary in length of words. The shortest title has 6 words and the longest title has 25 words with average of 13.12 words. The number of noun phrases that construct the titles varies from 2 to 7 noun phrases. The analysis of the text includes the structure of noun phrases. The longest noun phrase has 6 words of nouns and adjectives. This form occurs only one time. The noun phrase which has 5 words occurs 8 times, all with the same form of four nouns followed by an adjective (N+N+N+N+ADJ). The most used phrases are formed from two, three, or four words of nouns and adjectives.

4. System Description

To achieve the aim of the paper, a complete transfer-based translation system is implemented. The system comprises Arabic lexicon, rules database, Arabic/English dictionary, and English lexicon. The system structure is given in Figure 1.

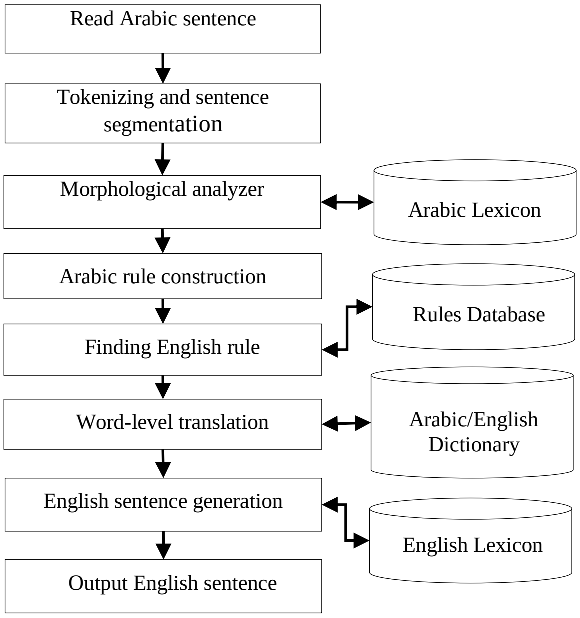

Figure 1. Overall structure of the system.

The Arabic sentence is entered to the system and it passes through the following steps:

- *Tokenizing and Sentence Segmentation*: The sentence is separated into tokens. Each token is a word or a separator. The separators divide the sentence into phrases. Each Arabic phrase is translated by itself to English phrase in a later stage of the system. For example, the sentence (أثر التضخم, في الأداء المالي 'the inflation effect in the financial performance') is divided into two phrases (أثرالتضخم, 'the inflation effect') and (الأداء المالي, 'the financial performance').
- *Morphological Analyzer*: If a word is not found in the Arabic lexicon, it will pass through a light stemming procedure. Stemming improves the performance of the system by reducing words variations [5]. Light stemming refers to a process of removing a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognize patterns and find roots [9]. Morphological analyzer is connected directly to Arabic lexicon. The lexicon holds all features of a word, which include broken plural form, part of speech (noun, proper noun, adjective, infinitive, pronoun, demonstrative pronoun, and separator), gender, and number (singular and plural).
- *Arabic Rule Constructor*: Creation of Arabic rule is the most crucial stage. The rule of the Arabic phrase is constructed from the information obtained in the previous step. For example, if the Arabic phrase to be translated is "أحمد طالب ذكي", the morphological analyzer will add the features of the words to the phrase as mentioned above. In this step, the system constructs an abstract rule expression that holds all required information for next step. The rule of this example will take the form PN+N(u)+ADJ(u) which means that our phrase is constructed from proper noun followed by undefined noun, and ended with undefined adjective. The argument 'u' is used for indefinite feature for nouns and adjectives. Other arguments that might be used in the rules are 'd' for definite, 'm' and 'f' for male and female gender feature, and 's' and 'p' for single and plural number feature.
- *Word level translator*: A direct translator gets the English words from bilingual dictionary.
- *English Rule Construction:* In this stage, the system searches the database for the English rule that matches the Arabic rule constructed in a previous step. The system has 43 rules that cover all forms of Arabic noun phrases found in the text with the corresponding English rules. English rule is the base of building English phrase in next step.
- *English Sentence Generation*: The English rule and the information got from the English lexicon are used for constructing English phrases. English lexicon contains the following features that attached with English words: plural, part of speech, gender, and number.

After constructing all English phrases, they are connected with particles to generate the English sentence. The example in Figure 2 explains these steps:

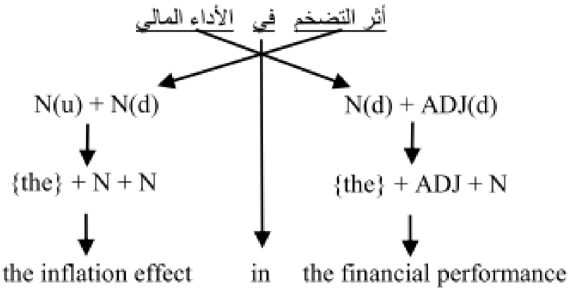

Figure 2. Steps of translating Arabic sentence.

5. Results and Discussion

The main aim of testing the translator is to find weak points in the proposed method of segmenting and translating the Arabic sentence into English. The errors that are considered in the test are two types: errors in mistranslating the meaning of the particles and errors that occur due to the context of using the particles. For achieving the aim of the test, 100 titles of the M.Sc. theses mentioned above are translated by the system and the results are summarized in Table 3.

The results show that most of the particles are translated accurately. With the conjunction 'و' one error occurs 8 times, and with the preposition 'ل' also one error occurs 9 times.

Table 3. Results of testing the system.

	في	و	ل	من	على	بعض
accurate	127	50	53	33	14	7
inaccurate	0	8	9	0	0	0

The conjunction "و" is normally used to connect two noun phrases. Sometimes it is used to connect two nouns in the same phrase. The statement " تكنلوجيا المعلومات و الاتصالات" is translated to "information technology and the communications" instead of "information and communications technology". On the other hand, the preposition 'ل' has two meanings, 'for' which is used by the system dictionary and produces 53 correct translations and 'of' which occurs 9 times in the tested sentences. There are some examples in Table 4 that show the output of the system for sentences include these particles.

Apart from the tested sentences, some other problems may occur due to the contexts in which a proposition is used. The preposition 'في' is sometimes gives more accurate meaning when translated to 'for' instead of its normal meaning 'in'. The same thing can be said for the preposition 'من' which can be translated to 'of' instead of its normal meaning 'from'.

Table 4. Results of translating Arabic sentences.

Arabic Sentence	دور تكنلوجيا المعلومات والاتصالات في تحقيق جودة الخدمة المصرفية
Translated Sentence	The role of information technology and the communications in achieving the quality of banking service
Arabic Sentence	أثر بعض مؤشرات فاعلية نظام المعلومات الإدارية في إقامة متطلبات نظام التحسين المستمر : دراسة في مديريات المؤسسة العامة لشؤون الألغام في العراق
Translated Sentence	The effect of some of the indicators of administrative information system effectiveness in setting the requirements of continuous improving system : a study in directorates of general establishment for the mines affairs in Iraq
Arabic Sentence	إمكانية تأهيل الموارد البشرية لمتطلبات الإدارة الألكترونية : دراسة إستطلاعية في عينة من المصارف التجارية في محافظة دهوك
Translated Sentence	the ability of qualifying the human resources for requirements of electronic management : an exploratory study in a sample of the commercial banks in governorate of Duhok

6. Conclusions

Machine translation systems cannot produce accurate translation as the human do. The problems increase as the sentence length increases. In this work, a transfer-based MT system is implemented for translating a long noun sentences from Arabic to English. Real titles of 100 theses from management and economy field are considered for analyzing noun sentences. The noun sentence is segmented into noun phrases separated by prepositions, conjunctions, or quantifiers and the separated phrases are translated individually.

The system gives one meaning for each particle, which is the most used meaning. The results of testing the system show that this method is efficient with most of the particles used in noun sentences. Two problems occur with two particles, the conjunction 'و' and the preposition 'ل'. The quality of translation can be improved with more investigation of sentence structure and word morphology and probably these improvements can be implemented in programming level of the system. The same method can be applied on other patterns of the language, such as verb sentence. However, there are more particles used in verb sentences. The systems that are implemented for different patterns of the language can be combined together for implementing more comprehensive system.

References

[1] Agiza H., Hassan A., and Salah N., "An English-to-Arabic Prototype Machine Translator for Statistical Sentences," *Intelligent Information Management*, vol. 4, pp. 13-23, 2012.

[2] Algani Z. and Omar N., "Arabic to English Machine Translation of Verb Phrases Using Rule-Based Approach," *Journal of Computer Science*, vol. 8, no. 3, pp. 277-286, 2012.

[3] Costa-Jussa M., Farrus M., Marino J., and Fonollosa J., "Study and Comparison of Rule-

based and Statistical Catalan-Spanish Machine Translation Systems," *Computing and Informatics*, vol. 31, vol. 2, pp. 245-270, 2012.

[4] Dastjerdi H. and Ghobadi S., "Does Technology Help? Google Translation Vs. Human Translation," *Intercontinental Journal of Educational Research*, vol. 2, no. 1, pp. 1-6, 2012.

[5] Dwivedi S. and Sukhadeve P., "Translation Rules for English to Hindi Machine Translation System: Homoeopathy Domain," *The International Arab Journal of Information Technology*, vol. 12, no. 6A, pp. 791-796, 2015.

[6] Hoyt F., *Lemma in The Encyclopedia of Arabic Lnaguage and Linguistics*, Brill, 2008.

[7] Hutchins J., "Machine Translation: A Concise History," Computer Aided Translation: Theory and Practice, Chinese University of Hong Kong, 2007.

[8] Kremers J., *The Arabic Noun Phrase: A Minimalist Approach*, Utrecht: LOT, 2003.

[9] Larkey L., Ballesteros L., and Connell M., "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-Occurrence Analysis," *in Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, pp. 275-282, 2002.

[10] Mohamed A., Machine Translation of Noun Phrases: From English to Arabic, Thesis, Cairo Unversity, 2000.

[11] Nwesri A., Tahaghoghi S., and Scholer F., "Stemming Arabic Conjunctions and Prepositions," *in String Processing and Information Retrieval*, Springer, 2005.

[12] Oliveira F., Wong F., and Hong I., "Systematic Processing of Long Sentences In Rule Based Portuguese-Chinese Machine Translation," *in Computational Linguistics and Intelligent Text Processing*, Springer Berlin Heidelberg, 2010.

[13] Roh Y., Seo Y., Lee K., and Choi S., "Long Sentence Partitioning using Structure Analysis for Machine Translation," *in Proceedings of Natural Language Processing Pacific Rim Symposium*, Tokyo, pp. 646-652, 2001.

[14] Sangeetha J., Jothilakshmi S., and Kumar R., "An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism," *International Journal of Engineering and Technology*, vol. 6, no. 4, pp. 1909-1919, 2014.

[15] Shaalan K., "Machine Translation of Arabic Interrogative Sentence into English," *in Proceedings of the 8th International Conference on Artificial Intelligence Applications (Egyptian Computer Society*, Egypt, pp. 473-483, 2000.

[16] Shaalan K., "Rule-Based Approach in Arabic Natural Language Processing," *International Journal on Information and Communication Technologies*, vol. 3, no. 3, pp. 11-19, 2010.

[17] Shirko O., Omar N., Arshad H., and Albared M., "Machine Translation Of Noun Phrases from Arabic to English Using Transfer-Based Approach," *Journal of Computer Science*, vol. 6, no. 3, pp. 350-356, 2010.

[18] Somers H., "Review Article: Example-Based Machine Translation," *Machine Translation*, vol. 14, no. 2, pp. 113-157, 1999.

[19] Tripathi S. and Sarkhel J., "Approaches to Machine Translations," *Annals of Library and Information Studies*, vol. 57, pp. 388-393, 2010.

Namiq Abdullah is a lecturer in Electrical and Computer Engineering at the University of Duhok, Iraq. Prior to joining Duhok University, he worked many years at University of Mosul in Iraq and Al-Hussein Bin Talal University in Jordan. He obtained his M.Sc. degree in Computer Engineering from University of Technology/Iraq (1993), and his B.Sc. in Electronics and Communications Engineering from University of Mosul/Iraq (1979). His research interests include Natural Language Processing and Microprocessor based systems.