Journal of University of Duhok., Vol. 21, No.1 (Pure & Eng. Sciences), Pp 115-120, 2018
DOI: https://doi.org/10.26682/sjuod.2018.21.1.11

ARABIC-ENGLISH TRANSLATOR OF SPECIFIC NOUN PHRASE WITH TRANSFER-BASED APPROACH

NAMIQ SULTAN ABDULLAH*
Dept. of Electrical and Computer Engineering, College of Engineering, University of Duhok, Kurdistan Region-Iraq

(Received: April 4, 2018; Accepted for Publication: July 9, 2018)

ABSTRACT

Machine translation between languages which have different morphology and syntactic features is generally a complex task. Rule-Based Machine Translation is machine translation based on linguistic information about source and target languages. This approach needs to design precise rules for both the source and the target languages in order to generate a correct translation. This paper presents a transfer-based approach for translating Arabic noun phrase to the English language. The idea is to start with noun phrases that are not including conjunctions, prepositions, and quantifier particles. The system was tested on 80 phrases taken from MSc thesis titles of management and economy domain. The system was evaluated by professional human evaluators. The accuracy of the result was 96.2%.

KEYWORDS: machine translation, rule-based approach, Arabic Language Processing, noun phrase

1. INTRODUCTION

Rule-Based Machine Translation is a classical approach of machine translation (MT) systems based on linguistic information that includes morphological, syntactic, and semantic of both the source language (SL) and target language (TL). The main approach of rule-based machine translation systems is based on linking the structure of the given source sentence with the structure of the target sentence, keeping the meaning of source sentence unchanged. There are three approaches being used for developing rule-based translation systems: direct translation that uses word-to-word translation, transfer-based approach that applies linguistic rules during the translation process, and interlingua-based translation that maps the SL to an abstract intermediate representation from which the TL is generated [1][2]. Vauquois triangle shown in Figure 1 is used to describe the different approaches of rule-based machine translation [3]. The bottom of the triangle is the direct approach and the top is the interlingual approach. It shows the increasing depth of analysis required as we move from the direct approach through transfer approach to interlingual approach. According to the model, each step up the triangle required greater effort in SL analysis and TL generation but reduced the effort involved in conversion between languages.

There are many advantages of the rule-based approach. No digitized texts are required for implementing the translation system. It is domain independent, therefore, same rules are applied to every translator. It is easy to debug the system for errors, as well as it is easy to extend the system by adding more rules. Furthermore, the linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system [4][5].

Most researchers in Arabic MT are concentrating more in English to Arabic translation because Arabic language is rich in morphology and it is difficult to define parts of speech of the language [6][7][8]. At present, there is not much work on Arabic noun phrases to English MT [9]. Shirko et al.

* E-mail: namiq.sultan@uod.ac

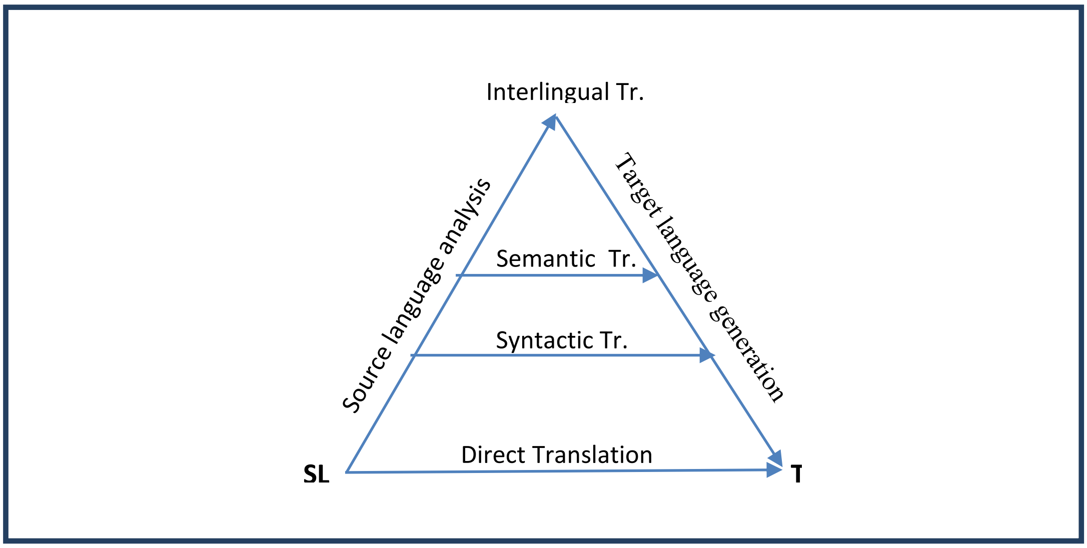

Fig. (1): Types of rule-based machine translation

(2010) developed an MT that translates Arabic noun phrases into English by using the transfer-based approach. The system is tested on noun phrases from the domain of computer science with an accuracy of 94.6% [9]. The score is given by a human expert in translation. Algani and Omar (2012) introduced a rule base translation system for Arabic verbal sentence of scientific text to English using transfer approach. The accuracy of the result of the designed system is 93% [10]. The methodology used for evaluating the system is based on making a comparison between the outputs of the designed system for the test examples and the human translation for the input sentences. Abo Shoqair et al. (2017) presented a transfer-based approach system to handle the translation of modern standard Arabic into English. The system has scored the accuracy of 96.6% [11]. It is not clear how the final score is calculated during the evaluation of the system accuracy.

The present work addresses the translation of Arabic noun phrase into English by using the transfer-based technique which is an important part of implementing a more general MT system. We concentrate here on noun phrases that are not including conjunctions, prepositions, and quantifier particles. The fundamental principles behind the design of our system are: (1) that it is possible to construct rules for different noun phrases of the SL and the corresponding phrases of the TL; (2) that the constructed rules are stored in a database and used during the process of translation; (3) that some post-processing can be used for finalizing the process of translation.

The rest of the paper is organized as follows. Section 2 gives an introduction to Arabic noun phrase. Section 3 explains the method of collecting and analyzing the noun phrases for constructing the rules that are used by the system. Section 4 describes the architecture of the system. Section 5 presents the evaluation method and discusses the results of the system. Section 6 concludes the paper.

Arabic Noun Phrase

The Arabic noun phrase (NP) consists of a head noun or a pronoun, optionally accompanied by one or more modifiers. Modifiers can be demonstratives, adjectives, nouns, and pronouns [12][13]. Nouns in Arabic may be definite or indefinite. The existing of the definite article (*al* "the") at the beginning of the noun is used for marking the definiteness feature of the word.

- **Demonstratives**: Arabic has two demonstratives (هذا, *hādhā*, "this") and (ذلكَ, *dhālika*, "that"). There are different forms of these demonstratives for single, plural, feminine, and masculine features. The demonstrative placed before the noun it modifies (هذا الكتابُ, *hādhā -l-kitābu* "This book"). If the definite article is omitted from the noun, the phrase (هذا كتابٌ, *hādhā kitāb-un* "This is a book") is considered as a sentence of a subject (مبتدأ, *mubtada'*) and a predicate (خبر, *khabar*). **Adjectives**: Adjectives always have a masculine and feminine forms. They agree with the noun in gender, case and number (كتابٌ كبيرٌ, *kitāb-un kabīr-un* "a large book", كتابان كبيران, *kitābāni kabīrāni* "two large books", كتبٌ كبيرةٌ, *kutub-un kabīrat-un* "large books").

• **Nouns**: A noun can be modified by one genitive noun in a form called genitive construct (إضافه, *iḍāfa*). The order of the two nouns in the genitive construct is always the head noun followed by the modifier noun. The head noun does not take definite article while the modifier must be marked for definiteness (كتاب الطالب, *kitābu -l-ṭālibi* "a student's book").

Nouns and adjectives in Arabic have singular, dual and plural forms. In Arabic we need only to add two letters, "ان *ān*" in the nominative case and "ين *aīn*" in the accusative and genitive cases, to the end of the singular form to express the dual form (كتابان, *kitāb-ān "two books"*). Plurals are of two types: broken plural and sound plural. Broken plural has no fixed rule for making it. Sound plurals are of two types: masculine sound plural and feminine sound plural. The masculine sound plural is formed by adding two letters to the end of the singular form, "ون *ūn*" in the nominative case and "ين *īn*" in the accusative and genitive cases. Feminine plurals are formed by adding the two letters "ات *āt*" to the end of the singular form of the word [14].

2.DATA COLLECTION AND RULES BUILDING

The first step for implementing the noun phrase translator from Arabic to English language is the collection of Arabic noun phrases from real texts. We investigated and analyzed 100 titles of MSc thesis in the field of management and economy. The investigated titles vary in length of words and number of noun phrases. The shortest title has 6 words and the longest title has 25 words. The considered noun phrases in this work are those which are not including prepositions, conjunctions, and quantifier particles. The investigated titles vary in length from 2 to 7 noun phrases and the total noun phrases are about 400 in all the titles.

The analysis of the text includes the structure of noun phrases. The longest noun phrase in the text, which occurs only one time, has 6 words of nouns and adjectives. 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.

The words are classified into two types: nouns and adjectives. There are three types of nouns: common noun (N), proper noun (PN), and infinitive (INF). The ordinal numbers are treated like adjectives for decreasing the number of rules. All nouns and adjectives can be defined or undefined words. Then we recognized 38 types of phrases which needs 38 rules.

The words order of Arabic language is different from the English. To change the words order in the translation process, the rule terms are joined with numbers that show the position of the word in the phrase. For example, the Arabic phrase (الصفُّ الثالثُ الأبتدائيُّ, *al-ṣaffu -l-thālithu -l-ʾibtidāʾiyu*) is translated to English phrase (the primary third class). The noun that occurs at beginning of Arabic phrase moves to the end of the English phrase and the adjective occurs at the end of Arabic phrase moves to the beginning of the English phrase. The rules that satisfy the translation of this phrase are (N1(d) ADJ2(d) ADJ3(d)) for the source phrase and ({the} ADJ3 ADJ2 N1) for the target phrase. The definiteness feature of the words in the Arabic phrase is denoted by a letter inside parenthesis; (u) for indefinite and (d) for definite words. The other features that are included in the Arabic rule are "m" and "f" for male and female gender, and "s" and "p" for single and plural number. These features are stored in an Arabic lexicon to be used during the morphological analysis of the Arabic words. In Table 1, samples of some rules are listed with examples of the source and target languages.

Table (1): Samples of phrases and their rules

Arabic phrase	Arabic rule	English rule	Translated phrase
دور المهارات الريادية	N1(u)N2(d)ADJ3(d)	{the}N1{of}ADJ3N2	The role of pioneer skills
تقدير أهمية الموارد البشرية	INF1(u)N2(u)N3(d)ADJ4(d)	INF1{the}N2{of}ADJ4N3	Estimating the importance of human resources
دور مستويات التفكير القيادي	N1(u)N2(u)N3(d)ADJ4(d)	{the}N1{of}ADJ4N3N2	The role of leading thinking levels
إنتاج المياه المعدنية	INF1(u)N2(d)ADJ3(d)	INF1{of the}ADJ3N2	production of the mineral water

2. SYSTEM DESCRIPTION

The translation system comprises Arabic lexicon, English lexicon, Arabic-English rules database, and Arabic-English word dictionary. The Arabic lexicon, English lexicon, and the Arabic-English dictionary are built by the author manually for satisfying the needs of this research. The Arabic lexicon structure includes the following properties for the word: part of speech (POS), gender, number, and plural. The POS property of the word has five types: noun, proper noun, adjective, infinitive, and pronoun. The gender property of the word is either masculine or feminine. The number property shows the case of the word if it is single or plural. The plural property is used to store the broken plural form of the word. In Arabic language, the broken plural is an irregular plural form of a noun or adjective. Therefore, it is difficult to use morphological techniques to derive broken plural.

The English lexicon includes two properties of each word: the POS which has same types of Arabic POS, and the plural form of the English word. The Arabic-English dictionary is a simple dictionary includes two fields: Arabic word and the corresponding English word.

The 38 rules that are produced from the analysis described in the previous section are arranged in a database table of two fields: Arabic rule and English rule. These rules cover the patterns required for translating all the phrases that are used in the titles of the theses. The overall system structure is shown in Figure 2. The Arabic phrase is entered into the system and it passes through the following steps:

Tokenization: The input Arabic phrase is separated into a list of words.

Morphological Analyze: The POS for each word in the list is determined by using the Arabic lexicon. If the word is not found in the Arabic lexicon, it will pass through a light stemming procedure. Light stemming refers to a process of removing a small set of prefixes and/or suffixes, without trying to deal with infixes[15]. The stem generated from this stage is associated with the properties found in the lexicon as well as features found during word analysis.

Arabic rule construction: 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 (أحمدُ طالبٌ ذكيٌّ, *Ahmed ṭālib-un dhakiy-un* “Ahmed is a clever student”), the system constructs an abstract rule expression that holds all required information. The rule of this example will take the form PN1 N2(u) ADJ3(u).

Word level translation: A direct translator gets the English words from the Arabic-English 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. English rule is the base of building English phrase in next step.

English phrase 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: part of speech, gender, number, and the plural form of the word.

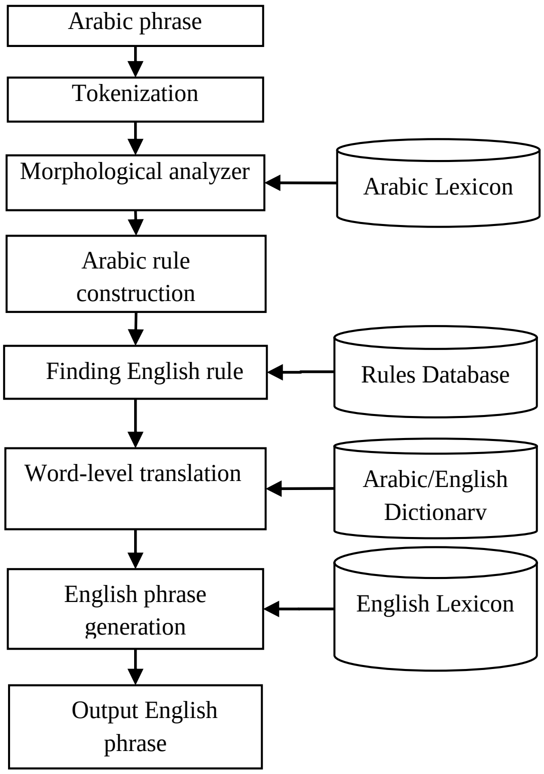

Figure 2: Overall structure of the system

3. RESULTS AND DISCUSSIONS

A major design goal of this system is to implement an MT for Arabic noun phrase to English. The system can be used as a base for more general long sentences translator. In order to evaluate the accuracy of our system, we used the human experience for comparing the exact translation with the results of the system. We selected 80 phrases randomly from 100 titles of the MSc thesis that include more than 400 phrases. The phrases are translated by the system and sent to three specialists in English language. The mother language of all the evaluators is Arabic and they hold high degrees in English language. Each evaluator reads the Arabic phrase and its English machine translation then he compares the translation with what emerges in his mind as the best translation. Consequently, the evaluator gives a score out of 10 to each phrase. The researcher sums the scores for each evaluator to get a score out of 100. After that the researcher sums the resulting scores for each evaluator and divides the total score by the number of evaluators. The system got an average mark of 96.2%. The problems that decrease the accuracy of the system can be classified into two types:

1) English word has different meanings, therefore it is impossible to be sure that the system selects the exact word from the dictionary. To decrease the effect of this problem, the system should be constrained on a special field of application. The semantic analysis also can be incorporated into the system to help in solving this problem.

2) Compound words are written in Arabic with a space in the middle even though they should be compound to produce single English word (رأس المال "capital"). This problem can be solved by joining the compound Arabic words with special characters, such as the hyphen, "-", character, so that the system can deal with these compound words as one item.

Table 2 shows some of the phrases translated by the system along with the human translation. The phrases are selected from the ones that received low scores in order to identify the problems behind the inaccuracy of the translator**.**

Table 2: Samples machine and human translated phrases

Arabic phrase	Machine translation	Human translation
التقويم المالي	the financial evaluation	financial assessment
الرضا الوظيفي	the functional satisfaction	job satisfaction
القوى العاملة	the operating power	man power/working force
تبني التسويق الألكتروني	adopting of the electronic marketing	the adoption of e-market
شؤون الألغام	affairs of the mines	mines affairs
دور مخرجات نظام المعلومات الإستراتيجية	outputs role of the strategic information system	the role of strategic information system outputs
قرار النفط مقابل الغذاء	resolution of the oil versus the food	oil-for-food resolution
كليات التعليم التقني	colleges of technical teaching	college of technical education

4. CONCLUSIONS

In this paper, an Arabic noun phrase to English MT system has been designed and implemented using transfer-based approach. The system receives Arabic noun phrase as an input. Morphological analysis applied to determine the part of speech and other features of each word of the phrase. The necessary rules have been applied to identify their structural representation which can be used to identify the English structure representations. Then the target English phrases can be generated from such representations after applying the required grammar rules by considering the relational grammar of both Arabic and English. There are several reasons that make transfer-based approach desired by MT community. The designed system can be very well incorporated with general MT systems for Arabic text.

Acknowledgment

I would like to thank Dr. Khalid M. Ismail (aburami3@qu.edu.qa), Dr. Salim Y. Fathi (salem _ yahya @yahoo), and Mr. Mahfood K. Mahmood (mahfood63j@hotmail.com) for accepting the task of evaluating the results of the system.

REFERENCES

-S. Tripathi and J. K. Sarkhel, “Approaches to machine translation,” *Ann. Libr. Inf. Stud.*, vol. 57, no. December, pp. 388–393, 2010.

-A. Alqudsi, N. Omar, and K. Shaker, “Arabic machine translation: a survey,” *Artif. Intell. Rev.*, vol. 42, no. 4, pp. 549–572, 2012.

-S. Bessou and M. Touahria, “Morphological Analysis and Generation for Machine Translation from and to Arabic,” *Int. J. Comput. Appl.*, vol. 18, no. 2, pp. 14–18, 2011.

-M. R. Costa-Jussà, M. Farrús, J. B. Marino, and J. A. R. Fonollosa, “Study and comparison of rule-based and statistical catalan-spanish machine translation systems,” *Comput. Informatics*, vol. 31, no. 2, pp. 245–270, 2012.

-K. Shaalan, “Rule-based Approach in Arabic Natural Language Processing,” *Int. J. Inf. Commun. Technol.*, vol. 3, no. 3, pp. 11–19, 2010.

-S. K. El-aini, K. Shaalan, A. Rafea, A. A. Moneim, and H. Baraka, “Machine translation of English noun phrases into Arabic,” *Int. J. Comput. Process. Lang.*, vol. 17, no. 02, pp. 121–134, 2004.

-A. A. El-Monem, “Machine Translation of Noun Phrases from English to Arabic,” Faculty of Engineering, Cairo University, Giza, 2000.

-M. M. Abu Shquier and T. M. T. Sembok, “Word agreement and ordering in English-Arabic machine translation,” *Proc. - Int. Symp. Inf. Technol. 2008, ITSim*, vol. 1, no. May 2014, 2008.

-O. Shirko, N. Omar, H. Arshad, and M. Albared, “Machine translation of noun phrases from Arabic to English using transfer-based approach,” *J. Comput. Sci.*, vol. 6, no. 3, pp. 350–356, 2010.

-Z. A. Algani and N. Omar, “Arabic to English machine translation of verb phrases using rule-based approach,” *J. Comput. Sci.*, vol. 8, no. 3, pp. 277–286, 2012.

-M. M. Abu Shquier, M. S. Atoum, and O. M. Abu Shqeer, “Arabic to English Machine Translation,” in *Proceedings of the New Trends in Information Technology (NTIT)*, 2017, pp. 118–124.

-J. Kremers, *The Arabic noun phrase: A minimalist approach*. Utrecht: LOT. 2003.

-L. Alkhazy, “Noun Phrases in Arabic: a Descriptive Study of Noun Phrases in Modern Standard Arabic and Najdi Arabic A,” Master Thesis in California State University, Northridge, 2016.

-J. Kremers, “Adjectival agreement in the Arabic noun phrase,” *Proc. Console XI. Available online http//www. sole. leidenuniv. nl/index. php3*, 2003.

-L. S. Larkey, L. Ballesteros, and M. E. Connell, “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*, 2002, pp. 275–282.