ISSN XXXX XXXX © 2019 IJESC

Research Article Volume 9 Issue No.1

Practical English to Arabic Machine Translation System by Applying Rules-Based Method

Dr.Khaled Elmenshawy[1], Ahmed El-Nady[2]
Department of Computer Science
Elshourok Academy, Cairo, Egypt[1]
Al-Azhar University, Cairo, Egypt[2]

Abstract:
English-Arabic Machine translation systems have been taking place in machine translation projects in recent years and so, many projects have been carried out to improve the quality of translation into and from Arabic. This research focuses on machine translation from the source language (English) to the target language (Arabic) using an English-Arabic electronic dictionary. The challenges of this research are the difficulty of delivering the appropriate meaning of the source language (SL) in the target language (TL), different sentence structure between two languages, word agreement, ordering problem, verbal forms and linguistic structures. The aim of this research was to design and build an automatic translation system from English to Arabic based on a dictionary of English roots using rule-based method. The proposed machine translation system uses a transfer strategy which is divided into three phases: analysis, transfer and generation of sentences in the target language. The system was evaluated by selecting a set of English language sentences to cover all the structures of sentences as a first stage, and then selecting another set of long sentences that contained more than one structure. All the results of the system were compared with the results of the various translation web on the Internet. that the most of natural language processing researchers build their systems on small or virtual dictionaries containing a few words and reserve few number of properties of those words.

Keywords: Machine translation, rule-based approach, Arabic language, English language, sentence structure, morphological analysis, to kenization

I. INTRODUCTION

Natural Languages Processing is an area that combines computer science and linguistics. It focuses on the problem of natural language formulation by computer. The natural language can be understood by combining knowledge of grammatical formulation and ability to understand and clarify logical relationships. This differs from the systems that are based on the keywords that called model harmonization systems in which the given models are to be compromised with the stored ones [1],[2]. Natural languages processing systems represent the ways in which the computer communicates with users of different natural languages. Figure (1) shows the main components of the machine translation systems.

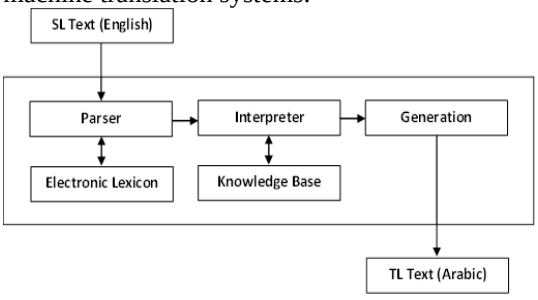

Figure.No:1

It is noted that the most of natural language processing researchers build their systems on small or virtual dictionaries containing a few words and reserve few number of properties of those words. Moreover, the study of the contents of these dictionaries confirms that there is no agreement with the nature of information that must contain the lexicon or how to represent them on the computer. Most machine translation systems deal with the lexicon as a necessary evil. So, most of these systems have not been given enough attention to work in the lexicon, either in theoretical view or in programming (2]. Grammar forms and the modern grammar in general emphasize on the importance of the lexicon. Therefore, we will try to focus on what data is required in the lexicon and how to represent it on the computer [3]. The above is one of the motivations and the other is dealing with the wide range of problems which faced machine translation systems including the problem of choosing the appropriate meaning of the word, the problem of the timing, and the problem of the source language text which contains many derivations, proper names, compound words and proverbs. The main aim of this paper is to create a model that combines the necessary various lexical resources for any translation system to form linguistic databases as basic units in the natural languages processing [6]. The lexicon resources include an English-Arabic dictionary, Arabic grammar, English grammar, conjugations, irregular verbs, affixes, etc., how to be represented on the computer and relationship among them, and using the latest information systems technologies to deal with these lexical resources with focusing on three main themes which are:

1. Existence of a full bilingual dictionary containing most of the required linguistic properties of any translation system for each word in the dictionary.
2. Affixes [prefixes and suffixes) for English and Arabic words and their rules.

3. Terms and compound words in the texts and how to detect and translate them.

Using computers to translate text or speech from one spoken human language (source language SL) to another spoken human language (target language TL) is what known as Machine Translation MT. Machine Translation has been defined as "an automatic process that translates from one human language to another language by using context information". In other words, machine translation is the translation of natural languages using a machine [7], [8]. Machine translation has many advantages that make it preferred for more than an accurate human translation. First, Machine translation systems are fast in providing output, which makes thema powerful tool. In fact, Machine translation systems can provide low quality translation in situations where any translation is better than no translation at all or where low quality translation of a large amount of documents delivered in minutes, is more useful than optimal translation in a few weeks. Furthermore, Machine translation systems produce a translation without bias, which can happen with human translators . Finally, Machine translation is significantly cheaper. Unlike human translation, Machine translation is a onetime cost i.e., the cost of the too land its translation [9].

SENTENCE STRUCTURE IN ARABICLANGUAGE:

One of the biggest challenges when translating to and from Arabic is mastering sentence structure. Unlike many European languages, Arabic's structure is not at all similar to English [10], for instance the default sentence structure for the Arabic sentence is VSO[11], whereas English follows the structure of SVO [12]. Arabic word order, parts of speech and grammar are also different. In Arabic, OSV and SOV sentence structures are not acceptable as a word order. The classical classification of Arabic sentences is: nominal sentences for the sentences containing no verb and verbal sentences that consist of a verb. For the verbal sentences structure, the different word orders of Subject-Verb-Object (SVO),Verb-Subject-Object (VSO), Verb-Object-Subject (VOS) and Object-Verb-Subject (OVS), are all acceptable in Arabic [13] [14].

COMPARISON BETWEEN ARABIC AND ENGLISH SENTENCE STRUCTUR:

As mentioned previously, Arabic has many various word orders. Whereas, in English sentences must put the subject first the verb second and finally the object. In the following section, we will present different examples of Arabic sentences and compare them with their English translations.

II. SYSTEM DESIGN AND ARCHITECTURE:

The overall process of the translation of English text into Arabic system based on a Rule Based approach is presented in Fig. No: 2. Generally, the system elements can be divided into three stages: analysis of the source language (English), transfer between two languages and generation of the target language (Arabic).However, the system involves the following three stages:

THE FIRST STAGE: ANALYSIS OF THE SOURCE LANUAGE (ENGLISH):

When the text to be translated from the source language "English" is entered into the target language "Arabic" into a text box and presses the Translate button, it is a sign of the beginning of the machine translation process. The first stage is the processing stage of the text to be translated. The text form and the lexical analysis and the loading of the text records of the main database of the lexicon as shown in Fig. No:3.

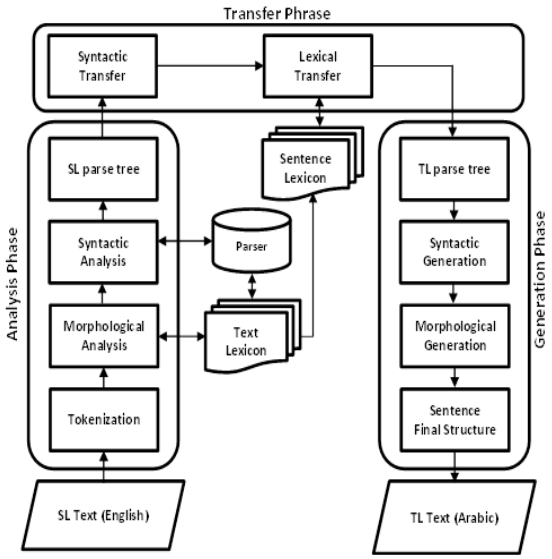

Figure.No:2 Process of the translation of English text into Arabic

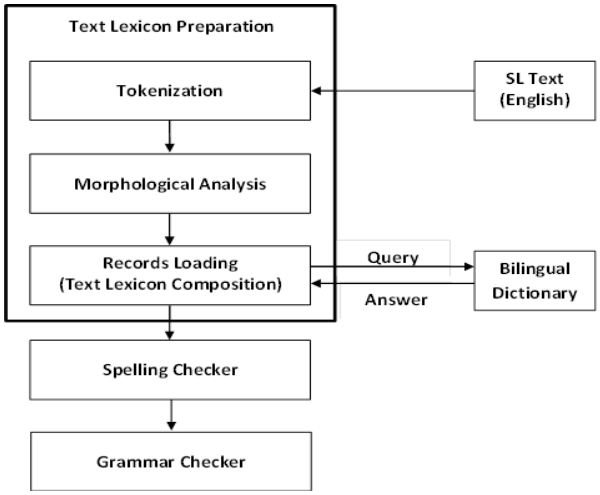

Figure. No. 3 Steps to Prepare a Text Dictionary

The stage of building the preparation of the text dictionary and testing the text consists of several steps:

Step (1): Tokenization: Tokenizing is a pre-process task and it is done by dividing the English input sentence into individual words as preparatory step for the following process. E.g. the tokenizing outcome of the sentence such as "He is a boy." is as follows: ["he", "is", "a", "boy"].

After entering the English text in the proposed machine translation system, it is divided into tokens on the basis of separating signs between words such as spaces, punctuation marks ... etc. where the **following processes are performed:**

1. Remove the special signs contained in the English text except the signs which may be used later, such as ";", ",", ".", ",", "/".
2. Determine the end of each sentence in the text and replace the end of the sentence with "#".

3. All words are converted to lowercase letters and the special signs are separated from the words.
4. The text is divided into sentences based on the positions of "#" sign.
5. Each sentence is separately processed. If a sentence contains a compound word such as verbs, adjectives, auxiliaries, propositions, conjunctions. This compound word is assembled into a single unit. The apostrophe is rewritten to standardize Form of writing to return to its origin.
6. The text is collected again and it is divided into Tokens.

Step (2): Morphological analysis: In this step, the morphological analyzer analyses each word of the English input text morphologically applies certain rules before implementing the derivation rules [15). Morphological rules depend on the input word features, such as gender and person features of the noun and the part of speech as well as the verb/adjective category. All of these features should be taken in account in order to get the correct derivation rules [16]. By applying the morphological rules, the input words are returned to their roots after the removing of prefixes or suffixes or both and then their effects are studied on the root and its meanings [17]. Entering Tokens into the lexicon that searches the lexicon for the root (s) of these words to obtain it if it exists in the dictionary or otherwise appears as it is.

Step (3): Lexicon composition: This step is for text dictionary composition, i.e., Load the records of the text tokens to the main internal database to create a lexicon Text. The English-Arabic Bi-lingual dictionary consists of 26 external databases where each one contains the words that start with an English alphabetic letter. According to the initial letter of each root, this letter database is loaded. The root record or records are uploaded to the main internal database and so for each token of the text. Therefore, all records of the English input text have been uploaded to the internal main database. All processes will be done on these records so it is called the "lexicon text".
Step (4): Spelling checker for the words of the English input text. An error message is to be displayed if the token does not exist.
Step (5): Grammar checker for each sentence in the English input text. If the text is free of common grammatical errors, the next stage begins.
For example: Suppose that the text to be translated is the **following English sentence:**
He won't play tennis.
After text reformation:
he will not play tennis.
After morphological analysis:
he will not play tennis.
The word "he" is checked and begins with the letter "h". The records of this word is stored in letter h database. The records of this database are uploaded to the internal database called "letter H". If the record or records of word "he" are loaded into the internal database, it will be transferred to the Main Internal Database and so for each word of the text.
The records of the processed sentence are loaded from the text lexicon into the temporary internal database to form the special lexicon called sentence lexicon. In this lexicon. From this limited number of records, the system identifies the meanings of the words, the part-of-speech of each word and other features such gender, location and appropriate use. Some of these features are used in the next stages such as the process of selecting the appropriate meaning of the word.

An example is the following English sentence:
The boys are playing a football.
We find that the word "playing" after removing its suffix "ing". It becomes "play". By searching in the lexicon, we find two meanings " يلعب – يعزف" for the word "play" when the part-of-speech is "verb" and other meaning "مسرحية" when part-of-speech is "noun". And then the system chooses the appropriate part-of-speech as a result of the syntactic analysis.
Step (6): Syntactic analysis: Syntactic analyzer utilizes the lexicon and grammar rules to check the English input text in terms of spelling and grammar then this information is used to produce the analysis of the text structure as an output (Parsing process). This process starts by assigning all possible POS for each word in English input sentence. After that it uses the rules to choose the POS which is suitable for combining of the all sentence words correctly. The next process is converting the English input sentence into a special data structure tree. TABLE NO. 2 shows the output sample of the syntactic analyzer**.**
The POS produced by Syntactic analyzer can be shown in Table 1.

Table.1. syntactic analyzer POS output

English word	Possible POS
The	Det
Boys	Noun
Are	Aux
Play	Noun, Verb
football	Noun

The final structure design created by syntactic analyzer can be represented as the following: The parser tree is:
The parser tree is: sentence(clause1(subclause3(phrase1(noun_phrase2("The","boys")),phrase2(verb_phrase2("are","playing")),phrase1(noun_phrase2("a","football")))))
POSList is: ["det", "noun", "aux", "verb", "det", "noun"]

THE SECOND STAGE: TRANSFER BETWEEN TWO LANGUAGES:
To specify the syntax of the sentence in the target language, there are two steps from the output of the sentence analysis in the source language:
Step (1): Structure transfer: This step deals with the structure and patterns of the target sentences. The task of this step is lining up the words of the target Arabic sentence based on the Arabic grammars rules.
Step (2): Lexical transfer: This step is for dictionary translation. The task of this step is using the English-Arabic Bi-lingual dictionary to look up the Arabic meaning for each word in the English parse. This process is done word by word maintaining the same order as the English source phrase. The output of this step is a list of English tokens and this equivalent Arabic translation.

THE THIRD STAGE: GENERATION OF THE TARGET LANGUAGE (ARABIC):
In this stage, the target language sentence is produced after passing through a series of analysis processes and applying the translation rules and Arabic grammar [17]. The target language sentence will be in its final version which should be correct in

terms of its grammar structure and meaning translation. There are two steps to be done in the generation stage which are: morphological generation and syntactic generation. The morphological generator utilizes Arabic grammar rules to construct the correct forms of the inflected Arabic words. However, the task of the syntactic generation is to generate the Arabic sentence in its final structure. For example, "I love them" consists of three words, but in translation it must become "أحبهم" according to Arabic grammar rules regardless of the structure of the source language. One of the main aspects of our system is the use of forms of the corresponding phrases between the two languages. The most of the forms of English language phrases and corresponding Arabic language forms were collected. In order to avoid, the use of complex programs for any change or manipulation in the Arabic sentence, since all the required processing would have already been done in the corresponding format.

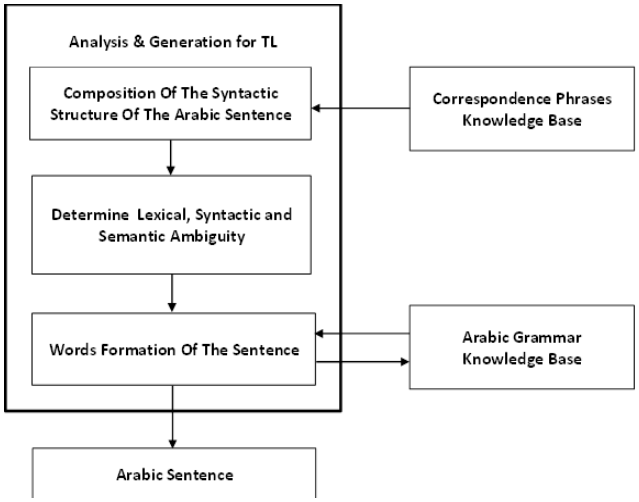

Figure.4. Steps of the transfer and generation stages

The second and the third stages consist of the following steps:

1. Determine the nature of the text.
2. Transfer the selected records of the sentence to the temporary database [2) depend on the nature and POS.
3. Determine the syntactic structure of the corresponding Arabic sentence.
4. The semantic analysis: selecting the appropriate meaning of each words in the processed sentence.
5. Determine the ambiguity of the sentence and its type.
6. Ensure that ambiguities are removed or even transferred to the output of the translation process.
7. Form the words of the Arabic sentence: Arabic grammar are applied in order to form the words according to their position in the Arabic sentence. For example: "The boys study hard.", by analyzing the word "study" and applying the rules of morphology, the verb form in the target language becomes "يدرسون" by adding "ون".

System Evaluation:

The purpose of the experiments is to explore whether machine translation systems specifically (Google and the proposed system) are accurate enough for the translation from English to Arabic. Table 2 represents set of the tested sentences. Five English sentences were selected to be translated using the proposed machine translation system as a test sample. The used evaluation method depends on the comparison between the system output and the translation resulting from various websites such as Google, Microsoft, Babel Fish, World Lingo, Free Translation, Rever so to evaluate the performance of the proposed machine translation system.

Table.2. Samples of Test Sentences

No.	Sentence
1	The technological advances witnessed in the computer industry are the result of a long chain of immense and successful efforts made by two major forces.
2	These are the academia, represented by university research centers, and the industry, represented by computer companies.
3	It is; however, fair to say that the current technological advances in the computer industry owe their inception to university research centers.
4	In order to appreciate the current technological advances in the computer industry, one has to trace back through the history of computers and their development.
5	The objective of such historical review is to understand the factors affecting computing as we know it today and hopefully to forecast the future of computation.

Table.3. Google & our system outcomes evaluation

Sentence #	(Google) Outcomes Evaluation	Our system Outcomes Evaluation
First sentence	7	9.5
Second sentence	7.5	9
Third sentence	8	9
Fourth sentence	6	8
Fifth sentence	7.6	9.5
Average rating	**7.22**	**9**

Quality is considered to be the correspondence between a machine translation's output and that of a human translation: "the closer a machine translation is to human translation, the better it is" [17].

Results: Table No. 3. Shows the evaluation of the outcomes of five sentences using Google and the proposed machine translation system: According to the results of the evaluation, the score average evaluation of the results for Google is 7.22, while the score average evaluation of the results for the proposed system is 9. It can be said that the proposed machine translation system is able to generate a better translation of Google translation from English to Arabic.

III. CONCLUSION:

This paper has been carried out with the aim of translating English text to Arabic by minimizing the need for user intervention to assist in the translation process. The main objective of this work is to design and implement a machine translation system which is able to translate text from English into Arabic using Rule-based approach .To achieve this goal, we designed a set of rules based on English and Arabic language

grammar. The designed rules dealt with different sentence structures between English and Arabic, Lexical, Syntactic and Semantic Ambiguity, the choice of the appropriate meaning of the word, Idioms and Proverbs translating, word agreement and ordering problem. The comparison method was used to evaluate the validity of the proposed machine translation system results. The comparison demonstrated that the output of the proposed machine translation system gives results with good accuracy better than Google when translating from English into Arabic. This is because the dictionary used contains suitable words for Arabic translation as well as the rules used to deal with the different sentence structures between English and Arabic.

IV. REFERENCES:

[1]. Owen Thomas, Eugene R. Kintgen: Transformational Grammar and the Teacher of English, Indiana University, Second Edition, 1974

[2]. Roderic A. Jacobs & Peter S. Rosenbaum: An Introduction to transformational Grarnmar, Ginn Company and A Xerox Company, 1987

[3]. Steven Bird, Ewan Klein, and Edward Loper: Natural Language Processing with Python, O'Reilly Media Publishers, 2009

[4]. Eduard Hovy and State of the Art/Machine Translation, (1993), "How MT Works", Byte.

[5]. Mohammed M. A. Ibrahim: A Fast and Expert Machine Translation System involving Arabic language, Ph.D. Thesis, Cranfield Institute of Technology, 1991

[6]. Eduard Hovy: State of the Art/Machine Translation, How MT Works, Byte, 1993

[7]. Daniel, J. and H. James, 2009. "Speech and Language Processing", Pearson Education, New Jersey.

[8]. H. A. Hebresha and M. J. Abd-Aziz: Classical Arabic English Machine Translation Using Rule-based Approach, Journal of Applied Sciences 13 (1): 79-86, 2013

[9]. Salem, Y.: A generic framework for Arabic to English machine translation of simplex sentence using the role and reference grammar linguistic model, MSc Thesis, Computing in the School of Information and Engineering, The Institute of Technology Blanchard town, Dublin, Ireland 2009

[10]. Hatem, A. and N. Omar: Syntactic reordering for Arabic-English phrase-based machine translation, Commun. Comput. Inform. Sci., 118: 198-206, 2010

[11]. Owen, J.: Modem linguistic theory and the Arabic grammatical tradition. Lingua, 1984

[12]. W. F. Clocksin · C.S. Mellish: Programming in Prolog, Fifth Edition, Springer-Verlag Berlin Heidelberg GmbH, 2012

[13]. Attia, MA: Repod on the introduction of Arabic to ParGram, Proceedngs of ParGram Fall Meeting. Dublin, Ireland. 2004

[14]. Alla, R., S. Richard and B Elabbas: The challenge of Arabic for NLP/MT, Challenges Processing Colloquial Arabic, University of Illinois at Urbana-Champaign, USA. No. 2005

[15]. Habash, N.: Arabic Morphological Representations for Machine Translation, Center for Computational Learning Systems, Columbia University, New York, USA, No. 2006.

[16]. Abu Shquier, M.M. and T.M.T. Sembok: Word agreement and ordering in English-Arabic machine translation. Proceedings of the Interactional Symposium on Information Technology Volume 1, IEEE Xplore Press, Kuala Lumpur, Malaysia, pp: 1-1 0., No. 26, August 2008,

[17]. Papineni, K., S. Roukos, T. Ward and W.J. Zhu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp: 31 1-318, No. 7 July 2002