An Automated System for English-Arabic Translation of Scientific Text (SEATS)

Hoda M.O. Mokhtar*, Nevin M. Darwish and Ahmed A. Rafea***
*Computer Science Dept., Faculty of Computers and Information,
**Computer Engineering Dept., Faculty of Engineering,
Cairo University.

Abstract

This work presents SEATS an Automatic System for English-Arabic Translation of Scientific Text. Automatic English-Arabic translation is still an active area of research since results are not completely satisfactory. SEATS adopts a transfer approach that employs unification based grammar, structural transformation rules and Arabic morphological synthesis rules. It consists of three main modules besides its input-output interface, namely; a preprocessing, an understanding, and an Arabic generation module. The preprocessing module is a simple scanner that adjusts some lexicals. The understanding module employs a subset of the English language grammar rules to analyze given text. The Arabic generation module includes a specialized semantic bilingual lexicon along with a set of English-Arabic transformation rules.

SEATS has been successfully implemented on a 64MB RAM machine using SWI-Prolog. Tests have been performed using a set of abstracts from the field of Artificial Intelligence. Results also include a comparison with the output of two general-purpose English-Arabic translation systems as well as human translation.

Keywords: Machine Translation, Natural language processing, English- Arabic Translation, Computational linguistics

1. Introduction

Machine Translation (MT) is an application of Natural Language Processing (NLP) that aims to produce fully automated systems capable of translating texts among various natural languages. It is the semantic mapping from one natural language to another. Machine Translation is therefore defined as " The transfer of meaning from one natural (human) language to another with the aid of a computer" (Goshawke et al 1987). Although interest in MT began by the end of the 1940's and the 1990's witnessed a lot of on-line translation systems together with lots of research and commercial systems, most systems available deal with languages with characteristics near to those of the English language. Few systems and research work dealt with languages like the Arabic language with its syntactic characteristics that are different from those of other Latin languages. For Arabic language, ATA software produced Al-Mutarjim Al-Arabey and Al-Wafi translation systems of English to and from Arabic (www-1). Also, the Sakhr CAT Translator is a computer-aided translation system supporting bi-directional bilingual translation between English and Arabic (www-2). Research in the use of the Arabic language is also limited and results obtained in this field are still not totally satisfactory.

This paper presents SEATS, a fully automatic machine translation system that is designed to produce fast, high quality translations for scientific English text to Arabic text. SEATS architecture, design, and implementation is herein, described, and results obtained are compared to the results of the two ATA software general-purpose translation systems and to manual translation.

2. SEATS system

SEATS is a Scientific English-Arabic Automatic Translation System implemented in SWI-Prolog. A Translation system consists, in general, of three basic blocks: an understanding module, to analyze input sentence; a transfer module, to translate source language structure and words to target language structures and words; and a generation module, to produce target language text. Based on this architecture SEATS is built of three main blocks that interact together to automatically translate an input abstract from a paper in artificial intelligence given in English to its Arabic equivalent text. SEATS is thus, so described in Figure (1) to consist of the following modules: *A PREPROCESSING, AN UNDERSTANDING, and A TRANSLATION MODULE* besides, the input-output interface.

2.1. Preprocessing Module and Input Mode

The overall module operation of preprocessing and input as shown in Figure (2) is as follows: first, an input text file containing an abstract of a scientific paper is to be chosen by the user. Currently, files containing abstracts of artificial intelligence papers are prepared for testing. Then, a text preprocessing procedure processes the source text before being presented to the translation system. This processing phase reformats the input text into a form understandable by the segmentation stage. A sort of morphological analysis is also applied. The analysis is mainly concerned with punctuation marks. For example, it eliminates the commas(,) in the presented text to produce flat sentences containing only a mark representing the sentence end, and any other punctuation marks that affect the sentence meaning like hyphens(-) or underscores(_) used to build combined words. The module also reserves spaces between words within the text so that these spaces act as word delimiters during the next segmentation stage, and reserves other punctuation marks as quotes (", '), brackets ([,(,),]), and braces ({,}). Finally, a segmentation stage takes place whose role is to read the input text and present it to the understanding module suitably.

2.2. Understanding Module

The Understanding module illustrated in Figure (3) performs morphological, syntax (parsing), and semantic analysis, and a part of the discourse integration phase. It employs a subset of English grammar rules that has been selected to cover almost all sentence cases that could compose an abstract encountered by the system. Being devoted to technical abstracts, the sentences are expected to be formal and restricted in structure. Basic grammar rules (of about 15 rules) were obtained from previous work (Covington 1994), whereas, additional rules (of about 25 rules) were augmented during system training. Rules were implemented using Phrase Structure Grammar (PSG). The syntax analysis is carried out through a combined syntax and semantics *Top-Down Parser*. The parser accepts a list of words building a sentence. It analyzes the sentence, and produces a list of parts of speech (Noun, Verb, Adjective, Preposition, Adverb, etc.) forming the considered sentence. During this operation, the parser makes use of a technical bilingual lexicon specially designed for the considered application. The parser performs combined *syntax and semantic analysis*. It adopts semantic features to finally accept sentences that are grammatically and semantically correct. The *UNIFICATION BASED GRAMMAR* (UBG) technique is adopted for parsing (Covington 1994). Thus, whenever a sentence-structure rule is applied, attributes -referred to as feature structures- in the rule have to unify with the corresponding feature structures. Some of these features are properties of the word and hence are explicitly listed in the "Lexical Entry" of the word in the lexicon (dictionary), the rest obtain their values through unification. This module also comprises a *morphological analysis* phase whose function is to handle different words' inflections producing their base form. By way of example, if the sentence being analyzed contains a verb like "addresses", the morphological analyzer will match the verb *"addresses"* with its base entry *"address"* listed in the lexicon. It will also return 'singular' as the verb's number, and 'present' as the verb's tense. Hence, the number of entries that need to be listed in the lexicon is reduced. Thus along with parsing the final output of the understanding module is a list containing the different parts of speech building the sentence. A final task also carried by the understanding module is a part of the *"Discourse Integration"* stage. The module keeps track of the nouns found in a sentence together with some of their semantic features. These entries maybe required later by the translation module to automatically solve reference problems, where a word in the currently considered sentence refers to another word in previous sentences. The word that references a previous word is checked for type and number and hence matching backwards the appropriate nearest noun is selected.

2.3.Translation Module

The Translation Module illustrated in Figure (4) is the final module in SEATS architecture. It accepts as its input both the list of sentence elements (*word_list*) and the corresponding list of parts-of-speech (*structure_list*). The output of the module is a sentence in target language. This module is designed to perform two main functions: transfer English words to their equivalent Arabic words, and generate Arabic text for input English text by applying appropriate rules (Nemah1977). It is noted that a number of English constructs (words' sequence) that are frequently encountered in an English sentence and whose Arabic equivalent needs special consideration is implemented with separate rules. These rules consider the translation of cases as those shown in the following examples.

- For the input string *"the system"*, the word_list will be *[the, system]* and the sentence structure_list will be [Det, Noun]. The use of the English-Arabic bilingual lexicon will translate each of these two elements as: *the* --- > *الـ* , *system* --- >*نظام* Thus, an absolute word to word translation will result in two Arabic words *"الـ نظام"*. This translation is inappropriate in Arabic language. The corresponding acceptable translation is the single defined Arabic word *"النظام"*. Thus, if the structure_list contains [Det, Noun], the 'Det' (Determinant) is seen if it is a definite article (e.g.: the), or an indefinite article (e.g.: a/an); whether it is singular (e.g.: a/an), or plural (e.g.: these/those), or number independent (e.g.: the).

Also, the structure [Det, Adj, Noun] as in " *the automatic system* ", requires a combination of two special cases (combining and inverting words) together with an extra processing step since Arabic rules state that the Noun (الموصوف) and the Adjective (الصفة) should both be either defined or undefined. Thus, the above translation case should be treated as if the structure is [Det, Adj, Det, Noun], and therefore, instead

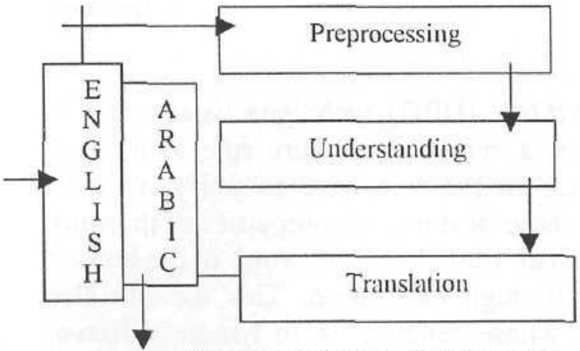

Figure (1): System Architecture

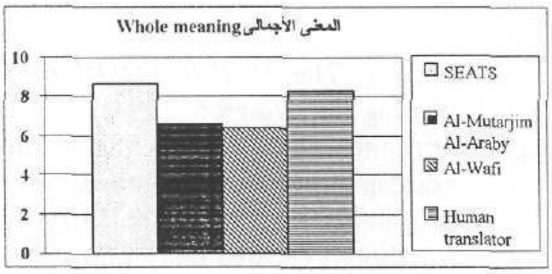

Figure (5): Whole abstract meaning

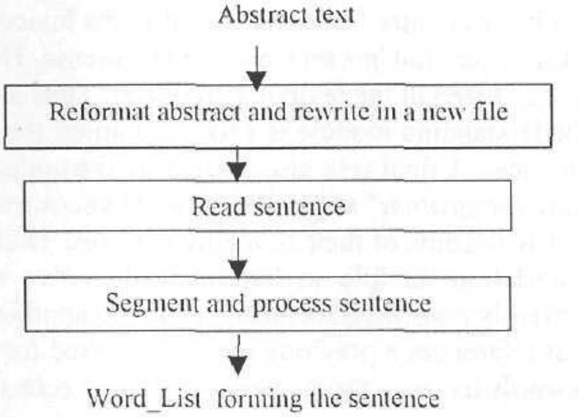
Abstract text
Reformat abstract and rewrite in a new file
Read sentence
Segment and process sentence
Word_List forming the sentence

Figure (2): Preprocessing and Input-Output Module Block Diagram

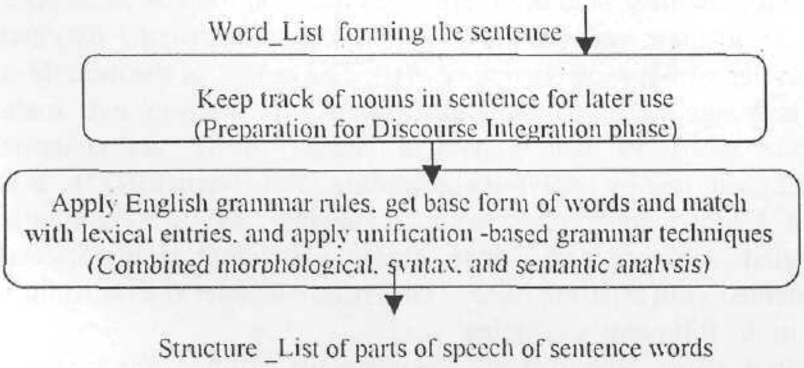
Word_List forming the sentence
Keep track of nouns in sentence for later use
(Preparation for Discourse Integration phase)
Apply English grammar rules. get base form of words and match with lexical entries. and apply unification -based grammar techniques
(Combined morphological. syntax. and semantic analysis)
Structure _List of parts of speech of sentence words

Figure (3): Understanding Module Block Diagram

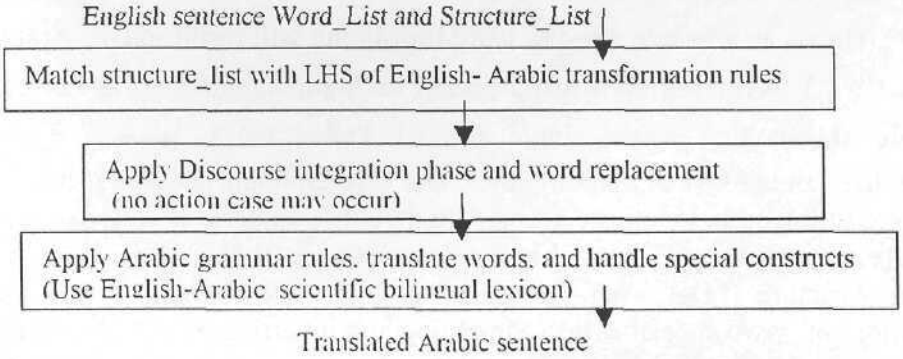
English sentence Word_List and Structure_List
Match structure_list with LHS of English- Arabic transformation rules
Apply Discourse integration phase and word replacement
(no action case may occur)
Apply Arabic grammar rules. translate words. and handle special constructs
(Use English-Arabic scientific bilingual lexicon)
Translated Arabic sentence

Figure (4): Translation Module Block Diagram

of obtaining the incorrect translation result "نظام الآلـى" we obtain "لنظام الآلـي".

Thus several special rules are also implemented for prepositions, verbs, and other encountered constructs. Finally a discourse integration stage is implemented to solve any reference problem that occurs.

3. System Results and Evaluation

SEATS has been successfully implemented on a 64MB RAM machine using SWI-Prolog. It has been evaluated through performing a number of successful tests using a set of abstracts from the field of Artificial Intelligence. Arabic translation results attained using SEATS have been compared to both translation software present in the market (Al-Mutarjim Al-Arabey and Al-Wafi), and to manual translation through translation by a recognized translation office. Results of translation by the four systems for a number of abstracts were circulated together with an evaluation check list containing 10 points of comparison from syntactic and semantic viewpoints among computer science specialists. Statistical analysis of this evaluation was carried out and a sample is given in Figure (5). As expected, the evaluation showed that SEATS translation results are superior in both syntax and semantics than the general purpose translation systems , but more important is that SEATS results came almost in the same order of evaluation of manual translation. For more details reader is directed to Mokhtar (2000).

4.Conclusions

SEATS is a novel approach in English-Arabic translation that narrow the gap in the field of scientific translations. It only required a training set of 10 abstracts that successfully worked for 30 abstracts in the field of Artificial Intelligence. The transfer approach employed in SEATS architecture that combines Unification Based Grammar, structure transformational rules, and Arabic morphological synthesis rules results in successful Arabic translations from scientific English text. SEATS is characterized by its translations that have high syntactic and semantic quality. In addition, its simplicity and modularity, enables future modification and extendibility with ease.

References

(Covington 1994) Covington M. A., "Natural Language Processing for Prolog Programmers", Prentice-Hall, 1994.

(Goshwake et al 1987) Goshawke, W., et-al, " Computer Translation of Natural Language", Sigma Press, 1987.

(Mokhtar 2000) Mokhtar, H., et-al, " An Automatic System for English-Arabic Translation of Scientific Text", M.Sc. Thesis,Computer Engineering Dept., Faculty of Engineering, Cairo University, July 2000.

(Nemah 1977) Nemah F., " "ملخص قواعد اللغة العربية, دار النشر للجامعات المصرية, 1977.

(www-1)ATA software products URL http://www.atasoft.com/products/index.htm

(www-2) Sakhr CAT URL http://www.translation.net/sakhrcat.html