ENGLISH TO ARABIC MACHINE TRANSLATION: A CRITICAL REVIEW AND SUGGESTIONS FOR DEVELOPMENT

Mostafa Aref, Muhammed Al-Mulhem & Husni Al-Muhtaseb
Information and Computer Science Department
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

ABSTRACT. Machine Translation (MT) is the use of computers in translating text from one natural language to another. This paper begins by an overview of the current research in machine translation (MT). Then, it presents a multi-level transformation approach to machine translation and the evaluation of the current research in MT according to the multi-level approach. Then, the paper refers to the research in English to Arabic Machine Translation (EAMT). Finally, the paper gives an outline for the direction of building a prototype for EAMT using AI techniques such as knowledge representation.

1. INTRODUCTION AND OVERVIEW

MT Research has started in mid 50's to translate from Russian to English. In the early stages of MT, word to word translation methods were used. In 1966, ALPAC [1] report from USA National Research Council concluded the impossibility of using computers in translation. Therefore, MT research stopped for more than a decade. In the 70's the MT research resumed as a solution to the multi-languages of Europe which was one of the difficulties facing European Community (EC). Also, in the 70's, Japan has translated many of the American and European periodicals to the Japanese language. In 1975 about 170,000 books have been translated to the Japanese language. In comparison, only 4028 books have been translated to Arabic language in twenty years (1948-1968) [2]. Japan gains a big jump by this translation in the advanced technology.

With the rapid increase of the published information, it is difficult to rely on human translation to translate such information from one language to another. For this reason, the idea of using computers in translation arises. In the 80's, Japan announced its fifth generation computers and MT was one of its components. The following subsections present brief summaries of the current research in machine translation.

1.1 Systran

Systran [3] provides translation for the USAF's Foreign Technology Division from 1970 to present. It has four stages of translation: preprocessing, analysis, transfer and synthesis. The preprocessing stage uses a look-up dictionary and applies morphological analysis to the whole text. The analysis stage consists of resolution of homographs, segmentation of sentences into main and subordinate clauses and determination of syntactic relations such as relations between nouns and modifiers. It also involves the identification of words related by enumerations, identification of subjects and objects, and identification of deep relations such as grammatical subjects in passive sentences. The transfer stage involves the lexical transfer of conditional idioms, translation of prepositions, and structural transfer using lexical routines. The synthesis stage has three parts. The first part is the assignment of the default

translation for any work that has not been already translated during the transfer stage. The second part is the morphological generation on the basis of structural information from the stem dictionary. The last part is the generation of target text. The USAF Russian-English system produces nearly 100,000 pages of text yearly since 1970. The majority of translated text is for information gathering and the reported estimated error is 5%.

1.2 Susy

Susy [4] is a multilingual German system (German, Russian, English and French) started in 1972 with a Russian-German prototype and ended in 1986 into Eurotra. Susy is a transfer system where analysis and generation are language-specific and transfer (lexical & structure dependency tree) is language-pair specific. Susy has a German analysis dictionary with 140,000 words, a German synthesis dictionary with 14,000 words, German semantic dictionary with 75,000 words, and bilingual dictionary (English - German) with 10,000 entries. Susy utilizes rescue (fail-soft) mechanism; if the input does not meet the minimum requirement then the mechanism relaxes some constrains. Translation phases include morphological analysis, phrasal analysis, semantic disambiguation, transfer, semantic syntheses, syntactical syntheses and morphological syntheses. Susy has a primitive architecture with low level programming language (FORTRAN). Many problems are resolved on the basis of ad-hoc rules.

1.3 Meteo

Meteo system [5], developed by the TAUM group in Montreal, has been in daily operation from 1977 to the present time. It translates weather bulletins from French to English (In 1989, an English to French system was developed). Meteo is a production system with single data structure. The translation method in Meteo depends on the transfer-based approach. In 1984 Meteo was replaced by Meteo-2 on a PC environment. Meteo depends on the following restrictions: weather reports have standard format, the vocabulary is fixed, restricted and predictable, only present or past participles verbs are allowed, no passives, no pronominal reference, no relative clauses and phrases are short. Meteo starts translation after receiving the information in a special format (separated words and punctuations). Meteo has three bilingual dictionaries: general and meteorological dictionary, place names dictionary, and idioms dictionary. It analyzes the input syntactically and generates an equivalent syntactic representation of the target language. It uses only five types of tree structures. Then from the generated syntax structure it produces morphological structures. No morphological analysis is done for the input because the main dictionary contains all morphological forms. It has been reported that Meteo is translating 37000 words every day at an accuracy of over 90%.

1.4 Ariane (GETA)

GETA(Ariane) System [6] developed between 1960-70 for three pairs of languages (Russian-French, German-French, and Japanese-French) at Grenoble university. The most important practical application of GETA was the Callope project in 1983 which was a French-English system for aeronautics and an English-French system for Computer Science and data processing. The translation process is a standard linguistic stratification: morphological analysis for source text, a multilevel analysis gives an intermediate source structure, lexical transfer of source-language lexical units (LUs) to the corresponding target-language LUs, a structural transfer produces an intermediate target structure, and finally a syntactic morphological generation. The Russian-French version uses a dictionary of 7000 Russian lexical units on abstracts in space science and metallurgy.

1.5 Eurotra

Eurotra [7] is one of the biggest MT projects with the planning stage approved by the European Community Council of Ministers in 1981. The aim of the project is to have advanced design capable of dealing with all official languages of EC (9 languages) and to overcome the limitation of Systran as a multilingual system. Eurotra has three phases: Linguistic & software specification; linguistic research & prototype; and large system with less restricted text. The project emphasizes on the output quality and the extendibility rather than on speed of performance. Eurotra was a transfer based system with 72 language pairs. Translation is viewed as a mapping using rules though several steps: source language, analysis, transfer and generation of target language. Eurotra is implemented using PROLOG on UNIX system with the average of 4000 lexical entities in monolingual and transfer dictionaries. The project covers a breadth of linguistic phenomena with significant advances in tense, modality and cross-linguistic differences. The simplistic approach of semantic and lack of knowledge bases lead to the failure to produce practical results.

1.6 Metal

METAL (MEchanical Translation and Analysis of Language) system [8] was released Commercially in 1989 by German Electronic Company Siemens of Munich. It translates Data processing and telecommunications information from German to English and vise versa. METAL Runs on Symbolics 36-series lisp machines, with batch processing and post-editing on workstations. The translation method in METAL was a transfer-based method. The system was implemented using Lisp. The basic system operations can be described as follows. The system accepts input from various resources, and separates textual data from diagrams, tables, charts, etc.. Then, using the dictionary database, the system produces unknown words list, unknown compounds, with suggested possible translations, and known technical words list to be checked by the user. The translation stage starts by morphological and lexical analysis to extract potential roots and affixes for every word in a sentence and to produce constructs of roots and affixes, then using phrase structure grammar the system produces alternative parsing ordered according to scores assigned to the rules of the grammar (syntactic analysis). In a next stage, transformation process operates from root of the trees downwards to dependent nodes. It uses lexical rules from the bilingual dictionary and structural transfer rules from the transfer parts generated by the analysis phase. Semantic features are used in this stage. The outputs of this stage are surface representations with full specification of word order and morphological constituents. From these representations the system produces morphological correct target language strings (generation). The system allows revision of translation with or without the original text and automatic re-insertion of textual data into diagrams, tables, charts, etc.. The system raw output is one word per second where 20% of the output does not need editing.

1.7 DLT

The preliminary study of Distributed Language Translation (DLT) [9] started in Netherlands, in 1979, to have the prototype version (English to French) in 1987, and the commercial version in 1993. The long-term aim was the translation between European Languages and for use on personal computers in data communication networks. PROLOG was used in the prototype for simplified English. DLT is based on Interlingual (Esperanto) approach. Esperanto is more like a "natural language" with its own independent structures and lexical items. DLT consists of two translation systems: source language to Esperanto and Esperanto to target language. In the prototype, all semantic and pragmatic processing takes place in the kernel interlingual component. In the commercial version, Bilingual Knowledge Bank is introduced. The basic processing stages of the English-French prototype system include source language parsing using ATN, bilingual tree transformations, semantic-pragmatic word

choice, disambiguation dialogue (source language), monolingual (interlingual) tree transformations, coding, network transmission, decoding, Esperanto parser, bilingual tree transformations, semantic-pragmatic word choice and tree linearisation (target language). In 1989, the Evaluation of DLT points out the lack of source language frequency information, the deficiencies and inconsistencies of the databases and poorly handling of structural ambiguity inter-sentence relations.

1.8 Unitran

UNITRAN [10] is an implemented machine translation system that translates Spanish, English and German bidirectionally. The translation is done by interaction between two levels. The syntactic level consists of the information necessary to accept or produce grammatically correct sentences. The lexical-semantic consists of the information necessary to provide an underlying lexical conceptual structure (LCS) and to match this structure to the appropriate target-language lexical items.

1.9 KBMT

KBMT (Knowledge-Based Machine Translation) [11] introduces AI techniques in machine translation at Carnegie-Mellon University Center for machine translation. KBMT is based on interlingual approach. The working prototype for English and Japanese, implemented in LISP, has 1500 concepts. The basic modules include: syntactic parser based on Lexical Function Grammar (LFG), semantic mapper, semantic generator and syntactic generator. The central core of the system is the representation of interlingual texts in concepts (e.g. events, individuals). A static knowledge base of these concepts are represented in a network of frames with slot values. The interlingual representation is instantiations of these concepts.

1.10 LOLITA

LOLITA (Large scale, Object-based, Linguistic Interactor, Translator and Analyzer) [12] natural language processing system developed at Durham University as a general domain-independent, natural language tool. LOLITA project started in 1987 and is still under development. The semantic network of LOLITA has over than 30000 nodes that can be easily expanded and modified using a natural language interface. The input text to LOLITA is parsed syntactically by first morphological analysis, then, a list of parse-trees for the input text is produced using deterministic grammar and parser model. The produced deep grammatical representation of the input is mapped onto nodes in the semantic network. The semantic analysis phase determines if a node already exists and how to build a new node and how to connect the new portion to the current network. In this stage, ambiguity of the grammatical parse tree is resolved using the semantic network knowledge. A natural language generation module is responsible to generate natural language text. There will be a separate translation module (generator) for each target language.

1.11 ALMUTARGEM

Al-Mutargem[13] is an English to Arabic machine prototype translator for political middle east news developed at the American University in Cairo. The prototype system uses a definite clause grammar to describe the structure of the common sentences in the used domain. Morphological analysis is done using morphological rules and a dictionary as part of the syntax analysis. After a sentence is analyzed syntactically, the prototype system replaces each English word to its Arabic meaning using semantic network and some semantic rules along with the dictionary. Then, the resulting structure is transformed into Arabic sentence internal representation which passes through a morphological generator that generates Arabic

sentences using some morphological synthesis rules. The prototype was designed to discover the problems of machine translation from English to Arabic. A limited efficiency test that included 150 sentences produced 118 correct translated sentences. Al-Mutargem was implemented using Prolog.

2. MULTI-LEVEL TRANSFORMATION

MT may be viewed as multi-level transformations. These transformations can be described as follows (shown in table 1).

2.1 Word to word transformation

It is a many-to-many transformation because a word in the source language has many meanings and a meaning can be expressed by many words in the target language. It is considered as level-0. The transformation is done without any pre-processing. Yet, word sorting, phrase sorting, sentence generation and sentence correction are required as post-processing.

2.2 Lexeme to lexeme transformation

In this level (level-1), the transformation is done after getting the lexical information of the word. Morphological analysis is required before the transformation. But also, phrase sorting, sentence generation and sentence correction are required as post-processing.

2.3 Phrase to phrase transformation

In this level, the transformation is done after getting the lexical and syntactical information of the phrase. Therefore, morphological and syntactical analysis are required before the transformation. Sentence generation and sentence correction are required as post-processing. This level is considered level-2.

Table 1. Machine Translation as Multi-level Transformations

SOURCE LANGUAGE		Pre-Transformation	Transformation	Post-Transformation	TARGET LANGUAGE
	Level-0	None	Word-to-word	Word sorting Phrase sorting Sentence Generation Sentence Correction	
	Level-1	Morphological	Word-to-word based on word types	Phrase sorting Sentence Generation Sentence Correction	
	Level-2	Morphological & Syntactical	Phrase to phrase	Sentence Generation Sentence Correction	
	Level-3	Morphological, Syntactical & Semantics	Sentence to sentence	Sentence Correction	
	Level-4	Semantic Representation	None	Sentence Generation	

2.4 Sentence to sentence transformation

Also, MT may be viewed as one-to-one mapping on the semantic level (level-3). It is a one-to-one transformation because a meaning (represented in a sentence) in the source language has only one meaning (represented in a sentence) in the target language. It requires morphological, syntactical and semantics analysis as pre-processing. And sentence generation is required as post-processing.

2.5 Zero transformation

The last level of this view is level-4, where no need for transformation from source language to target language. The input text would be represented in a semantic representation. The output text would be extracted from the semantic representation in the target language.

Considering the previous view of machine translation, the current research may be classified as follows.

1) Level 0: Most of the work done before ALPAC report was in this level.
2) Level 1: Morphological analysis is introduced before doing the transformation. Systran is considered in this level.
3) Level 2: Systems using morphology, surface syntax, and deep syntax analysis. Examples of such systems are Susy, Meteo, Ariane, Eurotra, Metal and Almutargem..
4) Level 3: Systems introduce semantic analysis as a pre-processing. Rosetta, DLT, and Unitran are considered in this level.
5) Level 4: Systems using AI knowledge representation techniques such as frame-based and domain knowledge formalisms are in this level. Examples of such systems are KBMT and LOLITA.

3. ENGLISH TO ARABIC MACHINE TRANSLATION

While the number of MT research and systems is increasing in USA, Europe and Japan specially in the last decade, MT research in Arabic Language is very limited [14]. This research includes: Linguistic problems in MT [15], difficulties in translating from English to Arabic and vice versa [16], Arabic to English Translation of figurative expressions [17], AL MUTARGEM for translating Middle east news [13], a system for automatic translation of Arabic language to English and vice versa[18], Torjomane: an Arabic to English Computer assisted translation system[19] and Arab Bureau of Education for the Gulf States research. The objectives of English to Arabic MT (EAMT) has been discussed in [20]. The directions of building a prototype for EAMT can be outlined as follows.

1- Semantic classification of Arabic words: since the semantic classification of English words has been already done long time ago[21].
2- Knowledge base of Arabic concepts: this knowledge base compiles all Arabic concepts and their relations to each others.
3- Language independent semantic representation: this representation captures the human concepts regardless of the language. Object oriented representation (OOR) is a good candidate for this representation. OOR for a concept will have different slot values for different languages.
4- Interpreter: it accepts the input text (in English) and carries morphological, syntactical and semantic analysis to get the corresponding language independent semantic representation.
5- Generator: it generates the target text (in Arabic) from the semantic representation.

4. CONCLUSION

This paper gave an overview of the current research in machine translation (MT). It presented machine translation as multi-level transformations. The Evaluation of the current research in MT according to this view was discussed. The directions of building a prototype for EAMT using AI techniques such as knowledge representation were given.

ACKNOWLEDGMENT

The authors wish to acknowledge King Fahd University of Petroleum and Minerals for utilizing the various facilities in preparation and presentation of this paper.

REFERENCES

[1] ALPAC, "Language and Machine: Computers in Translation and Linguistics," National Academy of Science, *National Research Council Publication 1416*, Washington, DC, 1966.

[2] مجموعة خبراء الهندسة الاجتماعية، " الترجمة قضايا ومشكلات وحلول 1-5. التخطيط الاجتماعي والتعليمي للترجمة"، مكتب التربية العربي لدول الخليج، الرياض، 1985.

[3] Huchins, W. , and Somers, H., *An Introduction to Machine Translation*, Academic Press, 1992.

[4] Maas, H. D., "The MT System Susy," in King, M. (Ed), *Machine translation today: state of the art*, (Edinburgh Information Technology Series 2), Edinburgh University press, pp. 209-246. 1987.

[5] Huchins, W. *Machine Translation: past, present, future*, Ellis Horwood, Chichester, 1986.

[6] Vauquois, B. and Boitet, C., "Automated translation at Grenoble University," Computational Linguistics 11, 28-36, 1985.

[7] Allegranza, V., et al. (Eds) Eurotra Special Issue, Machine Translation 6, Nos. 2/3, 1991.

[8] Thurmair, G. "Complex Lexical Transfer in Metal," Proceedings of the third Inter. Conf. On Theoretical and Methodological Issues in Machine Translation of Natural Languages (Austin, TX), pp. 91-107, 1990..

[9] Schubert, k. "The Architecture of DLT - Interlingual or Double Direct?," in Maxwell et al. (eds.), *New Directions in Machine Translation*, (Distributed Language Translation. 4), Foris, Dordrech. pp.131-144. 1988.

[10] Dorr, B., " Machine Translation: A Principle-Based Approach," in Winston, P., and Shellard, S., (eds.), *Artificial Intelligence at MIT: Expanding Frontiers*, MIT Press, Cambridge, Massachusetts, 1990

[11] Carbonell, J.& Tomita, M. "Knowledge-based machine translation, the CMU approach," in S. Nirenburg (ed.) *Machine Translation: Theoretical and Methodological Issues*, Cambridge Univ. press, pp. 68-89, 1987.

[12] Long, D., Garigliano, R., "*Reasoning By Analogy and Causality: A model and application",* Ellis Horwood Limited ,UK 1994.

[13] Rafea, A., et al., "Al-Mutargem: a Machine Translator for Middle East News ", Proceedings of the 3rd International Conference and Exhibition on Multi-lingual Computing, DEC 1992, Durham, UK

[14] Ebrahim, M., et al., "Arabic in Machine Translation," Proc. of The Seminar on Bilingual Comp.in Arabic and English, Univ. of Cambridge, UK, Sept. 1989.

[15] Bishai, W., "Linguistic Problems in Computer Aided Translation from English to Arabic, *Proceeding of the Inter. Workshop on Computer Aided Translation*, KACST, Riyadh, 1985.

[16] داود عبده، "بعض الصعوبات في الترجمة الآلية من الإنجليزية إلى العربية ومن العربية إلى الإنجليزية"، وقائع المؤتمر الثاني حول اللغويات الحسابية العربية، الكويت، نوفمبر 1989.

[17] Ebrahim, M. and Clarke, J., "Arabic-English Machine Translation of Figurative Expressions," Proc. of Second Cambridge Conf.: Bilingual Comp. in Arabic and English, UK Sept. 1990.

[18] Mashhour, A., "Automatic Translation of Arabic Language to English and Vice Versa", Proceedings of the 4th International Conference and Exhibition on Multi-lingual Computing, April 1994, London, UK

[19] Labed, L., et al., "Torjomane: An Arabic to English Computer Assisted Translation System," Proc. of Second Cambridge Conf.: Bilingual Comp. in Arabic and English, UK Sept. 1990.

[20] Al-Muhtaseb, H., Aref, M, & Almulhem, M., "Machine Translation from English to Arabic: Objectives, Plans and Steps," Proceedings of The first Sym. on Computer Applications, Bahrain, May 1994.

[21] Browning, D., *Roget's Thesaurus of English words and Phrases*, Octopus Books Limited, London, 1991.