J. King Saud Univ., Vol. 12, *Comp. & Info. Sci*., pp. 117-144 (A.H. 1420/2000)

Arabic Machine Translation: A Strategic Choice for the Arab World

Rached Zantout and Ahmed Guessoum
College of Computer and Information Sciences, King Saud University
P.O. Box 51178, Riyadh 11543, Saudi Arabia

(Received 13 March 1999; accepted for publication 14 September 1999)

Abstract. The widespread use of computers and computer networks, not least of which is the Internet, and the impact this has had on the quantity of information production is putting insurmountable pressure on nations whose first language is not English. Keeping up with technological, economic, social, cultural, and political advancements in the first world is becoming an extremely laborious and costly, and ultimately impossible, task. This impossibility makes the task culturally alienating, since a nation will have to espouse the "language carrier", namely English, and its cultural "load". This paper is an attempt to draw the attention of all the Arab scientists as well as administrators and thinkers to the important, in fact *strategic*, nature of Machine Translation. The paper surveys the state of the technology as well as its western products and contrasts them to the state of Arabic Machine Translation. This is then used constructively to put forward a proposal for the development of Machine Translation and Computational Linguistics in the Arab World.

1. Introduction

Ever since human peoples and tribes started travelling around the world and experiencing the need for communication and trade, humankind has realized the need to understand each other's language. This need has become more urgent with the increased contacts and trade between different peoples. The intellectual development that took place in various civilizations such as the Greek, Egyptian, Islamic, and Western civilizations, to name but a few, has made this need even more important. Since the industrial revolution, this communication and trade trend has expanded many times to an all-time peak with the information technology revolution. Indeed, the widespread use of computers in all walks of life, networked together into the Internet (and the worldwide web), and the phenomenon of globalization of the world economy has created more pressure in terms of needs for communication.

Given that the United States has emerged as *the* superpower of the end of this century and that Britain is one of the major world powers, English has become the *de-facto* language of international trade and communication [1]. This is a situation with far-reaching consequences: linguistic, cultural, political, economic, and social.

One of the advantages of the existence of one language that is conventionally used for trade and communication is that it represents a common platform or medium which is learned and used, by default, by all the people. This was in fact one of the reasons behind the development of Esperanto [2]. In addition, English as a common medium, makes it possible to benefit from the technological and economic lead of the Anglo-Saxon world. However, this is a shortsighted assessment of a serious problem. That one language or culture dominates the world impoverishes local languages and cultures and makes scientific, educational, social, economic, and political dependence more acute. This explains why a world power like France, backed by the world of "Francophonie" (i.e. the French-speaking countries), spends so much effort and money to try to resist the overwhelming use of English. According to [3], The French president Jacques Chirac declared in 1996: "If the new media, our language, our programs, our creations are not strongly present, the young generation of our country will be economically and culturally marginalized".

In the Arab world, there has been quite a great deal of awareness of the need for Arabization in all walks of life [4; 5]. Nevertheless, the task today has become overwhelming, human translation seriously failing to cope with the pace and scale of information production, especially in English. This is why we stress in this paper the importance, in fact the strategic nature, of Machine Translation (MT). This area has been the target of governmental, academic, and private institutions in the West and has undergone a lot of development, helped by an important effort to advance the state of computational linguistics. In the Arab world, unfortunately, research and development of Machine Translation and computational linguistics for Arabic has remained limited with almost no involvement of governmental institutions to support it. This is probably due to a lack of understanding of the full potential of MT [2], Natural Language Processing [6], and Artificial Intelligence [7] more generally.

This paper aims at scrutinizing the state of Machine Translation from and to Arabic, which we will call Arabic Machine Translation (AMT) in the sequel. It also aims at emphasizing the strategic nature of AMT for the Arab world. In Section 2, we introduce MT, its market, use, and importance. In Section 3, we shed some light on the state of non-Arabic MT, introducing the major systems that exist, presenting the efforts put on the development of the area, and discussing the area of evaluation of (non-Arabic) MT systems. In Section 4, we introduce Arabic MT and the main systems that exist on the market and discuss the gap that exists between the state of non-Arabic MT and Arabic MT and the reasons behind it. This places us in a good position to put forward,

in Section 5, some specific ideas on how to develop the state of AMT in the Arab world. Section 6 concludes the paper.

2. The Importance of Machine Translation

2.1 The machine translation market

Since its inception, Machine Translation (MT) has been a hot area of applied research. This is mainly due to the trend, since the early years of this century, towards globalization and inter-cultural relations. Translation (done by humans) is in and of itself a very lucrative business in many areas of the world. The number of documents that have been translated is minimal when compared to the number of documents that need to be translated. As early as 1970, UNESCO had recorded a 4½-fold increase in translation efforts since 1948. Scientific, medical, and technical journals are translated wholesale in the USA and the (ex-) Soviet Union. As early as 1967, hundreds of scientific journals were being translated annually. The USA, the (ex-) USSR, Germany, Canada and Japan have been the countries where most investments have been made in the "translation industry". According to [8], 70% of information about a technology is available in patents. Currently most of the patents are written in English and Japanese in the technology area (50% alone are US patents and 300,000 patent applications are filed yearly in Japan [8]). This means that any non-English speaking country which wishes to follow technological developments must have a way to translate such a huge amount of literature on a continuous basis.

In general, the MT world market is currently considered underdeveloped compared to the potential it could achieve. The estimated value of the world market in 1991 was U.S. $20 billion according to the Gartner Group (Stamford, CT, USA). As of 1994 the MT tools market was at U.S. $20 Million [9], growing 20 to 30 percent annually. This situation is expected to change in the near future. This is due to the fact that the MT market is due to develop to its fullest potential since its associated technologies have become more useful, accessible and affordable. Mainly this is because of the increase in the complexity of hardware and software tools that are available to the average developer and the decrease in price of the hardware needed by a user to run machine translation software. For example, *Metal,* one of the most advanced operational MT systems, when it first appeared during the 1960s, ran on a Symbolics machine and cost U.S. $170,000. In 1994, it was reported to run on a SUN SPARCstation and cost U.S. $40,000 [9].

More than 50 companies, as reported in [8], are involved in the MT market. In addition, two major initiatives currently exist that underline the importance of such an industry. The first is by the European Community, the U.S. $30 million Eurotra project. The other is by Japan, The Asian Multilingual Machine Translation System which was initiated in 1987 to translate between major Asian languages but is not yet commercially

available. Among the international corporations that are currently interested in the Machine Translation market are such giants as Microsoft, Siemens, Fujitsu, Hitachi, Toshiba, Oki, NEC, Mitsubishi, and Sharp. The list of users of Machine Translation software is even more impressive. It includes the United Nations, the US Government including many of its civil and military agencies. It also includes the Patent and Trademark Office, the National Science Foundation, and many important international organizations worldwide. In addition, Europe, Canada, Russia, China, Korea, Malaysia, Indonesia, and Thailand are but a few among the many countries interested in MT [8;10]. Virtually any company, government agency or non-governmental organization that needs to communicate in more than one language would be interested in using Machine Translation software.

The MT market is still unstable as different quality products are selling at prices with large variations. Generally, translation of documents today is increasingly being done with the help of machine translation software worldwide. The software used ranges in cost from U.S. $19.95 to a high-end of U.S. $100,000. The U.S. $100,000 is a full enterprise-wide license across many sites. Current MT software completes between 30 and 50 percent of a Machine translation task automatically [9;11]. Metal, a piece of MT software that translates whole sentences and outputs good draft-quality translation among English, German and Spanish costs U.S. $19,195 per one-way language pair as of 1994 [12]. *Intelligent translator* from Logos Corp. costs around U.S. $100,000 and translates between English and French. *LogoVista* from Language Engineering Corp. provides first approximations to translations from English to Japanese and costs U.S. $1995.

MT software is expected to cut the cost of translation by two thirds by transforming the process from a manual translation to semi-automatic translation, i.e. involving human post-editing of the machine output. When compared to human translators output, the increase provided by MT is impressive. It was reported in [11] that some companies have claimed to be translating up to 30,000 pages per year; Xerox uses the MT system Systran to translate up to 60,000 pages per year. As to the translation costs, there are now many companies on the internet that offer (human) translation of web pages charging at least U.S. $0.15 per word, with a waiting period of over 24 hours. MT is expected to change such numbers drastically. Currently existing MT software can bring down the cost of translating a document considerably while minimizing the translation time. Sietec claims that its MT software, Eurolang, saves 40% of the translation budget. However, expectations for the efficiency and correctness of the MT system should not be very high. In the near future, any MT system would still require human interference to be able to produce acceptable quality translations.

2.2 The uses of MT systems

MT systems can be used for different reasons, some of which are given below.

1. As a means of translating scientific and technical documents and textbooks, commercial and business transactions, administrative memoranda, legal documentation, instruction manuals, industrial patents, publicity leaflets, and newspaper reports.

2. As a commercial opportunity used for launching products abroad. MT is *par excellence* a tool that facilitates world trade. The latter is indeed getting increasingly global, especially with the trend for English to be its "official" language. In order to reverse this trend and enable non-English speaking people to take benefit of world trade, non-English speaking countries must find a way of translating documents produced in their languages to the languages of their trade partners (which do not have to be English-speaking), and vice versa. In fact, in order for trade to be completely globalized, either the whole world should speak one language, which is absurd, or an easy way of translating between different world languages, at least the major ones, should be found.

3. As a technology that would enable non-English speakers to be part of the Internet. The Internet has become a vital part of commerce and information exchange all over the world. Non-English speakers are seriously disadvantaged due to the fact that most of the information available on the Internet is in English. The new information technology tools have turned the translation effort into a bottleneck of the international trade and development efforts. Among such tools are computers with all the software packages that are developed and used worldwide. Other important tools are computer networks with the example of the Internet being the tool that is revolutionizing many aspects of world life and trade. The World Wide Web (WWW) is a prominent example of the explosive growth of the Internet. At the end of 1992, there were only 50 WWW servers. One year later they were 250. A year later more than 2000 servers were connected to the WWW. As early as the end of 1996, 100,000 WWW servers were operational with hundreds being added daily. As to an idea about the use of the WWW, one might cite the example of Intel's WWW site, which was reported to have gotten more than 300,000 hits per week. Forty percent of those hits originated from outside the United States [10].

4. As a tool that helps any community preserve its culture and present it to other peoples. Any culture that strives to survive in the global village of ours should invest a great deal of effort and resources in Machine Translation. Language is more than a vehicle of expression between people. Any language carries in its intricacies subtle characteristics of the culture that stands behind it. As such, using a given language entails being sensitive, or even permeable, to its underlying culture. Of course, we do not want to give the wrong perception that MT can counter, on its own, the effects of cultural aggression. Nevertheless, it can surely be used as a means to alleviate the feeling of absolute need for some other language, English in particular, to survive in this techno-economic world of ours. MT is also useful to

translate from a culture's language into other languages in order for this culture to be able to project its richness and subtleties to others.

5. As a high-tech industry, that will create high-skill jobs for unemployed citizens. It can also be used as a catalyst for processes that will bring broader benefits for the society in health, sciences, technology and environmental issues, in a country that would otherwise fall behind in such vital areas.

6. As a common platform for developing artificial intelligence research, knowledge processing, computational linguistics (parsers, lexical and morphological analyzers, etc.), speech recognition, unconstrained text search through multilingual databases, database abstraction, report generation, general information management, and specialized application shells.

2.3 Arab investment in MT: A top priority!

Although only 20% of the world population know English (as a native or second language), it was reported in [8] that around 50% of science and technology literature is published in English. Arabic accounted for a minimal part of the 50% that was not published in English. This means that an Arab who cannot read English will be deprived of more than half of the available literature in technology. In addition, Arabs that cannot communicate in English will be excluded from the growing area of collaborative international projects. English has become the de-facto language that most people involved in technology, international trade and politics must understand. Assuming we could use the Internet as a sample of the world population, we find that, currently, French is in three sites out of 10 (a 92% increase since last year). Other languages, excluding English, have a far smaller percentage of presence on the Internet where English has also become the de-facto language. For example, the use of French on the Internet is only 7% [13]. In a survey performed at King Saud University in Riyadh [14], over 54% of a random sample of Arab academics (in Saudi Arabia) questioned, do not use the Internet because of their weakness in English. It is obvious that this ratio would be much bigger if non-academics had been considered in the survey. This means that an appallingly large proportion of the Arab people are simply ignoring a whole wealth of information that would bring them large benefits, for the mere reason of not understanding English.

Arabs should realize the importance of machine translation in relation to their language and their culture. This is due to the reasons discussed above as well as other reasons specific to the Arab people who represent a sizeable and geo-strategically important ethnic block in the world today. Machine translation from other languages into Arabic is obviously crucial. This is in order to enable Arabs to keep up with the advancements in the world and in catching up with the rest of the world at all levels.

Machine translation of Arabic into other languages is also an important area since it enables Arabs to disseminate information more efficiently and make their culture and products accessible to more people. Many Arab countries are moving into the industrial phase and as such will need, in the near future, to export even more products to the rest of the world.

Arabs should invest in AMT for the following additional reasons. AMT will allow them to:

1. Face the acute shortage in human translators from and to Arabic, especially given that the whole world, including the Arab world, faces the ever-increasing pressure for the translation of massive amounts of information from very different disciplines. According to [8], Arabic is one of the languages that "professional technical translators in the US are largely unable to translate".

2. Keep up to date with the technological, economic, and financial developments in the world. According to [8] "a country unable to assimilate a high volume of potentially useful information from abroad could lack timely, accurate data and lose its edge in international business, diplomacy, military readiness, and academic research." Additionally, "whoever controls the information industry will decide who has access to what." It turns out that MT is the only way, available today, to screen massive amounts of material while it is still useful. Human translators are excessively slow and too scarce to keep up with the information explosion. Accelerating the efforts for translation of information databases and media contents from and to Arabic requires the availability of translation methods that are faster than the currently existing human translation. It is obvious therefore, that MT will boost the technology transfer efforts to make more information about new technologies available to Arabs in their native language.

3. Counter the trend towards a global linguistic and therefore cultural, political, economic, and social domination of the world by one language/culture. As an important by-product of this effort, Arabs will have served the Arabic language. Indeed, the development of AMT requires research on ways of computerizing the use of Arabic, bringing to light, in modern terms, the extensively rich literature on Arabic words, structures, and semantics. It will also require new research on lexicons, lexical analyzers, parsers, semantic analyzers, and pragmatic tools for Arabic.

4. Develop the discipline of AMT. Human translation from and to Arabic is by far less institutionalized than its counterparts in the West. This fact

might explain why the efforts in Arabic machine translation within or outside the Arab world remain minimal. Moreover, in order to give Arabic literature and culture more exposure outside Arab circles, it is necessary to develop the translation capabilities from Arabic into English and other important languages.

5. Enable the fast translation between Arabic and Islamic languages (non-Arabic languages used by Muslims such as Urdu and Malay), which will facilitate integration and exchange among Muslims in general. In addition, AMT will allow the translation of a large body of Islamic literature to Islamic as well as other languages. This would be the best way of portraying the true, shining picture of Islam.

6. Automate the preparation of bilingual and multi-lingual documents especially for those Arab countries, which have large numbers of non-Arab expatriates.

7. Give Arab inventors a chance to access the patents in the US, Japan and the European union. This will not only enrich their technical expertise, but it will also help them register their patents and, hence, preserve their rights as inventors.

It is noteworthy that most of the demand for translation related to the Arabic language is for software to translate from English to Arabic (90% of the demand in 1986 according to [15]). Therefore, we think that Arabs should first concentrate on developing good systems that can translate from English into Arabic. Such systems should be developed, keeping in mind the possibility of expansion into more languages in the future. Translation from Arabic to English is a second priority that should benefit from the efforts of developing a good English to Arabic translation system but is, at the present time, not as urgent as an English to Arabic one.

3. The State of MT for Non-Arabic Languages

3.1 Approaches to MT

Machine translation systems can be classified into three categories regarding their design: Direct, Transfer-based, and Interlingua-based systems. The first approach, which has now been largely given up by MT researchers, consists of incorporating all the details for some specific pair of languages in one translation direction. Translation using the Direct approach is done in almost a word-to-word manner. The second approach makes use of internal syntactic representations where knowledge is represented after disambiguation. The source text is first translated into an internal representation of

the source language; this is then converted into an internal representation of the target language, which is finally used for the generation of text in the target language. The third approach makes use of an Interlingua, that is, an entirely independent language. The Interlingua is used as a common representation for the parsed source text as well as for the text to be generated in the target language. This approach, by far the most ambitious, has the advantage that it simplifies adding to a given MT system the capability of translating between additional pairs of languages. The addition of any language capability to the system would only require the addition of two new modules: one for analysis and one for generation.

Machine translation systems can also be classified based on the level of interaction during the translation process, with the user. Systems based on pre-editing require the user to edit the original document in order to remove any ambiguity. Systems based on post-editing require the user to do the same only after the document has been translated. Interactive systems would involve the user in the translation process throughout the translation. The program asks the user questions when it needs more information to perform the translation. All of the previous systems would be classified as Human-Assisted Machine Translators. When the level of involvement of the human user becomes more than that of the machine, the system is classified as a Machine-Assisted Translator. Many of the currently existing MT software are machine-assisted-translators, a fact that appears to be one of the facts that are currently improving the sale of MT systems.

In order to make the problem of translation more computationally tractable and increase the accuracy of MT systems, researchers have resorted to limiting the vocabulary of the source language. This limitation has allowed researchers to circumvent many yet unsolved problems in Machine Translation such as disambiguation and context-dependent translation. However, this leads to another categorization of MT systems related to their capability of translating texts written in a language or only in some subset of that language.

A discussion of different types of MT systems will be incomplete if the concepts of translation for assimilation and translation for dissemination are not mentioned. Translation for assimilation is targeted towards giving the user a summary of the most important ideas in the translated text. Therefore it is usually not concerned to a high degree with the style of the text or its correctness, as long as the translated text conveys the idea(s) present in the source text in a satisfactory manner. Translation for dissemination, on the other hand, would be more concerned with producing a translated text that would have a good style and would be grammatically and syntactically correct in the target language. This is mainly because the translation done by any MT software cannot currently produce a 100% correct output. Therefore, and to make the output of a MT software most useful, it is beneficial to consider the purpose of the translation when evaluating the output. Users of systems that are used to translating for assimilation can

tolerate low quality translations as long as the general meaning is not lost in the translation. However, they require the MT system to work with open domains. Generally, translation for assimilation is used to help humans index and store documents, analyze events, and scan documents for relevance. In contrast, users of MT systems that are designed with the purpose of translating for dissemination require the system to be able to deal with high-volume documents that should be translated with the highest available quality.

3.2 Review of non-Arabic MT research and products

Research and development in the area of machine translation for non-Arabic systems (non-Arabic MT) has been an active discipline for almost half a century. In 1949, Warren Weaver suggested that computers could be used to perform text translation. Since then, (non-Arabic) MT[1] has gone through a lot of development and has benefited of all the research and products of natural language processing.

One of the early MT systems is *Systran*, which targeted in the 1950s Russian-to-English translation. In the late 1960s and early 1970s, interest concentrated on Russian-to-German translation. Today, Systran is a professional, fully automatic machine translation system that deals with scientific and technical text. It is available as translation software between 10 pairs of the European Union Languages. Other versions are currently being developed for another 13 pairs of languages, including English to Arabic. Systran has undergone intensive scrutiny; in October 1976 and June 1978, it underwent major evaluations. In terms of intelligibility (clarity and comprehensibility), the system scored 78%, whereas the correction rate was 36%. Systran has been extensively used by major agencies and companies including General Motors of Canada, the NATO headquarters in Brussels, The German National Railways, the German Nuclear Research Center in Karlsruhe, the International Atomic Energy Authority, and the French company Aérospatiale. The company Xerox uses Systran to translate some 60,000 pages per year. Most of these companies and agencies use Systran to produce raw output for information purposes, rather than for output that would later be post-edited [2]. A professional, commercial version of Systran is available at the cost of U.S. $1495. It allows translation from English to any of French, German, Italian, Portuguese, and Spanish. It also allows translation from any of French, German, Italian, Japanese, Russian, Portuguese and Spanish to English.

Susy is another multi-language translation system that involves German, Russian, English and French, though the emphasis has been on the German language. Susy initially started with the unsuccessful attempt to adapt Systran to perform Russian-to-German translation. It was then developed during the 1970s at the Universitat des

[1] In the sequel, MT will denote non-Arabic Machine Translation, whereas AMT will denote Machine Translation from and to Arabic.

Saarlandes in Saarbrucken, as a project on its own, funded by the governmental German Research Association. Developed in the programming language Fortran, Susy, despite some interesting features, has suffered from the serious limitations caused by Fortran. Nevertheless, the system was used with some success by various agencies and institutions. For instance, it is worth noting the coupling of Susy with the MT system Titran at the Japanese Kyoto University to produce translation of titles from German to Japanese, using English as an intermediary language.

Météo is a system that translates public weather forecasts from English to French. At the request of the Canadian government, Météo was developed by the Machine Translation group at the University of Montréal. First made available in 1976, it has been successfully in use since 1977. An improved version, Météo-2, was implemented on microcomputers and made available in 1984. Météo has been very successful, but its narrow domain (weather forecasts) and the fact that it has been extensively refined to be well tuned for this specific application may explain its success. However, currently, it is unclear whether *Météo* is still used for MT.

Ariane was developed at the University of Grenoble and was first made available in 1975 with a newer version in 1984. More versions have since been released. Research and development on the Ariane project mainly focused on German-French and Russian-French systems. However, other languages have also been investigated. These include Portuguese, Malay, Japanese, and Chinese. Ariane has been very influential in the machine translation community. Many Japanese MT systems are similar to Ariane in design.

Eurotra began as an ambitious research and development project funded by the European Community from 1983 to 1993 to develop a multilingual MT system based on the latest advances in computational linguistics. As of 1992, work was still being done on an industrial prototype that had good linguistic and computational performance. However, this work has not led yet to a single working system. All what it has produced are some useful by-products and an increased experience in Natural Language Processing (NLP) for a range of different languages. This helped those languages to survive. The main problem facing the Eurotra project is the difficulty of coordinating such a large international project. However, the amount of investment that was put into Eurotra does emphasize the importance of MT and the extent to which Europe feels that it needs MT.

Metal is one of the most advanced operational MT systems. Based on research at the University of Texas at Austin that began in the 1960s, its first release was in 1989. Metal translates from German to English and is being extended to English-German, Dutch-French, French-Dutch, German-Spanish, German-French, and German-Danish. When it first appeared, it ran on a Symbolics machine and cost U.S. $170,000. In 1994, it was reported to run on a SPARCstation and cost U.S. $40,000 [9]. It was also reported

to translate whole sentences at a time and output good draft-quality translation among English, German and Spanish at a cost of U.S. $19,195 per one-way language pair as of 1994 [16]. The same company (Sietec, Canada) offers the *Eurolang Optimizer* as an extension to Metal. The Eurolang Optimizer performs what Sietec terms pre-translation. It produces a color-marked document that a user can then post-edit into a translated version of the original document. Sietec claims that Eurolang saves 40% of the translation budget.

Work on *Rosetta* started in 1982 at the Phillips Research Laboratories in Eindhoven, Netherlands. Three versions have been produced so far. Work on a robust, application-oriented fourth version was started in 1989. Rosetta is an experimental system in that it attempts to devise as interlingual representation using a novel grammar formalism. It has concentrated on the Dutch-English and Dutch-Spanish pairs.

The *KANT* system, developed at Carnegie-Melon University, is an Interlingua-based machine translation system that mainly uses the Knowledge-Based approach. In addition, KANT uses statistical methods for Corpus Analysis and dictionary creation. KANT is a development from an earlier system, KBMT-89, designed by the same group [17]. It translates from a subset of the English language (controlled English, in KANT terminology) with a fixed set of about 8,000 to 14,000 words. The use of this controlled language enables KANT to produce a full translation of the source text, in effect substituting post-editing with pre-editing. The KANT system has been successfully used by the Caterpillar Company to translate its technical documentation (about 10,000 pages a year) into many non-English languages.

Currently SYSTRAN and METAL are the MT systems that dominate the Western Hemisphere market [8]. However, a number of other pieces of MT software exist. *Intelligent translator* from Logos Corp. uses semantic tables with around 15,000 rules to translate between English and French. It costs around U.S. $100,000. *LogoVista* from Language Engineering Corp. provides a first approximation to translation from English to Japanese and costs U.S. $2000. Two major initiatives exist currently. The first is by the European Community, the U.S. $30 million Eurotra project. The other is by Japan , the Asian Multilingual Machine Translation System, which was initiated in 1987 to translate between major Asian languages; however, it is not yet commercially available [1]. In 1997, several companies revealed new Machine Translation software. Intergraph Corporation introduced *Transcend,* a U.S. $500, PC-based MT software that translates between English, French, German, Spanish, Italian and Portuguese. Fujitsu also introduced the *Atlas* Machine Translation software that was originally designed to help Mazda translate its car owner manuals from Japanese to English. Currently, Atlas is accessible through the Internet on Fujitsu's Web page. *Accent* offers translation involving five European languages at a cost of U.S. $300 per language pair. *Ambassador* is a U.S. $99 translation-based correspondence system. *GTS-Power* allows translation

involving English, French, Spanish, and Japanese. The system costs U.S. $1400 for the Unix-based version or U.S. $780 for the PC-based version.

It is beyond the scope of this paper to review all the software that exists today. The reader is referred to [47] for a detailed evaluation of MT technology and products. Generally, the translation of documents is increasingly being done with the help of machine translation software worldwide. The software used ranges in cost from U.S. $99.95 to U.S. $150,000 and complete between 30 and 50 percent of a machine translation task automatically [9]. Although the discipline has had time to mature and its methods and tools have been refined, there remains a lot to be done. The primary reason for this is that Machine Translation represents a crossroads between human translation (as a field on its own), Artificial Intelligence, and Linguistics. The tools in these areas, though formalized, are still undergoing research and development. Natural Language Processing problems, especially those dealing with disambiguation and contextual reasoning, turn out to be intrinsically complex problems to comprehend and, subsequently, to model in a computationally viable way. In fact, according to [18], as of 1996, MT "under controlled conditions and using pre-tested texts... is only 80% accurate, with abysmal syntax and style."

Based on the above review of various (non-Arabic) MT systems, it is worth emphasizing the heavy involvement of governmental agencies in funding and pushing to fruition the MT enterprise. Such names as the US government, the Canadian government, the Japanese government and the European Union are main contributors to the funding of MT projects. The private sector has also played a major role in this respect, especially Japanese companies such as Fujitsu, Hitachi, Toshiba, Oki, NEC, Mitsubishi, Sharp, and others. A large number of universities and research centers have been also involved in this major effort. Some examples are Carnegie Mellon University, the MCC research center in Texas, The University of Montréal, Japan's Center for the International Co-operation for Computerization, the University of Saarbrucken and the German Center for Artificial Intelligence (DFKI), UMIST in Manchester, and the University of Grenoble.

3.3 Lexical resources for machine translation

Central to any MT system, and more generally to any linguistic system, are its lexical resources. This includes mainly the information associated with individual words in its lexicons (dictionaries). Monolingual dictionaries for the human user usually contain information such as irregular inflected forms, a definition of meaning in some form, and, often, some indication of the history of the word. The lexical entries for MT contain the information needed for syntactic and semantic processing, and this has to be far more explicit and detailed: grammatical category, morphological type, sub-categorization features, valency information, case frames, semantic features, and selection restrictions [2].

Substantial work has been devoted to the issues related to the construction of (non-Arabic) lexicons [19-21]. In [19] a detailed account of how word senses are represented is given. The essential nature of the intimate interconnection between the meaning representation, the ontology, and the lexicon is stressed; and the way this connection is accomplished in the MT system MikroKosmos [22]. One finds in [20] a good special issue of the *Machine Translation* journal, devoted to the building of lexicons for MT. Papers in this special issue addressed major research areas like the lexical levels (syntactic, lexical, semantic, ontological, etc.) required by a MT system and the interdependence between these levels. Automatic procedures for the construction of lexical representations; semi-automatic methods for the acquisition of lexical knowledge were also discussed. Use of existing resources and aids for transforming these resources into appropriate representations for MT was also discussed as well as accommodation of MT divergence and mismatches in the lexicon. Another collection of papers devoted to the Lexicon and MT is [21], a volume in the series "Lecture Notes in Artificial Intelligence". A major issue tackled in this volume is the re-use of lexicons and their combinations to build larger ones to support MT. The extraction of class-hierarchies from dictionaries versus the merging of multiple lexical resources was one of the subjects discussed. Another subject was the acquisition of technical terms from corpora. Other subjects that were discussed were the use of machine learning for lexical acquisition and standards for re-use and management of lexical data in commercial systems.

3.4 Evaluation of MT systems

A serious issue to consider when dealing with a MT system is the ability to evaluate it scientifically. This is in order to establish its usefulness and compare it with other already available systems. This evaluation should cover both its computational and linguistic features and should be informative, especially to potential purchasers and users of the system, as to its capabilities and updatability. However, an important part of the evaluation is to be performed by the developers. This is in order for developers of MT systems to check that the system performs as intended, that it produces acceptable translation, and what changes to the system are possible without radical changes to its programs and facilities.

Evaluating software is important not only for its potential users and buyers, but also to researchers and developers. This is also crucially true for MT software. It is obviously important to scrutinize any MT system, analyzing the quality of its output, classifying errors it makes, and improving it based on such an evaluation. The evaluation can thus be concerned with the technical quality of the system; it can also be concerned with the limitations of the system, the software engineering aspects, and/or the costs and benefits [2].

Various types of evaluation have been developed. Among them are:

1. Black-box evaluation: evaluating a system by just looking at its output and comparing it with what would have been expected from what has been input to the system. This is often carried out as a comparison with human translations. In other words, the system output is compared to what human experts would produce. This is a very subjective way of evaluating a MT system; however, it gives developers, as well as potential users, an indication of the general weaknesses of the system.

2. Glass-box evaluation: evaluating a system by performing evaluations on each of its constituent components. This type is usually done by the developers of a system in order to measure the improvements introduced to their product because of certain changes.

3. Evaluation based on customer criteria: determining the usefulness of a system in the actual environment in which it would work.

The Evaluation of MT systems has attracted the interest of funding agencies since the early steps of MT development. Some effort has already been devoted to the development of evaluation criteria, metrics, and methodologies. In [2], one finds a good survey of the various kinds of evaluations, namely:

i) quality assessment in terms of accuracy, intelligibility, and style;
ii) error analysis; and
iii) benchmark tests.

The evaluation should also target an assessment of the computational limitations of the MT system as well as its cost and benefits for the potential purchasers and users. A number of contributions exist such as [23; 24] on the evaluation by users and [25-27] on methodologies for MT evaluation. [28] is a collection of papers which includes a number of discussions of methods for MT system evaluation. Notable evaluations of MT systems are those of Systran [29; 30] and of Logos [31; 32]. Major projects exist such as the DARPA project [33], the project DIET [34] at DFKI (Germany), and the European project Eagles [35] for the development of diagnostic and evaluation tools for Natural Language Processing (NLP) applications.

A more recent contribution is that of the Machine Translation group at Carnegie Mellon University [36;37] where a methodology was presented, which allowed a component-based evaluation of a MT system. The authors of the study introduced completeness, correctness, and stylistic criteria, which were defined in terms of the lexicon, grammar, and the mapping rules.

1. Completeness: which measures the ability of the MT system to assign an output string to each input string. It can be broken down into three basic types:
   i) Lexical: which measures the existence of source and target lexicon entries for each word in the domain.
   ii) Grammatical: which measures the ability of the MT system to analyze all grammatical structures in the input and generate grammatical structures at the output.
   iii) Mapping rule: which measures the ability of the MT system to assign at least one output structure to every input structure.

2. Correctness: which measures the ratio of correct output strings to the number of strings taken as input.
   i) Lexical: which measures the ratio of correct words at the output.
   ii) Syntactic: which measures the ratio of correct grammatical structures at output.
   iii) Semantic: which measures the ratio of complete meanings of the output sentences to the complete meanings of the input sentences.

3. Stylistics: Measuring the understandability of output
   i) Lexical Appropriateness: which measures the ability of the MT system to choose the most appropriate word depending on the context.
   ii) Syntactic Style: which measures the ability of the MT system to choose the most appropriate grammatical style.
   iii) Usage Appropriateness: which measures the ability of the MT system to choose the most appropriate expression in the output.
   iv) Level of text difficulty: which measures the level of difficulty for the output text.

These criteria were actually translated into metrics which could measure the *Analysis coverage* and the *Generation coverage* of any MT system. Those metrics were then combined, using simple mathematical formulae, into a collective measure of the system called *Translation correctness* [36].

Despite the fact that some interesting methodologies for (non-Arabic) MT system evaluation have been put forward, we believe that they still lack well-foundedness. For instance, the valuation and combination of the various kinds of errors should be done based on some theoretical grounds, such as Statistics, Certainty Factors, Fuzzy Logic, Dempster-Shafer theory, etc. Thus, we believe that work remains to be done on the development of a principled methodology for evaluating MT systems. Moreover, all of the evaluation methodologies that we currently know of have a lot of subjective information included in the data on which the evaluation is being made. The evaluation criteria should be independent of the person(s) performing the evaluation and therefore

depend on an objective metrics that depend only on the output of the MT system. Such an objective evaluation would necessitate the development of objective criteria that will depend on the contents of the output rather than the opinion of whomever is performing the evaluation. This points to the need for developing tools for the automated, or semi-automated, evaluation of MT systems in particular, and NLP systems more generally.

3.5 The frail nature of MT

The increased volume and complexities of world trade and communication have put increasing pressure on the required translation effort. However, research in MT has not been able to match the increase in requirements for translation, neither in the quality nor in the quantity of translated texts. As mentioned earlier, according to [18], as of 1996 MT "under controlled conditions and using pre-tested texts... is only 80% accurate, with abysmal syntax and style". The adopted approach of much existing software to get good-quality translation is to limit the area of applicability of a given software. Such a limitation will limit the vocabulary and somewhat the grammar, which makes the problem of translation acceptably feasible for computer programming. Another adopted approach is to regard the translation software as an aid rather than a replacement of the human translator. Post-editing of machine-translated documents by human translators is common nowadays. In fact, it seems that the majority of today's machine translation software pitch themselves as aids to human translators and explicitly state the need for post-editing or pre-editing as well as on-line help from human translators.

The weaknesses of MT as observed today point, in our opinion, to the limits research in computational linguistics has reached. More precisely, what is currently needed is the enrichment of MT systems with the recent developments in Pragmatics as well as the increased power of Artificial Intelligence inference tools. What NLP and MT tools need most is an ability to achieve contextual reasoning. Indeed, being able to reason in context, such systems will be able to select the right word meaning, grammatical structure, referent for pronoun resolution, and more generally, the correct sentence interpretation in the given context.

4. The State of Arabic Machine Translation

4.1 Existing arabic machine translation systems

Compared to its American, European, and Asian counterparts, Arabic Machine Translation is still in its infancy stage. Nevertheless, the area has recently been gaining momentum. As first fruits of this endeavor, commercial software has been produced by companies such as *Apptek*, *ArabTrans*, *CIMOS* and *ATA*.

In 1990, Apptek (Virginia, U.S.A.) started developing an AMT system. This led to the English-to-Arabic machine translation system, *Transphere*, which can be used for translating both general and specialized domain texts. The software makes use of a lexicon of more than 100,000 entries. It can be used for translating documents and, in conjunction with word processing, as a tool for computer-assisted learning of Arabic. It has reportedly been used as a translation engine in large complex products related to air traffic, transportation, communications and control. The software runs currently under UNIX and Windows.

CIMOS (Paris, France) produces a software package *Al-Nakeel* intended for translation between different languages. It currently translates from English to both Arabic and French and from French to both Arabic and English. It is intended to assist, not replace, human translators in a wide range of areas (e.g. science, technology, commerce, banking, computer and petroleum). Each area has its own customized dictionary and translation-memory database. The software is capable of learning new rules and information using sentence analysis and semantic connections. *Al-Nakeel* costs U.S. $1000 and runs under MS Windows. CIMOS claims that *Al-Nakeel* produces translations at the speed of 20,000 words per hour.

ATA (London, U.K.) produces *Al-Mutarjim Al-Araby*, which is heralded as the first PC-based professional English-Arabic translation system. This system was first introduced in 1995 with a reported minimum speed of 1000 words per minute. It comes with a 300,000 English word and phrase dictionary, and add-on modules that target specific areas such as Science, Technology, Commerce, Finance, law, Oil industry, Agriculture, Medicine, Military, etc. The software also displays alternative meanings, when they exist, for the user to make an informed choice. It has a transliteration option for non-dictionary names and proper nouns and performs translation of abbreviations, e.g. UK, UN. The same company produces Al-Wafy which is a scaled-down version of Al-Mutarjim Al-Araby. The prices of Al-Mutarjim Al-Araby and Al-Wafy on the market are U.S. $500 and U.S. $50, respectively.

Al-Alamiyah (Riyadh, Saudi Arabia) is also working on Machine Translation software between English and Arabic. As of October 1996, it had completed around 70% of the software by building a 10,000 root English lexicon, a Morphological Analyzer and Grammar rules. Inside sources have recently revealed that Al-Alamiyah is still having problems with its Morphological Analyzer that is preventing it from introducing its own product on the market. The Aim of Al-Alamiyah is to produce MT software that would translate general texts and not only those restricted to certain areas.

As to research related to AMT, one can mention two recent pieces of work. In [4], an expert system was presented for English-to-Arabic MT. The system is transfer-based and, according to its developers, is easy to use and allows a fast handling of input English text. Khabeer [4], an expert system developed to support Arabic applications,

was used as a machine translation tool. Being Object-Oriented, the Khabeer production system was presented as simplifying lexical and morphological analysis and generation. There has also been some work on research issues related to Arabic machine translation. The first KFUPM Workshop on Information and Computer Science, Dhahran, June 1996, was devoted to Machine Translation with a specific emphasis on the Arabic language.

Despite the existing efforts in the area of AMT, by and large, Arab efforts towards developing the research in such a vital area have been minimal. Most noticeable is the absence of Arab governmental institutions and Pan Arab institutions involvement in the efforts towards building such a vital area of technology. In fact, were it not for scattered researchers across the Arab and Non-Arab world, Arabs would not have seen any MT software today. Most of the commercial products available now have been developed by enthusiastic researchers who had realized the importance of the area and who, often, happen to be based in the West!

As to the evaluation of Arabic MT systems, to the best of our knowledge, work remains non-existent, were it not for the articles [38-40]. These articles present very brief surveys of a number of Arabic MT systems including Transphere, Arabtrans, and Al-Wafy. Nevertheless, these brief assessments were carried out in a very ad-hoc way and thus failed to convey any scientifically convincing conclusion as to the quality of the evaluated systems.

With respect to tools that support the processing of Arabic, a lexicon for Arabic was presented in [41]. The lexicon consists of four modules: a noun inflection module, a verb inflection module, a module for irregular verbs, and a module for pronouns and miscellaneous. Some restrictions were put on the verb forms that could be handled as well as on a number of technical issues related to the various modules. The authors of [42] stressed the importance of a semantic analysis that describes derivation from a semantic viewpoint. They showed how generalization, based on some concept classification (or taxonomy) could be used in Arabic Lexicalization, i.e. in mapping concepts into lexical entries. In [43], a paradigm was presented for the automatic generation of all the Arabic words. The goal of the work was to study the ways Arabic words are generated from their roots. This lead to the development of a knowledge base that contained all the rules used in the analysis and generation of words. The implemented tool allowed an automatic construction of a database of words based on the paradigm.

4.2 The gap between (non-arabic) MT and AMT

Despite the fact that (non-Arabic) MT systems still display weaknesses and the translated texts they output still require fairly substantial post-processing, their state is by far better than that of AMT systems. The following discussion shows the extent of the gap between the two.

In our opinion, the prime reason for the existence of a wide gap between non-Arabic MT and AMT is the attention non-Arabic MT has received. This is because MT has been selected as a *strategic choice* by the officials in the USA, the European Union, Japan and Canada [8; 11]. Such an awareness of the crucial and strategic nature of MT, is still lacking in the Arab world. We are yet to see some official national Arab or Pan-Arab initiative that endeavors to give a boost to AMT, promoting all the related areas that can support its development. While one finds major projects in the aforementioned nations dealing with various aspects of MT and/or NLP and involving tens of experts from related disciplines [8; 17; 34; 35], this effort remains appallingly dismal in the Arab world. In fact, the existing work is mainly based on individual researchers in universities and research centers or on commercial companies, many of which happen to be located in the West.

The previous factor has been detrimental in that funding has been put on research and development of (non-Arabic) MT systems. A number of the early efforts have also attracted funding from security institutions, such as the ministries of defense, in the USA and in Europe to develop MT between European languages, mainly English and Russian. The primary reason for such funding was linked to the cold war. The primary source of funding in the US is the US government. However, in Japan, major funding, research, and development are carried out by private companies. The reason for this notable difference seems to lie in the nature of the perception of the strategic nature of MT. In the US, the alarm was sounded to attract the government's attention to the danger of not monitoring the new technological advances of other nations. This was initially directed to the USSR; however, after the fall of the USSR the main target became Japan. It was therefore logical that the government would put all its weight behind the strengthening of MT research and development [2;8]. In Japan, however, private companies willing to expand their markets of cars, electronics, and other products have heavily invested in MT so that they could quickly produce manuals and documentation to support this expansion [8]. Efforts in France, however, seem to be mainly geared by the country's institutions that fight for the survival of French, a language largely overcome by the supremacy of English [3]. The European Union (EU) has funded a number of ESPRIT projects related to MT and NLP [34; 35]. This was done by the EU for reasons similar to those of the US and Japan as well as for building a united Europe tied by strong links of political, economic, financial, educational, and cultural exchanges. Unfortunately, governmental or private funding of AMT in the Arab world remains negligible, especially given the strategic nature of the area!

Given that (non-Arabic) MT has been recognized very early as a strategic area, and given that funding has poured in to help its development, the discipline has been active for half a century already. This effort has been strengthened during the last 10 to 15 years. Therefore, one finds a number of mature systems, some of which, such as KANT, are using the Interlingua approach. Moreover, a sophisticated "tool box" exists.

These tools include lexicon builders, morphological analyzers, parsers, generators, as well as various tools to solve other NLP problems such as pronoun resolution, semantic disambiguation, and pragmatics-related problems. In the Arab world, including research on AMT and ANLP done in the West, the products remain in their infancy. Grammatical and linguistic tools are very few and yet to be acknowledged as being of a definite quality. While, as mentioned above, the evaluation of MT and NLP tools has been accepted as a sub-area with its own rights, evaluation of AMT is still very sparse and very much non-systematic [38-40]. These attempts at evaluating AMT systems, did not rely on any firm or formal evaluation methodology. Despite such ad hoc methodology, the evaluations have shown some serious shortcomings of the available AMT systems. For instance, a very brief evaluation of Al-Wafy [40] was presented. The author of that article gave examples of English-to-Arabic translations performed by Al-Wafy: the results were quite disappointing. There are still serious contextual reasoning problems with getting the word order or verb tenses right, choosing the right word or expression out of a number of alternatives or translating ambiguous sentences. As a result of a number of drawbacks, translation done by Al-Wafy was shown to produce texts which are not faithful or accurate with respect to the source text, and which are often quite unintelligible. One example of the dismal state of Arabic Machine Translation software is the manual that Compaq (the PC manufacturer and retailer) sent with its personal computers to one of the Saudi governmental institutions. Reading through one of those manuals, the user would have a hard time understanding what the manual is trying to communicate. In fact, the output reflected many deficiencies in translating even basic computer terms. Three AMT systems (Al-Wafy, Al-Mutarjim Al-Arabey, and Arabtrans) were recently evaluated. The evaluation was based on a methodology developed in [44]. The results as detailed in [45] show very severe shortcomings especially related to the grammar and meanings of the translated sentences.

The above comparison, which may sound very pessimistic, is meant to give all researchers and developers interested in AMT, the right dose of realism so that a change is initiated. It is true that the 15 years or so in the life of AMT is a minor in the life of a discipline. Nevertheless, it is important to realize now, and before it becomes too late, that major efforts are urgently needed.

5. Issues Toward the Development of AMT

In order for Arabic Machine translation to get to the stage which MT for other languages has reached, it is essential that Arab public and private institutions get involved in supporting basic research and development of AMT tools. Thus, the very first step towards the development of AMT is to work towards an increased *awareness* of the strategic nature of AMT. This effort should target primarily governmental institutions, private companies, universities, and research centers. This awareness will be

the driving force that will encourage and fund research groups and projects in all areas supporting the development of AMT. Once the awareness exists, the remaining steps can be classified into two categories: technical and organizational.

Technical work for the development of AMT

The technical areas on which Arabs should concentrate are:

1. Research in Natural Language Processing and Artificial Intelligence. This should tackle the following items:
   - i) lexical analysis, including work on tools that help build lexicons and dictionaries.
   - ii) syntactic analysis
   - iii) semantic analysis
   - iv) pragmatics: in particular, the use of intelligent tools for reasoning in context, disambiguation and interpretation.
   - v) knowledge-based tools: for the purposes of inference and reasoning. This would help in problems like automated text summarization, goal recognition, topic shifts, etc.
   - vi) neural networks: could be used to deal with some aspects of the translation task. This would be particularly useful if some patterns of translation, even if in specific settings, can be singled out. It has yet to be proven that Neural Networks are indeed as promising as statistical techniques in general for the field of MT.
   - vii) data mining: the automation of the extraction and classification of objects from texts could then be used by the knowledge-based tools for problems related to pragmatics.
   - viii) corpora: setting up a database of texts and their (model) translations.
   - ix) evaluation of NLP and MT tools: developing methodologies for automated, or semi-automated evaluation.

2. Research on Arabic Natural Language Processing should concentrate on:
   - i) studying and analyzing the Arabic language for purposes of computerization. This mainly involves presenting Arabic morphology and grammar that can be used by computer programmers.
   - ii) developing on-line Arabic dictionaries
   - iii) building lexical, syntactic, and semantic analyzers for Arabic.
   - iv) developing tools that deal with the pragmatics of Arabic.
   - v) collecting corpora of Arabic text classified according to the application domains they cover.
   - vi) agreeing on standards for Arabic, especially for coding Arabic characters.

3. Research on Arabic Machine Translation:

i) developing a methodology for the evaluation of AMT systems
ii) evaluating existing AMT systems and identifying their shortcomings. This should be done on a continuous basis.
iii) working on the development of *transfer* tools for the pairs Arabic-English (primarily) and Arabic-French.
iv) working on the development of an *Interlingua* that can cater for the complexity and expressiveness of Arabic as well as the main language targets, namely English and French.
v) developing prototypes of commercial AMT systems based on the various MT approaches (transfer and Interlingua).
vi) working on commercial products that could be generated as by-products of research on AMT. These could be products like grammar checkers, vowelizers, language tutoring systems, etc. Revenues generated from such products should be used for sponsoring research, making it self-supporting.

Organizational work for the development of AMT

In the organizational areas, Arabs should:

1. Work on a resolution to be supported by the Arab league, decreeing AMT as an area of strategic importance for the Arab world.

2. Establish an Arab Center for ANLP and AMT. Having passed the above resolution or given the existence of an Arab state aware of the strategic nature of AMT, the next step would be the establishment of an Arab center for ANLP and AMT logistics. The center will play a pivotal role in crystallizing the aforementioned targets. More specifically, the center will:

i) gather Arab researchers working on linguistics, computational linguistics, Artificial Intelligence, and Computer Science with Arabic being a prime target. In particular, encourage collaboration between the disciplines of Linguistics and Computer Science in the Arab world. Both disciplines currently have different outlooks and as such do not work for a common goal. A collaboration between researchers and instructors in both areas could produce a new breed of Linguists who have enough formal training in computers and computing to be able to take a real part in the development of MT systems.
ii) host and sponsor research on all items listed under “Technical Work” above.
iii) dedicate manpower for the building of a central database and an archive center. The former will be a repository of software, lexicons, and corpora, which would support ANLP and AMT. The latter would

collect any literature related to NLP or MT, with ANLP and AMT being prime targets.

3. Build stronger relations between universities, research centers, and the industry.

4. Sponsor pan-Arab Projects. Such projects could be conceived at the image of the European ESPRIT projects which bring into collaboration researchers and developers from universities, research centers and private and state companies, from all over Europe, to work on problems of common interest. Implementing this suggestion should be done carefully so that it does not fall into the same mistakes as in the ESPRIT project. Instead Arabs should learn from the mistakes of Europe. The main problem to avoid is that such a pan-Arab project should gather researchers based on their expertise and their projected benefit to the area of AMT rather than selecting researchers based on a criteria of equal country representation.

5. Sponsor conferences, exhibitions, and trade shows. Such gatherings, related to ANLP and AMT, would provide researchers and developers with the opportunity to meet and discuss their findings, their products, and the research problems they face.

6. Sponsor Arab researchers in AMT, and Arabic NLP in general, to spend periods of time in countries which have strong research programs in MT or NLP. This way, existing Arab expertise can learn, from the source, how others have developed the field of MT. In addition, a good idea might be to sponsor researchers from non-Arab countries, who have a proven record in the field of MT, to come to Arab countries and train Arab experts

6. Conclusion

In this paper, we have tried to draw the reader's attention to the importance of Machine Translation for the Arab world. The importance of AMT stems from the fact that MT outside the Arab world is already a huge industry. The American, European, Japanese and Asian governments and private companies have, long ago, realized the importance of Machine Translation. We pointed out that the MT market is an important one and that it is yet to expand with the growing trend towards globalization[46]. The fact that English has become the de-facto language of trade and communications puts increased pressure on the non-English speaking nations to develop tools, especially computerized ones, so as to keep up with the scale of the task. This is in order to enable them to benefit from the scientific, economic, political, and cultural benefits of globalization.

In this paper, we have highlighted many reasons why governments and private companies should invest in Machine Translation. MT is important indeed for any nation/country, no matter where it stands on the development scale. Different reasons drive different nations/countries; still, they should all realize that MT, from and into their own languages, is a strategic area that they should develop.

Arabs, in particular, should concentrate more on AMT. More specifically, governments and Pan Arab organizations should give AMT a boost by establishing computational linguistics centers. Such centers would enrich Arabic research in Natural Language Processing and AMT and provide researchers with the necessary tools for progress in Arabic NLP. Operation between Arab academics and industrialists should be increased. A wide gap already exists between the current AMT systems and their non-Arabic counterparts. In our assessment, the primary reason for the existence of such a gap is that the officials in countries like the US, EU, Japan, and Canada have designated MT as a strategic choice. This has not been done yet in any Arab country and governmental and private funding of AMT in the Arab world remains negligible.

Among the specific areas that would benefit the development of AMT as well as its users is the evaluation of AMT systems. Developing a methodology for evaluating AMT systems is important in scrutinizing any MT system, analyzing the quality of its output, classifying errors it makes, and possibly improving it as a result of such an evaluation. In fact, the evaluation of MT systems is a research area in itself with its own conferences and literature. The evaluation of MT systems should also involve the evaluation of the lexicons of such systems which has grown into a separate area of research.

We have finally outlined in this paper some basic steps that should urgently be taken by Arab governments and Pan-Arab institutions in order to raise the level of AMT to where MT for other languages is now and beyond.

References

[1] Lockard, Joe. “Resisting Cyber-English.” *Bad Subjects*, No. 24 (Feb. 1996).
[2] Hutchins, W. John and Somers, Harold L. *An Introduction to Machine Translation*. Academic Press, 1992.
[3] “Language Restrictions Create Own Tangled Web.” *In: The Detroit News*, 20 February 1997.
[4] *Proceedings of the First KFUPM Workshop on Information and Computer Science: Machine Translation,* Dhahran (1996), 23-44.
[5] *Proceedings of the KSU Conference on the Generalization of Arabization and the Development of Translation in the Kingdom of Saudi Arabia,* Riyadh, 1998.
[6] Allen, James. *Natural Language Processing*, The Benjamin/Cummings Publishing Company, 2nd ed., 1995.

[7] Russell, Stuart and Norvig, Peter. *Artificial Intelligence, A Modern Approach*, Prentice Hall, 1995.
[8] White, Robert M. *Machine Translation Technology: A Potential Key to the Information Age*. Report of the FCCSET Committee on Industry and Technology, PB-93-134336, Office of Science and Technology Policy, Washington, D.C, January 1993.
[9] Hedberg, Sara. "Machine Translation Comes of Age." *AI Expert*, 9, No. 10 (October 1994), 37-41.
[10] Machrone, Bill. "Publishing on the Web is Good Business." *PCWEEK*, 12, No. 8 (February 27, 1995), pp. 60.
[11] Equipe Consortium Limited. "Survey of Machine Translation Products and Services." Summary of a report to the European Commission, September 1996.
(see http://www2.echo.lu/langeng/reps/mtsurvey/mtsurvey.html)
[12] Johnson, R. Colin. "Machines Attack the Language Barrier: Translation of Written Text Possible, but with Limited Fluency." *Electronic Engineering Times*, No. 814 (September 12, 1994), 37-38.
[13] Ramdani, B. "French Fails to Compete with English on the Internet." *Al-Sharq Al-Awsat Newspaper*, 1997.
[14] Al-Fantoukh, A. and Al-Badr, B. "A Survey on the Use of the Internet Using the Arabic Language: Preliminary Results". *Proceedings of the First Workshop on the Arabization of the Internet.* King Saud University, Riyadh, 1998 (12 Muharram 1418H).
[15] Darke, Diane. " Machine Translation for Arabic". *Language Monthly* (January 1986), pp.10.
[16] Johnson, R. Colin, "Machines Attack the Language Barrier: Translation of Written Text Possible, but with Limited Fluency." *Electronic Engineering Times*, No. 814 (September 12, 1994), 37-38.
[17] Carbonell, Jaime G., Mitamura, Teruko and Nyberg, Eric H. 3rd ed., *The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics,* Center for Machine Translation, Carnegie Mellon University, PA.
[18] AlKarmi, Aroud. "Transcend Translates from English to Five Languages: The First Program for Instantaneous Translation from Fujitsu and Xerox Studies." *Al-Computers, Communication and Electronics,* Jan. 1997.
[19] Onyshkevich, B. and Nirenburg, S. "A Lexicon for Knowledge-Based MT." *Machine Translation*, pp. 5-57, Vol. 10, 1995.
[20] Dorr, B. J. and Klavans, J. *Journal of Machine Translation, Special Issue on Building Lexicons for Machine Translation,* Kluwer, 1995.
[21] Steffens (Ed.) "Machine Translation and the Lexicon." In: *Lecture Notes in Computer Science,* Carbonell, G. and Siekmann, J. (Eds.), Vol. 898, Springer-Verlag, 1995.
[22] Nirenburg, S. and Levin, L. "Syntax-Driven and Ontology-Driven Lexical Semantics." In: *Lexical Semantics and Knowledge Representation*. Pustejovsky, J. (Ed.), Springer-Verlag, Heidelberg, 1992.
[23] Lehrberger, J. and Bourbeau, L. "Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation." *Linguisticae Investigationes*, Vol. 15, 1998.
[24] Dyson, M.C. and Hannah, J. " Towards a Methodology for the Evaluation of Machine-Assisted Translation Systems." *Computers and Translation*, Vol. 2 (1987), 163-176.
[25] Melby, A.K. "Lexical Transfer: Between a Source Rock and a Hard Target." In: *Proceedings of the International Conference on Computational Linguistics,* (1988), 411-419.
[26] Nagao, M. " Evaluation of the Quality of Machine-Translated Sentences and the Control of Language. " *Journal of the Information Processing Society of Japan*, 26, No. 10 (1985), 1197-1202.
[27] King, M. and Falkedal, K. "Using Test Suites in Evaluation of Machine Translation Systems." In: *Proceedings of the International Conference on Computational Linguistics,* (1990), 211-216.
[28] Vasconcellos, M. (Ed.), *Technology as Translation Strategy*, American Translators Association Scholarly Series, Vol. 2, 1988.
[29] Van, Slype. "Systran: Evaluation of the 1978 Version of the Systran English-French Automatic System of the Commission of the European Communities." *The Incorporated Linguist,* pp. 86-89, Vol. 18, 1979.
[30] Wilks, Y. *Systran: It Obviously Works but How Much Can It Be Improved?*. Report MCCS-91-215, Computer Research Laboratory, New Mexico State University, Las Cruces, 1991.
[31] Sinaiko, H. W. and Klare, G. R. "Further Experiments in Language Translation: Readability of Computer Translations." *ITL*, Vol. 15 (1972), 1-29.

[32] Sinaiko, H. W. and Klare, G. R. “Further Experiments in Language Translation: A Second Evaluation of the Readability of Computer Translations." *ITL*, Vol. 19 (1973), 29-52.
[33] White, J.S. “Machine Translation Program: 3Q94 Evaluation.” Advanced Research Projects Agency, 1994. (http://ursula.georgetown.edu/mt_web/3Q94FR.htm).
[34] Klein, J. and Lehmann, S., Netter, K. and Wegst, T. “DiET in the Context of MT Evaluation.“ KONVENS98, 5-7 October 1998.
[35] Bevan N. *et al*. Proceedings of the Second EAGLES II Workshop on Evaluation in Human Language Technology, Geneva, 8-9th September 1998.
[36] Nyberg, Eric H. 3rd, Mitamura, Teruko and Carbonell, Jaime G. “*Evaluation Metrics for Knowledge-Based Machine Translation."* Center for Machine Translation, Carnegie Mellon University, PA.
[37] Carbonell, Jaime G., Mitamura, Teruko and Nyberg, Eric H. 3rd, *"Evaluating KBMT in the Large."* Japan-US Workshop on Machine-Aided Translation. Nov. 22-24, Washington D.C., 1993.
[38] Jihad, A. “Has the Arabic Machine Translation Era Started?.” *Byte-Middle East*, November 1996, (in Arabic).
[39] Qendelft, G. “The Translation Program Al-Wafy is Useful for Getting a General Understanding of Letter Written in English.” *Al-Hayat newspaper*, Number 12657, 25 October 1997.
[40] "The Machine Translator: Al-Wafy." *Arabuter,* 8, No. 71 (September 1996). (In Arabic).
[41] AlNeami, A. and Gregor, J. J. “The Arabic Computational Lexicon.” In: *Proceedings of the 5th International Conference and Exhibition on Multi-Lingual Computing*, (1996), 3.3.1-3.3.22.
[42] Al-Jabri, S. and Mellish, C. “Using Classification for Mapping Semantic Representations into Arabic Lexical Entries.” In: *Proceedings of the First KFUPM Workshop on Information and Computer Science: Machine Translation*, Dhahran (1996), 23-44.
[43] Al-Hannach, M. “A Formal Linguistic Tool for Machine Translation Programs.” In: *Proceedings of the First KFUPM Workshop on Information and Computer Science: Machine Translation.* Dhahran (1996), 1-22.
[44] Guessoum, A. and Zantout, R. "Machine Evaluation of Machine Translation System Lexicons." In preparation.
[45] Al-Sikhan, A. and Zantout, R. and Guessoum, A. “Automating the Evaluation of Machine Translation System Lexicons.” *Proceedings of the 7th International Conference on the Applications of Artificial Intelligence*, Cairo, 1999.
[46] *Globalization: Creating New Markets with Translation Technology*, OVUM Ltd., 1995.
[47] *Ovum Evaluates: Translation Technology Products*. OVUM Ltd., June 1995.

الترجمة الآلية العربية : خيار استراتيجي للعالم العربي

راشد زنتوت و أحمد قسوم

كلية علوم الحاسب والمعلومات، جامعة الملك سعود،

ص.ب ٥١١٧٨، الرياض ١١٥٤٣، المملكة العربية السعودية

(قدّم للنشر في ١٩٩٩/٣/١٣م؛ وقبل للنشر في ١٩٩٩/٩/١٤م)

ملخص البحث . لقد نتج عن الانتشار الواسع لاستعمال الحاسوب والشبكات الحاسوبية – بما فيها الإنترنت – ضغط هائل على الدول والأمم غير الناطقة بالإنجليزية. وأصبح من الصعب والمكلف جدا – إن لم نقل من المستحيل – مواكبة التطورات التقنية والإقتصادية والإجتماعية والثقافية والسياسية. وهذا الواقع يضع هذه الدول والأمم في حالة من الإستلاب الثقافي، إذ أنها أصبحت ملزمة بقبول "الوسيط اللغوي"، أي الإنجليزية، وما يصاحبها من "حمولة" ثقافية

إن الهدف من هذه الورقة هو محاولة جلب انتباه كل العلميين والإداريين والمفكرين العرب إلى أهمية، بل البعد الإستراتيجي، للترجمة الآلية. ومن خلال هذه الورقة، نلقي نظرة عامة على المستوى التقني الذي بلغته الترجمة الآلية في الغرب، بالإضافة إلى البرامج التي طُورت هناك ونقارن وضعها بوضع الترجمة الآلية العربية. ومن هذه المقارنة ننطلق إلى تقديم مقترح يهدف إلى تطوير الترجمة الآلية واللسانيات الحاسوبية في الوطن العربي.