MACHINE TRANSLATION: A SURVEY.

Ioan – Lucian POPA
Bacau University, Romania

1. Introduction

This paper aims at providing the reader with a comprehensive survey of the human quest for devising software for the translation of natural languages. This human quest resulted in what is called 'machine translation' (MT). It has left its imprint on the evolution of linguistics and computer science in the latter half of the 20th century and has an ever increasing impact on the developments in the fields in the early days of the 21st.

Ideally, the objective of MT would be to bring together all the information necessary for translation in one computer program, so that a text could be translated without human intervention.

MT should not be mistaken for CAT (computer-assisted translation) that employs a variety of *computer-based translation tools* i.e. electronic dictionaries and glossaries, terminology management systems, translation databases, and translator workstations to assist the translator's in his/her work and ensure accuracy and speed, i.e. quality and quantity.

The evolution of human society has made it necessary to find solutions for more efficient methods of translation as the demand for translations cannot be met because there are not enough human translators and also because of a series of misconceptions about translation. Human translation is expensive since it requires very serious training, the productivity of the human translator is essentially limited (2,000 words per day) and the processes it needs to duplicate are complex.

According to W. John Hutchins (2006), one can identify three major types of MT: (1) *machine translation for dissemination (MTd)*, (2) *machine translation for assimilation (MTa)*, and (3) *machine translation for communication (MTc)*. In the case of MTd, the quality of the translation is paramount. Such translations, after an initial MT processing, must be subjected to a process of editing by human translators, with a view to obtaining the optimum quality. This mode is identified as *human-aided machine translation* (HAMT). In the cases of MTa and MTc the quality is of secondary importance.

2. Beginnings

The origins of MT can be traced back to the seventeenth century ideas of universal and philosophical languages and of 'mechanical' dictionaries; nevertheless, the first practical solutions were suggested in the 20th century, in 1933, by two individuals: Georges Artsrouni in France, and Piotr Smirnov-Trojanskij in Russia. Artsrouni's idea was that of a general-purpose machine which could also be utilized as a mechanical multilingual dictionary. Trojanskij's idea was also a mechanical dictionary, but he developed the initial idea with proposals for coding and interpreting grammatical functions using sets of universal symbols with a view to creating a multilingual translation device.

In 1946 and 1947, Andrew Booth and Warren Weaver suggested that the newly invented computers could be used for translating natural languages. In 1948, Booth co-operated with Richard H. Richens (Cambridge, UK) on a model of morphological analysis for a mechanical dictionary.

BUPT

In July 1949, Warren Weaver (Rockefeller Foundation) approached the specific problem of ambiguity and offered solutions that were based on his knowledge of cryptography, statistics, information theory, logic and language universals.

In May 1951, the philosopher, mathematician and linguist Yehoshua Bar-Hillel started full-time research on MT at the Massachusetts Institute of Technology, USA, and he organized, in June 1952, the first MT conference. This was a seminal moment as debates made it clear that full automation of translation and optimum results were impossible, and that human involvement either before or after computer processes, i.e. pre-editing and post-editing, respectively, would be indispensable.

On January 7, 1954 Léon Dostert (Georgetown University) joined efforts with Peter Sheridan (IBM) and Paul Garvin (Georgetown), in a project which resulted in the first demonstration of an MT system. The demonstration was sufficiently impressive to stimulate large-scale funding of MT research in the USA and trigger the launch of MT projects al over the world.

In the same year, the journal *Mechanical Translation* was initiated by William Locke and by Victor Yngve; it would assemble in its pages some of the most significant papers on the topic until 1970.

In 1954 and 1955, MT research groups were set up in Cambridge, England, (Margaret Masterman), in Milan (Silvio Ceccato), then the first Soviet, Chinese, and Japanese projects were initiated.

In 1955, the first MT book was published, *Machine Translation of Languages,* a collection edited by A.D. Booth and W.N. Locke.

3. Breaking new ground

The period 1956-1966 was dedicated to breaking new ground; the expectations were high, the developments were extremely numerous and we will have to limit ourselves to mentioning only the most important ones.

MT research, in its early days, could barely rely upon linguistics and the state it was in then and, consequently, the research methods ranged from empirical trial-and-error approaches to what was to become 'computational linguistics'.

There were three basic approaches to MT in the decade under scrutiny. The *'direct translation' model* (the programming rules were conceived and developed for translation from one source language (SL), into a target language (TL), with a minimal amount of linguistic theory) led to an oversimplification that is not characteristic of natural languages. This model was favored especially by empiricists. They resorted to statistical analyses of actual texts with a view to identify dictionary rules that were all too frequent "of an ad hoc nature, with little or no theoretical foundation" (Hutchins, 2006).

The so-called *interlingual approach* or *pivot language strategy* concentrates "on a trade or transfer of meaning based on an analysis of one language pair alone" (Gross, 1992) and operates on the theory that one can devise a set of abstract language-neutral representations such as codes or symbols independent of both SL and TL, i.e. an intermediate 'language'- in at least one case, a form of Esperanto. Such a language can encode sufficient linguistic information to serve as a universal intermediate stage - or *pivot point* - enabling translation back and forth between numerous pairs of languages, despite linguistic or cultural differences.

The third approach was the *transfer approach*: texts underwent a conversion through a transfer stage from disambiguated representations of SL texts to equivalent TL representations; translation consisted of three phases: *analysis* (it describes the SL document linguistically and utilizes an SL dictionary), *transfer* (it transforms the results of the analysis stage and establishes the linguistic and structural equivalents between the two languages), and *generation* or *synthesis* (it produces a document in the TL on the strength of the linguistic data of the SL using a TL dictionary).

The 'interlingua' model and the transfer approach were favored by various groups of researchers; the more pragmatic groups postponed the solution of semantic problems.

In Europe and the Soviet Union, research groups were compelled to concentrate on theoretical issues due to the inadequacy of the technical means.

Erwin Reifler and his research group (University of Washington, Seattle) developed the *dictionary-based 'direct' approach* that entailed the compilation of large bilingual dictionaries where lexicographic information was not used solely for identifying TL lexical equivalents for SL items but also for solving grammatical problems without the use of syntactic analysis.

Another direction was that of developing methods based on the analysis of language corpora. A group of researchers at the RAND Corporation carried out statistical analyses of a large corpus of Russian physics texts, compiled bilingual glossaries, and gathered grammatical information that were later used for a translation computer program.

Researchers at Georgetown University developed the *Georgetown Automatic Translation (GAT)* that had three levels of analysis: morphological, syntagmatic, and syntactic. GAT was initially put into operation on the SERNA system and, then, a new version was successfully installed by Euratom at Ispra (Italy) in 1963 and by the US Atomic Energy Commission in 1964.

From 1954 to 1960, Anthony Oettinger and his group at Harvard University compiled a substantial Russian-English dictionary considered a precursor of the now common computer-based dictionary aids.

From 1959 onwards a new direction for research could be identified: the *predictive syntactic analyzer* - a system for the identification of acceptable sequences of grammatical items and the probabilistic prediction of following items (initiator: National Bureau of Standards, Ida Rhodes); successive improvements led later to William Woods' familiar *Augmented Transition Network* parser).

From1953 until 1965, at MIT, under Victor Yngve, syntax was placed at the centre of MT research but, the conclusion was that a semantic barrier was reached and further progress in the field was extremely difficult.

The Linguistic Research Center at the University of Texas, (initiator Winfried Lehmannin, 1958), focused on basic syntactic research and reversible grammars in order to realize bi-directional translation within the framework of an essentially syntactic transfer approach. At the University of California, Berkeley, Sydney Lamb developed his 'stratificational grammar', with an architecture of relationships paralleling that of computers and he also highlighted the importance of developing efficient electronic dictionaries and a linguistic theory appropriate for MT.

The interlingua approach was not popular in the US; but it was elsewhere in the world: Margaret Masterman and the Cambridge Language Research Unit worked along two basic lines of research: the development of a prototype interlingua system and of tools for improving and refining MT output; Silvio Ceccato (Milan) aimed at developing an interlingua based on cognitive processes and he is considered one of the precursors of the neural networks.

In the Soviet Union, research was somewhat similar to that in the United States (amplitude and mix of empirical and basic theoretical approaches): D.Y. Panov's group's research on English-Russian translation; basic research at the Steklov Mathematical

BUPT

Institute (A. Ljapunov, O. Kulagina and I. Melčuk); the interlingua approach was represented especially by the group of Nikolaj Andreev (the Leningrad State University).

By the end of the period under discussion, MT research groups had been established in many countries throughout the world, including most European countries, China, Mexico, and Japan.

The 1950s was a very optimistic period when some researchers even predicted the imminent implementation of fully automatic MT systems. In 1960, Bar-Hillel stated that the creation of *fully automatic high quality translation* (FAHQT) systems was not realistic and impossible in principle, but his pessimism was not completely justified.

In 1964, the Automatic Language Processing Advisory Committee (ALPAC) examined the situation of MT and, in 1966, the ALPAC report concluded that "there is no immediate or predictable prospect of useful machine translation". The committee recommended the development of machine aids for translators, such as automatic dictionaries, and the continued support of basic research in computational linguistics; the most important effect of the ALPAC report in the US and elsewhere was that the research changed its focus from the direct translation approaches, to the indirect models, both interlingua and transfer-based.

4. Gaining speed

In the interval 1967-1976, while in the United States manifest interest in MT diminished, especially in Canada and Europe the situation was different. Due to bilingualism in Canada and multilingualism in the European Community (E.C., i.e. today's European Union) the problems of translation were acute and development in MT continued.

In 1970, the *TAUM* (*Traduction Automatique de l'Université de Montréal*) project, and the *Météo* system for translating weather forecasts (successful operation since 1976) produced results. Besides *Météo*, two other sublanguage systems appeared: in 1970, *TITUS* (at the Institut Textile de France), a multilingual system for translating abstracts written in a controlled language, and, in 1972, *CULT* (designed for translating mathematics texts from Chinese into English) of the Chinese University of Hong Kong.

The main experiments of the decade focused on approaches from the essentially interlingua perspective (e.g. 1960 - 1971, Bernard Vauquois of Grenoble University, France and his group developed a system for translating mathematics and physics texts from Russian into French). Towards the end of the period, the prospects of the interlingua approach seemed to become narrower and narrower and the less ambitious transfer approach seemed to offer better prospects.

The period 1976-1989 is characterized mainly by the implementation on a large scale of previously developed systems and the launching of several operational and commercial systems and less by important theoretical breakthroughs: 1976, the Commission of the EC purchased an English-French version of the *Systran* MT system that had already been used at the US Department of Defense; NATO, International Atomic Energy Authority, General Motors, Dornier, Aérospatiale, etc., a system from the Logos Corporation (Bernard E. Scott), and, at the end of the 1980s, the commercial *METAL* German-English system. Special-purpose systems were also developed, e.g. systems for the Pan American Health Organization (Washington, D.C.). In Japan, most of the computer companies developed software for computer-aided translation. The languages of choice were English, Korean, and Chinese and the majority of them were low-level direct or transfer systems.

A new direction emerges in the 1980s: systems designed for personal computers,

BUPT

i.e. *Weidner* (1981) and *ALPS* (1983). Towards the end of the decade, *PC-Translator* from Linguistic Products, *GTS* from Globalink and the *Language Assistant* series from MicroTac).

5. Renewed interest in the theoretical aspects of MT

The period 1976 - 1989 is characterized by a renewed general interest in the theoretical aspects of MT. The main tendency was the quasi-general adoption of the *three-stage transfer-based approach.*

The Grenoble group (*GETA - Groupe d'Etudes pour la Traduction Automatique*) switched allegiances from the interlingua system and embarked on the development of a system which was soon considered as the prototype of the new generation of linguistics-based transfer systems. The *Mu* system was developed at the University of Kyoto and, since 1986, it has been converted into an operational system used by the Japanese Information Center for Science and Technology for the translation of abstracts. In the mid-1970s, MT research at Saarbrücken (Germany) resulted in the development of a multilingual transfer system *SUSY (Saarbrücker Übersetzungssystem).*

In the late 1980s, one can identify a renewed interest in interlingua systems with notable results: in the Netherlands, the *DLT (Distributed Language Translation)* system of BSO, Utrecht, led to the construction of large lexical databases, and the *Rosetta* project focused upon the use of Montague grammar in interlingual representations and the exploration of the reversibility of grammars.

Another direction in MT research was the implementation of the results of explorations in artificial intelligence (AI): Yorick Wilks' work on 'preference semantics' and 'semantic templates', in the mid-1970s, and Roger Schank's research at Yale University, and the projects that applied knowledge-based approaches at Carnegie-Mellon University in Pittsburgh. In the late 1980s, the Carnegie-Mellon group elaborated its *KANT* prototype system and could implement it in an operational knowledge-based system.

One must also mention a major change of perspective: the syntactic orientation was replaced by 'lexicalist' approaches, the results being the *Electronic Dictionary Research* project (late 1980s, in Japan), the massive database and dictionary resources (the Linguistic Data Consortium, United States, and the European Language Resources Association), all these being available to our days.

During the 1980s, computer aids were developed, and in the early 1990s, all these and more were integrated in the so-called 'translator's workstation/workbench': *Translator's Workbench* (Trados), *Transit* (STAR AG), *Translation Manager* (IBM), *Eurolang Optimizer*, followed in the 1990s and early 2000s by many others: *Déjà Vu* (Atril), the *SDLX* system (SDL), *XMS* (Xerox), *LogiTerm* (Terminotix), *MultiTrans* (MultiCorpora), *WordFast* (Champollion), *MetaTexis*, and *ProMemoria.*

And yet, in spite of progress in MT, at the beginning of the 1990s the translation market was as follows (Loffler-Laurian, 1996): Europe and the United States - human translation (HT) 300 million pages as compared to MT 2.5 million pages; Japan - HT 150 million pages versus MT 3.5 million pages. Market analysts predicted that this percentage will not change radically by 2007.

6. Research since 1989

Until the end of the 1980s, MT research favored the rule-based approach, but since 1989, the supremacy of the rule-based approach has been successfully challenged by the *corpus-based* approaches. Since then, *statistical machine translation (SMT)* has

BUPT

become the major focal point of various research groups. Another corpus-based approach is the *example-based/memory-based approach (EBMT)* that was initiated by Japanese researchers.

The rule-based approaches have not been discarded altogether: the *CAT2* system (Saarbrücken), the *PaTrans* system (for Danish/English translation of patents), the *LMT* (*Logic-programming Machine Translation*) project (IBM research centers in Germany, Spain, Israel and the USA). Neither was the interlingua approach: the *CATALYST* system at Carnegie-Mellon University, the *ULTRA* system (New Mexico State University), the *UNITRAN* system (University of Maryland), and the *Pangloss* (the universities of Southern California, New Mexico State and Carnegie-Mellon). In the US, this period marks the end of the negative impact of the ALPAC report.

Speech translation is another of the most significant and ambitious developments since the late 1980s. The pioneers were British Telecom (late 1980s, a spoken language phrasebook type system, ATR, Japan (1986, a system for telephone registrations at international conferences and for telephone booking of hotel accommodation). Other important projects are *JANUS* (Alex Waibel, Carnegie-Mellon University), *C-STAR* (Consortium for Speech Translation Advanced Research) and VoxTec for the US military, the *Phraselator* (a one-way phrase-based voice-to-voice machine translator into numerous languages).

There has been an accelerated increase in the use of MT systems since the 1990s. One of the fastest growing areas is *software localization* and *localization of web pages* on company sites. Also, systems for specialized domains and users, the so-called *controlled language systems* have been developed. A. Gross (1992) identifies "a more extreme form of pre-editing, *controlled language,* the severely limited vocabulary which is used by a few companies to make MT as foolproof as possible". John Newton (1992:40), identifies one of the most successful machine translation applications at Perkins Engines, as an example of what can be achieved when a system is introduced in a thoroughly planned and methodical way into a restricted domain environment to process controlled-language source texts.

Since the early 1990s, scores of systems for personal computers have appeared: among the first systems to be sold extensively for use on personal computers were *PC-Translator* (Linguistic Products, Texas) and *Power Translator* (Globalink).

In order to provide an illustration of the evolutions in the field of commercial machine translation systems and computer-aided translation support tools we should mention the fact that if, in its preliminary draft edition of 1999 the *Compendium of Translation Software,* (a directory of commercial machine translation systems and computer-aided translation support tools compiled by J. Hutchins on behalf of the European Association for Machine Translation and the International Association for Machine Translation) had 69 pages; in its 13th edition of June 2007, it has 126.

7. MT on the Internet

Since the mid-1990s, when the Internet became more and more available to ever increasing numbers of users, it has exercised a great influence upon MT development.

First, MT software products specifically designed for offline translation of web pages and e-mail messages made their debut.

Then, Internet-based online translation services for on-demand translation have been offered. A precursor was the service offered in France by Systran on the Minitel network during the 1980s. Then, in 1995, the idea was revived by CompuServe. Then, *Babelfish* was offered by AltaVista and various other online services offered online

BUPT

versions of their systems: Softissimo with *Reverso*, LogoMedia with *LogoVista* and *PARS*, etc. This is a developing story and it needs a separate survey.

8. Assessment of MT

MT evaluation has become a major concern for all those involved in the activity and in the 1990s there were numerous workshops dedicated specifically to the problems of evaluating MT, e.g. Falkedal 1994, Vasconcellos 1994, and the workshops attached to many MT conferences. Coherent methodologies have been developed by organizations such as, for instance, Japan Electronic Industry Development Association (JEIDA 1992) and Advanced Research Projects Agency in the US (ARPA 1994). At first, MT quality assessment was performed by humans and they focused upon comprehensibility, intelligibility, fluency, accuracy and appropriateness but this seemed to be time-consuming and too laborious; especially since 2000, automatic or semi-automatic methods have been developed that rely on the application of statistical analysis to the automatic evaluation of MT systems. We can mention some initial successes: *BLEU* (IBM) and *NIST* (National Institute for Standards and Techniques); nevertheless, one must mention that there is no general agreement upon the reliability of such means of evaluation, and research upon devising infallible ones continues and intensifies.

9. Conclusion

In its early days, MT's implicit goal was the automatic translation of documents at a quality equal to that of the output of best human translators, notwithstanding their nature and complexity. It soon became obvious that this objective was unfeasible in the predictable future. Revision by specialized humans of MT output was indispensable if the results were meant for publication. Nevertheless, MT was not doomed as it was found that the rough and ready MT output could be of use to those who wanted a shortcut to the gist of a text written in a language that was not accessible to them. This use of MT as a tool of assimilation, for information gathering and monitoring was ignored or disregarded. MT was unrealistically supposed to produce human-quality translations for dissemination and for publication. Fortunately though, for many years, the potential of HAMT has been put to use by multinational corporations and other multilingual institutions such as for instance the European Union, and has contributed to the solving of massive translation tasks.

The first obvious strong point of MT is productivity. An MT system can process huge amounts of information at speeds unattainable by human translators (a decade or so ago, there have been reports of 700 pages translated per day, by one system). These performances, of course, are dependent upon how well the system has been conceived with respect to the translation task: the more specific, the better defined the task is, the better the results. This brings us to another an enormous advantage of MT.

A huge amount of work has already been spent on codifying specialized vocabulary and entering it into the computer's dictionary. Thus, translations of specialized texts can reach high levels of accuracy due to it. If the texts to be translated are not specialized, if they are literature or other types that require massive operations of disambiguation, the levels of accuracy will be deficient, the post-editing will be time-consuming.

There is another plus for MT: uniformity of vocabulary within the same translated text. Once the dictionary-building task has been correctly performed, the computer will constantly offer the same translation of the same structure, of the same term.

BUPT

MT is thus cost-effective for quantitative and/or rapid translation of specialized documentation and software localization materials, while the human translator has been and will be the unavoidable solution for non-repetitive linguistically complicated texts (literary, legal, etc.), and even for individual texts using highly specialized vocabularies.

The solutions to MT problems that have been devised have turned out to be invaluable tools for human translators, e.g. computer-based translation support tools. All these tools are combined in translator workstations that frequently incorporate full MT systems as well.

The coming of online translation on the Internet has given the kiss of life to MT and has triggered most significant changes, with potentially extensive implications for the future. Massive availability of information in many languages has created a rapidly growing demand for real-time online availability of rough translations to quench instant communication and information needs. These services have already evolved from providing low-quality MT to providing add-on human translation services such as post-editing or full translation.

A close scrutiny of the evolutions in MT makes it all too obvious that the adoption of multiple approaches will lead to success in achieving good-quality automatic translation and also that MT continuously adapts to its users' needs and is still light years away from exhausting its potential.

References

1. Gross, Alex, "Limitations of computers as translation tools " in Newton, John, (ed.) *Computers in Translation: A Practical Appraisal,* New York: Routledge, 1992
2. Hutchins, John "Compendium of Translation Software", *http://www.eamt.org/compendium.html,* various editions, 1999 – 2007
3. Hutchins, W. John, "Machine translation: a concise history" *http://ourworld.compuserve.com/homepages/ WJHutchins,* 2006
4. Loffler-Laurian, Anne-Marie. *La traduction automatique.* Villeneuve d'Ascq: Presse Universitaire du Septentrion, 1996.
5. Newton, John, (ed.) *Computers in Translation: A Practical Appraisal,* New York: Routledge, 1992.
6. Newton, John,, "The Perkins Experience", in Newton, John, (ed.) *Computers in Translation: A Practical Appraisal,* New York: Routledge, 1992
7. Robinson, Douglas, *Performative Linguistics: Speaking and Translating as Doing Things with Words.* London: Routledge, 2003.

BUPT