

Artificial Intelligence and Machine Translation: present developments and future prospects

Josette M. Coughlin

Artificial Intelligence (AI) accomplishments are well known to the general public in the field of mechanical engineering and medicine. Robots equipped with sophisticated tactile and visual sensors able to perform very precise and complicated manipulations are favorite subjects for the news media. Medical diagnostic programs such as Caduceus (Miller 82), which suggest additional tests to be performed on a patient before they predict which disease or diseases the patient may have, receive a great deal of public attention. While the field of robotics and expert systems has been highly publicized, AI research in natural language comprehension may well be less spectacular but no less important. Machine Translation (MT) which is a subarea of AI research in natural language processing has not been at the forefront of research and development in AI. However it has proven to be an excellent testbed for AI theories and has benefited from several AI applications.

Our purpose here will be to study improvement in MT due to AI strategies. Originally language analysis in AI was not concerned with the translation of texts from one language to another but rather with more general problems of information processing such as the development of question and answer programs or the machine's ability to extract information in order to summarize it or organize it in a given manner. To accomplish these tasks AI research has developed several techniques including the use of semantic parsing, the consultation of expert systems, and knowledge databases as well as the creation of high level programming languages designed for symbolic rather than numeric computing such as LISP (Winston 84) and PROLOG (Kowalski 1985). All the above mentioned AI techniques will be described in detail when applied within various MT systems.

In an attempt to define AI, a researcher stated: "Artificial Intelligence (AI) is the study of how to make computers do things that people are better at" (Rich 83). This general definition of AI could have been written specifically for the field of MT, for there is no doubt in any professional translator's mind that human translation is superior to MT. MT may be an evil, but it is a necessary evil given the ever increasing amount of information needing to be disseminated world wide. From these premises it is logical to conclude that every effort should be made to improve the quality of MT including the use of AI strategies if they prove to be helpful.

Interestingly it was MT vocal detractors who first mapped out the area of language processing which was to become the domain of AI. The Bar-Hillel's report in 1960 (Bar-Hillel) and the now famous ALPAC report in 1966 (ALPAC 1966) described the "semantic barrier" which made high quality MT unattainable. Both discussed in detail the inability of the computer to "understand" natural language and the lack of universal encyclopedias available for consultation in case of lexical ambiguities.

This study of both operational and experimental MT systems will show how AI semantic parsing and analysis are being used to overcome

the "semantic barrier" and how knowledge databases are being developed to answer the machines inquiries about homograph resolution or the problem of words with multiple meanings.

When reviewing MT in relation to AI strategies we need to distinguish several categories of computerized translation programs. These programs can be fully automatic, requiring no human intervention. They can also be interactive, meaning that human intervention takes place before, during or after machine translation has occurred. The degree of interaction between the computer and the human translator varies greatly depending on the system. At the lower level of human intervention are systems called human-aided MT, while at the other end of the scale are computer-aided translation systems. In the latter case the human translator performs the translating act aided by computerized devices such as dictionaries, data banks, word processing, printing, etc. In other cases the machine produces a "raw" translation of a text, that is to say a very rough draft which requires more or less human post-editing. Pre-editing occurs mainly when a relatively short text requires translation into a large number of languages. The degree of human intervention in interactive systems can vary greatly and the line separating human-aided and computer-aided systems remains blurry. If in the late fifties MT researchers were hoping to develop Fully Automatic High Quality Machine Translation (FAHQMT), it is generally admitted today that MT requires human intervention to be brought up to the quality level of work produced by a professional translator. Nonetheless, MT researchers are continuing to investigate conventional and AI strategies to improve the quality of raw translations and diminish the need for human intervention. More realistic than their predecessors, these scientists are not aiming for FAHQMT, but hope to reduce textual ambiguities with more sophisticated syntactic and semantic parsing.

Our study will include two parts and will encompass syntax-based MT systems which have added AI features to an already functional system, and AI-based experimental MT systems. It will also briefly review systems which freely mix so called conventional and AI methods.

Chronologically, the first MT systems to become operational were syntax-based direct translation systems. The best known of them is Systran designed initially as a Russian-English system and later adapted for English-French for the European Economic Communities (1978). Direct translation systems versus indirect translation systems were initially designed for translation in one language pair only. In direct systems, analysis of the source language (SL) is limited to disambiguation necessary for one given target language (TL). Indirect translation systems such as EUROTRA (King 1982), which are designed from the beginning to be multilingual, will be reviewed later. In indirect translation systems, analysis of SL is exhaustive since it is intended for a multiplicity of TL.

Systran which was started as a US military project is now a commercial product used by XEROX, General Motors of Canada, Aerospatiale and many others in a number of language pairs including Spanish, German, Italian, Portuguese, etc. The system has evolved considerably and is now regarded as a direct-transfer system, meaning that programs of structural analysis of SL and synthesis of TL have become more independent from each other. The main characteristic of the system is that the translation process is largely driven by the SL-TL dictionaries. Of interest to this study are AI features which have been incorporated to the dictionaries. These AI strategies consist of a "Limited Semantics Dictionary" and a "Conditional Limited Semantics Dictionary." The main advantage of semantics dictionaries is that they

help the machine resolve problems of words with multiple meanings called "homograph resolution." The term "homograph" is used by MT researchers to include what linguists call homonyms and polysemes. Homonyms are words which have two or more unrelated meanings, such as "bank" : "geological feature" or "financial institution." Polysemes are words such as "raise" which reflects different shades of meaning depending on context in sentences such as "to raise one's hand" "to raise a question" or "to raise money." The seemingly impossible task of "homograph resolution" has been the bane of MT and was the main argument used by Bar-Hillel when he attempted to demonstrate the non feasibility of quality MT. His example which has become famous uses the term "pen" as an example of a homograph and reads as follows:

ex. "The box was in the pen"

which does not make sense even to a translator until it is placed in its context:

"Little John was looking for his toy box. Finally, he found it.
The box was in the pen. John was very happy."

Bar-Hillel used the above example to show that not only does human knowledge store vast amounts of facts but also draws an infinite number of inferences from given facts. His conclusion was that the computer would have to be supplied not only with dictionaries but a universal encyclopedia, which was in his words, "utterly chimerical and hardly deserves any further discussion" (Bar-Hillel 1960). It must be stated that Bar-Hillel was striving for perfection and had in mind a previously mentioned system called Fully Automatic High Quality Machine Translation (FAHQMT). Today, however, most researchers have accepted the necessity of post-editing for the foreseeable future. Nonetheless, semantic dictionaries, without being universal encyclopedias, incorporate increasing amounts of knowledge data which contributes to successful homograph resolution. Not only do these semantics dictionaries handle idiomatic expressions such as "hold one's tongue or make away with", they also include the use of contextual information which is common or essential in a sentence or a text. They include what are called semantic "universals" or "primitive elements" such as "human," "animate," "liquid," etc. but also semantic relations such as "agent-action" or "cause-effect." While earlier direct systems would print a list of two or three possible translations for the term "bank" meaning "geological feature" or "financial institution", present systems making use of contextual information will make a selection using semantic categorization which include subject-fields able to distinguish between the world of finance and geology. Obviously such semantic aids are not foolproof, and in a sentence such as:

"Banks refused to finance the Outer Bank Islands conservation project"

the system would most probably categorize both "bank" terms in the financial subject field unless "Outer Bank" were entered as an idiom. Semantics dictionaries' purpose is not FAHQMT, but a higher percentage of correct homograph resolution. Semantics strategies added to operational direct-transfer systems cannot be said to make the semantic barrier come tumbling down, but it would be justified to state that these AI features are slowly chipping away at the now famous barrier.

Of considerable interest for MT researchers is the EEC Eurotra project meant to replace Systran, whose potential as a multilingual system is limited. The project will regroup the efforts of approximately eighty noted researchers including a large number of nationalities and research centers. Eurotra is probably unique since

it has been designed since its inception (1978) as a multilingual system. It was agreed that it would be based on the most recent techniques and that be a robust operational system. At this point Eurotra is a transfer system whose transfer elements are called "Euroversals" since they are of common historical origin. Of interest to our study is the fact that even though Eurotra planning and design included the possibility of incorporating most recent linguistic and AI techniques, neither knowledge data bases nor inference mechanisms have been included in the system so far. [Inference rules which are conditional and probabilistic (if all elephants have a trunk, and Jumbo has a trunk, then, Jumbo is probably an elephant) allow the system to work out conclusions from contextual data.

The second part of this study includes semantics based systems which at this point remain experimental. As their name indicates these programs apply semantic parsing and semantic analysis approaches as a first step to natural language understanding followed, when necessary, by syntactic analysis. This approach is a reversal of the method reviewed in Systran, for instance, where syntactic parsing came first and was followed by semantic parsing where needed for disambiguation. In general terms, semantic parsing is a means of going beyond the sentence structure. It is a method for creating cognitive or conceptual representations which form knowledge data bases. These cognitive or conceptual frameworks are often called "schema" in AI jargon. They are also called "frames," "templates," "scripts" and "primitives" by various MT researchers. This is not to suggest that the above mentioned terms are synonymous. They are all knowledge structures, but the amount of knowledge they represent can be extensive and detailed. This is the case for Charniak's "frames" about shopping in a supermarket (Charniak 1975) or Shank's "scripts" of what happens in car accidents etc. Or, as is the case for Wilks's "primitives" or "templates," they can be as brief as "Man Have Thing" (Wilks 1973). Semantic roles are expressed, for instance, by "agent" of a transaction or "beneficiary" of a transaction and deal with the equivalences of roles played by Eve (agent) and Adam (beneficiary) in the two following sentences having the same meaning but exhibiting two different structures:

"Eve gave the apple to Adam" and "Adam was given the apple by Eve."

As is pointed out by researchers, semantic based systems do not translate but interpret or paraphrase the text since general meaning is retained but textual structure often is not. An example taken from the Yale experiment by Shank (1975) will illustrate this point. Shank's approach establishes "scripts" which are composed of conceptual representation about what happens in car accidents, ambulances, hospitals, etc. Using such "scripts" the sentence

"Friday evening a car swerved off Route 69. The vehicle struck a tree."

becomes the following in Spanish

"El viernes al anochecer un auto choco contra un arbol."

The paraphrase into Spanish uses the correct verb "chocar" for "hit" in this given context rather than "pegar" or "golpear" but it also "retells" the event instead of translating it. These few examples give an idea of the complexity of semantics based MT and explain why at this point there are no large-scale systems in operation. It should be mentioned in passing that the Japanese are predicting for the 1990s "intelligent computers that will be able to converse with humans in natural language

and understand speech and pictures..." (Feigenbaum and McCorduck 1984).

Most of the recent efforts in MT in the eighties have combined both linguistics-based and AI-based approaches and many systems defy any neat classification. Such are Susy at the University of the Saar, Geta of the University of Grenoble, Metal of the University of Texas and Eurotra of the European Economic Communities. AI strategies including semantic parsers, knowledge databases, expert systems and inference routines have all been incorporated more or less extensively into disambiguation routines.

Almost thirty years ago Bar-Hillel pointed out the unique ability of the human mind to tap its real world knowledge to understand a text to be translated and the translator's ability to make inferences from known facts and situations when faced with a new context. Today MT is a testbed for AI use of knowledge data bases and inference routines in particular. However, given the complexity of real world mirrored in natural language comprehension AI inference routines are far from foolproof and AI-aided MT output will continue to require some post-editing for possibly the next twenty years to attain human translation quality.

BIBLIOGRAPHY

- ALPAC 1966. *Language and Machines: Computers in Translations and Linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences-National Research Council.
- Bar-Hillel, Y. 1960. The present status of automatic translation of languages. *Advances in Computers* 1, 91-163.
- Charniak, E., and McDermott, D. 1984. *Introduction to Artificial Intelligence*. Reading, Mass.: Addison-Wesley.
- Chomsky, N. 1975. *The Logical Structure of Linguistic Theory*. Chicago: University of Chicago Press.
- Connell, Ch. 1986. Machine Translation poised for growth. *High Technology*, June 1986, 53-55.
- Davis, D. 1986. Artificial Intelligence enters the main stream. *High Technology*, June 1986, 16-23
- Feigenbaum, E.A. and McCorduck, P. 1984. *The Fifth Generation: Artificial Intelligence and Japan's Challenge to the World*. London: Joseph.
- Fillmore, C.J. 1968. The case for case. In: Bach, E. & Harms, R.T. ed. *Universals in linguistic* (New York: Holt, Rinehart and Winston), 1-88.
- Hutchins, W.J. 1986. *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood Limited.
- Lawson, V. (ed.) 1982. *Practical Experience of Machine Translation: Proceedings of a Conference, London, 5-6 November 1981*. Amsterdam: North-Holland.
- Lehmann, W.P., and Bennett, W.S. 1985. Human language and computers. *Computers and the Humanities*. Vol. 19, No. 2, April-June 1985. 983. *Artificial Intelligence*. McGraw-Hill.
- Rich, E. 1983. Artificial Intelligence and the Humanities. *Computers and the Humanities*. 19/2, April-June 1985.
- Shank, R. (ed.) 1975. *Conceptual Information Processing*. Amsterdam: North-Holland.
- Shank, R, and Colby, K.M. (eds) 1973. *Computer Models of Thought and Language* San Francisco: Freeman.
- Slocum, J. 1985. Machine Translation. *Computers and the Humanities*. 19/2, April-June 1985. 4. 1975. *Translating and the Computer*. Amsterdam: North-Holland Publishing Company

- Snell, B. M. 1975. *Translating and the Computer*. Amsterdam: North-Holland Publishing Company.
- Wilks, Y. 1973. Analytical intelligence approach to machine translation. In: Schank and Colby (1973), 114-151.
- Winston, P. and Predergast, K. 1984. *The AI Business*. MIT Press.
- Witkam, A.P.M. 1983. *Distributed Language Translation: Feasibility Study of a Multilingual Facility for Videotex Information Networks*. Utrecht: BSO.