Hybrid Approaches in Machine Translation: From Craft to Linguistic Engineering

Claude Coulombe
Machina Sapiens inc.

Summary

To attain the status of a true technology, MT must move beyond mere demonstration of pure theories and clever tinkering to linguistic engineering.

On the one hand, MT-applied linguistic engineering must abandon craft, its toy-systems that will never become true applications due to insufficient theoretical foundations.

On the other hand, linguistic engineering must also avoid approaches that mainly look to reproduce how the human mind operates for the same reasons that airplanes do not fly like birds. Linguistic engineering is characterized by its eclecticism, rejecting the infamous theoretical "schools" and theories that are supported by ideologies.

MT-applied linguistic engineering systematically and incrementally builds a set of tools, methods and know-how based on proven theoretical results.

Linguistic engineering integrates the results of various theories and paradigms into an essentially hybrid approach such as the use of large-coverage symbolic parsers, integrated semantic knowledge that is partially symbolic and partially based on statistical models and, finally, the user's interactive participation.

Introduction

Approaches in MT are very diversified. Some researchers see MT as a means of demonstrating their theories or formalisms, with their measure of success based on whether or not the system is an accurate model of the human mind or simply an elegant theory. The search for a universal grammar for all languages or translation based solely on neural nets to simulate the human mind fall under these theoretical approaches.

Other researchers concentrate only on applying formulas lacking theoretical grounding. The systems they produce are often gadgets that cannot be developed into large-scale systems (Carbonell 1992; Fuchs 1993). Weizenbaum's Eliza system and Terry Winograd's SHRDLU system are good examples of such an essentially empirical approach. In reality, the method used in linguistic engineering basically consists of seeking a convergence of theory and practice.

A great deal of effort is needed to create a functional MT program, and sources of inspiration abound. They include experimentation and theories or paradigms that are symbolic, connectionist, cognitivist, or psychological. Research in MT is still, above all, experimental but guided by solid theoretical foundations.

Its sole performance criterion is to obtain results for a well-defined need. There is no global solution or panacea. For every need, a fitting solution must be found. The goal of linguistic engineering is to respond to well-defined needs of users while balancing resource, cost and time constraint matters.

However, certain myths must be debunked to create a new generation of MT programs.

The myth of the perfect translator

From a theoretical point of view, a text must be thoroughly understood for it to be properly translated. This position assumes that a human translator starts with a perfect understanding of a text and then translates it into another language. Is this really what a translator does? If so, we will not see worthwhile machine translation before the advent of true artificial intelligence. In support of such a position, the same old academic examples are propounded: "John saw the horse with the telescope." Q.E.D.

It is a rude awakening, similar to the aftermath of the Alpac report. Because there is no perfect and inexpensive solution, the problem is impossible to resolve. One can almost hear echoes of the past expounding on the impossibility of men learning to fly or moving faster than100 km/h.

The same can be said about the grammar correction of texts. Even though less sensitive to semantic aspects than translation, the perfect correction of a text cannot be done without a thorough understanding of the text. The authors of this article are very familiar with this fact, because the firm they work for successfully markets Correcteur 101, a system capable of correcting over 80% of errors in a French text. It is always amusing to recall that before Correcteur 101 was created, some 'specialists' warned us that we would fail in our endeavors. It is legitimate to wonder if translation tools will ever be 100% reliable.

In a typical translated text (one taken randomly from a large corpus), there will always be translation errors of an essentially semantic nature, which will be difficult to correct without integrating a great deal of world knowledge. In fact, it is relatively easy to trick translation programs with specially constructed texts full of ambiguities that are practically unresolvable, even for a human being. From a practical point of view, we prefer to imagine the user as being co-operative rather than trying to trick the programs.

It is easy to convince oneself that there will always be interpretation errors, which can never be avoided without a thorough understanding of the context, that is, errors that a human translator, who is not the author of the source text, could make. We can take comfort knowing that no human translator can be 100% reliable, which is what justifies the work of the proofreader.

Could we not be satisfied with translation software that is 80% or 95% reliable? Nothing engineered is without fault, including aircraft, cars, computers, and bridges. Nothing is 100% reliable under all imaginable conditions of use, so why require translation to be perfect? Why is there this double standard?

Meeting well-defined needs

Engineering is driven by a practical work hypothesis, where for each need there is an appropriate solution. For example, it is pointless to swat a fly with a hammer or a steamroller.

There is no panacea or universal solution. For every translation need, there is an adapted MT solution that considers expected results and constraints on resources, cost and time.

First, the type of text to be translated must be considered (technical manuals, business letters, contracts, e-mail, novels, or poems). Second, a distinction must be made between "instant translation" for the purpose of quickly understanding a text in a foreign language, translation-distribution of technical documents and translation-adaptation of literary works (Carbonell 1992). From a practical and purely commercial point of view, we feel that literary translation is beyond the capabilities of MT. Along the same lines, the failure of MT is often proclaimed, but what kind of MT was used and for what purpose?

There is no denying that presently there are MT systems capable of meeting the limited needs of the assisted reading of foreign language texts (also known as comprehension aids or "gist" translation). These systems are particularly useful when the source language is very different from the user's language. For text comprehension, commercially available MT is already sufficient, as well as quick and inexpensive. A user can always decide to entrust the meticulous translation of a text to a professional translator. Therefore, the myth of MT as an all-out failure must be debunked.

Furthermore, powerful MT systems for sublanguages exist, such as the well-known METEO system, which translates hundreds of Canadian weather bulletins every day with 95% accuracy and no human involvement (Chandioux 1988; Macklovitch 1990).

The need to be met involves the translation-distribution of technical documents, business letters, and factual reports whose content must be translated without flourish. The process involves a cycle of translation-revision-distribution. We believe that the general acceptance of MT as a practical tool for translation-revision-distribution would be obtained with an 80% success rate (i.e. 80% of sentences translated are of a suitable quality). Existing systems are currently performing at 60%, which means that only a 20% gain is needed!

Another characteristic of linguistic engineering involves the importance given to the quantification and measurement of results. Measuring results gives a better understanding of the state of the technology and allows development efforts to be better directed in order to maximize the results.

Symbolism versus connectionism…

For several years, the field of artificial intelligence has been divided into two theoretical camps: symbolism and connectionism. They have become two schools of thought, and without disrespect, can even be described as two ivory towers.

Symbolism

The basic position of a symbolic approach to AI is that a physical symbol system is necessary and sufficient to simulate behaviors of intelligence (Newell 1972). It is the primary position of expert systems. Symbols are manipulated using explicit rules established by a linguistic engineer. A symbolic system can only solve a problem if a human being defines the problem with a formalism that facilitates the search for the solution. Furthermore, despite several technical improvements such as fuzzy or weighted logic (to introduce the notion of probability) and meta-level reasoning (to modify the inference strategy), symbolic systems have a tendency to get lost in combinatorial explosions. However, the main drawback with the symbolic approach lies in its dependence on laborious hand coding by expert manpower. The development and maintenance of large knowledge bases represent other challenges.

The symbolic approach seems well suited to simulating the most developed faculties of the human mind, such as problem solving, logical reasoning and planning.

Connectionism

Connectionism refers to neural net computations. A neural net involves parallel processing done by a network of interconnected elementary processing elements. The power of a neural net is above all due to interactions within the network. A neural net can be seen as a rudimentary model of the human mind. The connectionist approach performs well in perception processes, pattern recognition, learning and adaptation to an environment. However, systems based on neural nets are often very slow and are still impenetrable black boxes.

Linguistic engineering, which scorns the idea of making a model based on the workings of the human mind, would opt for implementing the learning functions of neural nets by using more efficient statistical methods. From a technological standpoint, this can be seen as the most efficient alternate implementation.

The power of hybrids

Far from being in opposition to each other, these two approaches are actually complementary. From a practical linguistic engineering point of view, the symbolic approach involves hand-coding knowledge while the connectionist approach relies on a certain form of automated learning from examples or a corpus. There is a tradition of hybrid systems in AI, which, for example, delegates perception to a connectionist layer (a neural net) and reasoning to a symbolic layer (a rule-based system). Such an approach is

often found in voice recognition and advanced robotics. The secret to good engineering practice consists of finding a fine balance between these architectures.

The first steps in properly parsing a source text

The quality of a translation depends heavily on the quality of the parse of the source text. A decrease in noise at input would necessarily create less noise at output (GIGO: Garbage In - Garbage Out). That is why it is important to have a good quality parse of source texts. Again, a basic engineering principle applies. In fact, there is absolutely no point in performing reasoning on the deep semantics of a text that has not been successfully parsed at the lexical then syntactic level. Indeed, many semantic relations are computed from a syntactic parse.

A very sophisticated parse is one of the main advantages of Machina Sapiens' technology (Correcteur 101 for French, CorText for English and El Corrector for Spanish). It consists of broad-coverage morphosyntactic parsers, which are considered among the most powerful to date (Coulombe 1991; Doll 1995). Machina Sapiens' parsers are hybrids on many levels. Their grammar formalism borrows from the linguistic theories of Igor Mel'cuk (Mel'cuk 1988), Maurice Gross (Gross 1975), and Noam Chomsky (Chomsky 1965). Their operation involves symbolic processing and probability models. They represent a convergence of know-how acquired through practice as well as grounded theoretical foundations, which has made these parsers perform as well as they do.

Interlingua's supposed theoretical superiority over transfer

Translation systems rely on three approaches: the direct method, the transfer method, and the interlingua method (Fuchs 1993; Boitet 2000). The direct method entails a morphological analysis of the source text, after which a bilingual dictionary provides a word-for-word translation. The word order is then rearranged and finally the target text is generated. The transfer method, on the other hand, entails a morphosyntactic parse of the source text, then a transfer module establishes equivalencies at the lexical level and transformations at the syntactic level, and then a generation module produces the target text. Finally, the interlingua method entails a morphosyntactic and semantic parse of the source text to produce an abstract representation (the interlingua)--drawing on semantic primitives--that is supposedly language-independent. The generation module then produces a text in the target language from the interlingua representation.

Another somewhat theoretical debate in the design of MT systems centers on opposing transfer-based systems and interlingua-based systems (Carbonell 1992). We believe that this debate is a non-issue. Again, the needs that we aim to fulfill must be well defined. In the case of a single-source and multi-target system, a transfer architecture proves to be particularly effective. On the other hand, for a multi-source and single-target system, it is preferable to use a direct translation architecture. Lastly, for a multi-source and multi-target system it is more cost-effective to use an interlingua architecture (Boitet 2000).

Many MT systems try to define a type of universal language, known as an interlingua, which serves as an intermediary in the translation process. This is appealing on a theoretical level, as the interlingua is often an attempt to avoid the work involved with multiple transfer modules. Theoretically, N*(N-1) transfer modules are needed for N languages. The transfer approach thus leads to exponential costs, at least in theory. This theoretical reasoning is alarmingly naive when the practical is considered. As demonstrated by Christian Boitet of GETA, it suffices to use the parser results of the most frequently translated language as an interlingua to reduce the number of transfer modules (Boitet 2000). However, the design of an interlingua translation system that goes beyond the stage of modeling small, closed domains is a particularly difficult challenge (Boitet 2000). It involves in-depth research at the level of semantic representation that can quickly turn into the creation of a world model. Unfortunately, to be truly honest, no theory can as of yet guarantee that any utterance in a natural language will have an interlingua representation that can generate the same utterance in another language. Consequently, from a linguistic engineering standpoint, the interlingua system cannot be relied on to create a serious application.

In principle, transfer systems are less ambitious (and more realistic) than interlingua systems (Fuchs 1993). Transfer-based systems will prove to be superior in processing similar languages. Transfer is facilitated for certain families of human languages that display the same syntactic characteristics. The closer two languages are syntactically, the easier it will be to implement a transfer-based MT system. Furthermore, resolving all ambiguities is not always necessary to process similar languages (Fuchs 1993). However, it is essential for correct translation when more dissimilar languages groups are involved.

We believe that the 80% quality objective we have set for ourselves will only be economically viable for an MT system that integrates a rather restricted number of related languages. Indeed, the translation process depends on a specific expertise representing a great amount of knowledge for each language pair, for each translation domain and for each translation direction. Proof of this lies in the fact that a human translator rarely translates successfully towards more than one language (basically their mother tongue) and from more than one or two languages. Furthermore, a quality translation will try, as much as possible, to maintain the structure of the source text, especially when the languages involved are similar. The use of an interlingua eliminates this structure and merely retains the meaning of the source text. On this level, the MT system does not translate; it simply generates paraphrases.

With a view to translating related languages, such as French and Spanish, we propose a transfer architecture enriched with semantic information. The semantic enhancement makes the proposed system an architecture that is once again a hybrid of a transfer and an interlingua system. While it may be true that an interlingua system cannot omit semantics, this does not rule out that a transfer-based system may also resort to semantics. This resorting to semantics is what characterizes a third-generation system. In the hybrid method of semantics-enhanced transfer, there is a morphosyntactic and semantic parse of the source text. Then, a transfer module determines equivalencies at the lexical level, makes certain inferences and disambiguations at the semantic level and performs transformations at the syntactic structure level. Finally, a generation module produces a target text.

In such a hybrid system, the distinction between an interlingua translation system and a transfer-based system is blurred. Indeed, semantics are no longer the sole prerogative of the interlingua approach, which can no longer presume to possess conceptual superiority in terms of the thoroughness of its parse. This approach is clearly situated in the wave of recent transfer systems that use semantic information for parsing and transfer. Moreover, Laurence Jacqmin has already stressed this tendency towards "crossbreeding" (Jacqmin 1993).

Semantics "à la carte"

The problem with current commercially available translation software is that these programs only perform a very cursory parse of source texts. Their analyses do not resolve semantic ambiguities (polysemy) or syntactic ambiguities (points of attachment).

A purist would argue that a quality translation relies on a world model. To attain absolute reliability, translation software would have to be equipped with a total comprehension of "common sense." However, the creation of an MT system equipped with common sense would require encoding an astronomical amount of encyclopedic knowledge, which is necessary to build a general world model. In fact, an evaluation based on a simple extrapolation of the work already done by Machina Sapiens suggests that an effort in the order of 500 to 1,000 person-years is needed to create a generalized bilingual MT system capable of translating any kind of text, with a hand-coded world model. A specialized bilingual system (in the order of 10,000 concepts) would require a minimum of 100 person-years. This is no mean feat! In fact, it is a daunting task that appears to be clearly insurmountable by a small team. It is simply unthinkable to enter all of this information by hand.

Fortunately, resolving all ambiguities is not always necessary when related languages are processed. A good quality translation is possible without a thorough semantic parse aimed at disambiguating sentences. Indeed, we can rely on the similarity between languages to try to preserve all of the source language's ambiguities in the translation. The ideal solution to the problem of having too much information to hand-code is to add only the semantic information necessary to resolve semantic ambiguities (polysemy) or syntactic ambiguities (points of attachment). We propose semantic enhancement where semantic information is added on the basis of necessity alone, with a view to solving problems of ambiguity rather than constructing a world model, as interesting as such a project may seem in other respects. An unsupervised learning approach using large corpora may provide the basics of commonsense knowledge at reasonable costs. Statistical processing of collocations could be employed to resolve anaphora and points of attachment. Systems enhanced with semantics can also be constructed for a specific or micro-domain application.

Engineering functions according to Pareto's principle of economic optimality, which postulates that 20% of the effort produces 80% of the results. If we apply Pareto's principle, we arrive at an estimate in the order of 100 to 200 person-years.

And what is wrong with interactive translation?

To obtain an acceptable quality translation, why can we not rely on the interactive involvement of the user? Let us reconsider the task of correcting texts written in French. Despite the fact that no program can guarantee 100% reliability, a cooperative user with a basic knowledge of grammar and spelling can, with the help of Correcteur 101, produce faultless texts, or at least texts that are considerably better linguistically. When we look at the bigger picture, moving from the computer only to a coupling of the computer and user, it becomes evident how user interaction can become the main means of resolving natural language ambiguities (Wehrli 1992). Because no program will be able to integrate enough world knowledge and common sense to automatically resolve all the ambiguities in any source text for many years to come, the idea then is to place the user at the end of the process. The user makes final decisions and resolves persisting ambiguities. This idea comes from the simple observation that the user possesses world knowledge. In any case, it is preferable that users intervene in the revision and final approval of texts, especially for important or complex texts.

Pre- and post-processing tools for translators can also be integrated with the translation process, including a standardization of terminology, simplification with controlled language tools and grammar correction (Boitet 2000).

The delay in the field of interactive translation may have something to do with the fact that the main translation programs date from the 1960s and 1970s. These programs have only recently been released for graphical platforms and they offer few possibilities for user interaction. There should be a major focus put on designing new user interfaces.

We can even conceive of programs for monolingual writers that can generate high-quality texts in a target language needing no revision from the source language text. To ensure the quality of the text produced, the program would use “retranslation.” In other words, the system translating from the source language to the target language would retranslate the translated text to the source language. This would enable the author to check the resulting text without having to know a single word of the target language (Boitet 1993). In effect, it involves applying a test frequently used to evaluate translation programs.

Conclusion

The future of MT is bright if we remain realistic – and modest. For example, well thought out MT tools could become a major asset in increasing the productivity of translators.

To obtain a translation of suitable quality, sterile theoretical frameworks that proclaim the impossibility of generating a perfect translation must be discarded. Hybrid and innovative approaches must be relied on. This includes using large-coverage symbolic parsers, semantic enhancement on an “as-needed” basis, as well as using statistical models and the user’s interactive participation.

Lastly, let us not forget what we learned in ecology lessons – hybrids are able to survive in hostile environments.

References

Boitet, C. (1993). "La TAO comme technologie scientifique : le cas de la traduction automatique fondée sur le dialogue" [chapter 3.5]. *La Traductique* [under the direction of Pierrette Bouillon and André Clas]. Montréal: Presses de l'université de Montréal, AUPELF/UREF.

Boitet, C. (2000). "Traduction assistée par ordinateur" [chapter12]. *Ingénierie des Langues* [under the direction of Jean-Marie Pierrel]. Paris : Hermes Science Europe.

Carbonell, J. G., Mitamura, T., H. Nyberg, E., 3rd (1992). "The KANT perspective: a critique of pure transfer (and pure interlingua, pure statistics. . .)". *Proceedings of TMI-92*. Center for Machine Translation, Carnegie Mellon University, Pittsburgh.

Chandioux, J. (1988). "Dix ans de METEO, dans Traduction assistée par ordinateur". *Actes du Séminaire International sur la TAO* [under the direction of A. Abbou]. OFIL (Observatoire français des industries de la langue), Paris.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Coulombe, C. (1991). "Les qualités attendues d'un correcteur orthographique et syntaxique". *Actes du Colloque* "*Problématiques 1995*", *Traitement Automatique de la Langue et Industries de l'Information*. OFIL (Observatoire français des industries de la langue), Paris.

Doll, F. (1995). "Du correcteur othographique au correcteur grammatical intelligent". *États Généraux de la Francophonie Scientifique*. Montréal: AUPELF/UREF.

Fuchs, C., in collaboration with Lacheret-Dujour, A., Victorri, B., Danlos, L, and Luzzati, D. (1993). *Linguistique et Traitements Automatiques des Langues*. Paris: Hachette Université, Collection Langue Linguistique Communication.

Gross, M. (1975). *Méthodes en Syntaxe*. Paris: Hermann Editeur.

Jacqmin, L. (1993). "Classification générale des systèmes de traduction automatique" [Chapter 3.1]. *La Traductique* [under the direction of Pierrette Bouillon and André Clas]. Montréal: Presses de l'université de Montréal, AUPELF/UREF.

Macklovitch, E., and Isabelle, P. (1990). "Les voies actuelles de la traduction automatique au Canada". *Tribune des Industries de la Langue*. OFIL (Observatoire français des industries de la langue), Paris.

Mel'cuk, I. (1988). *Dependency Syntax: Theory and Practice*. SUNY, New York.

Newell, A., and Simon, H. (1972). *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice-Hall.

Wehrli, É. (1992). "The IPS System". *Proceedings de la Conférence COLING-92*. Nantes.