A Direction of MT Development[1]

Sergei Nirenburg
Center for Machine Translation
Carnegie Mellon University

I would like to discuss just one among the many potentially fruitful directions of development of our field — the introduction of multi-engine system architectures which would facilitate the choice of the best way to translate a source text passage given a set of system resources.

1.   Toward Extended MT Architectures

Recent years have witnessed a shift in the balance of scientific and technological efforts in the area of machine translation. All the latest methodological novelties in this field are essentially technology-oriented and do not aim at advancing our knowledge about either basic mechanisms of text comprehension and production or computer models simulating such mechanisms.

The two most recently popular technological paradigms in machine translation — example-based translation (EBMT) and statistics-based translation (SBMT) — require knowledge about language only as an afterthought. While the representatives of the above paradigms are still at the stage of either building toy systems (e.g., Furuse and Iida, 1992; McLean, 1992, Jones, 1992, Maruyama and Watanabe, 1992) or struggling with the natural constraints of approaches that eschew the study of language as such (e.g., Brown et al., 1990), a number of proposals have come up for some hybridization of MT. In some such approaches, corpus analysis is used for tuning analysis and transfer grammars (e.g., Su and Chang, 1992). In others, a standard transfer-based approach (TBMT) is followed using traditional analysis and generation techniques but having transfer component based on aligned bilingual corpora (Grishman and Kosaka, 1992). Still in others, it was suggested that statistical information be used as the source of preference assignment during text disambiguation (e.g., the outline presented in Lehmann and Ott, 1992). Indeed, hybrid MT systems were a central topic of the latest Conference on Theoretical and Methodological Issues in MT.

It is important to recognize, though, that most of these hybridization proposals were driven essentially by technological concerns. Of course, machine translation *is* an applied field and it is quite appropriate that the impetus for progress comes in a large part from extra-scientific sources. However, machine translation is a special application. Unlike most other areas, it is a very good

---

and comprehensive testbed for linguistic theories (of syntax, semantics, pragmatics and discourse), computational-linguistic methods (algorithms for parsing, semantic and pragmatic interpretation and text generation), descriptive linguistics (lexicons and grammars for particular languages), modeling human reasoning processes (knowledge representation and manipulation) and translation studies. In fact, we believe that it is an ideal environment for developing and testing treatments of what we have referred to as "microtheories" (Nirenburg et al., 1992a), that is, treatments of a wide range of specific language phenomena such as semantic head-modifier dependencies, aspect or quantification, as well as larger, more comprehensive computational theories of language use.

Current MT projects — both "pure" and hybrid, both predominantly technology-oriented and scientific are single-engine projects, capable of one particular type of source text analysis, one particular method of finding target language correspondences for source language elements and one prescribed method of generating the target language text. While such projects can be quite useful, we believe that it is time to make the next step in the design of machine translation systems and to move toward adaptive, multiple-engine systems.

In addition, current systems have been developed for a particular text type (e.g., weather report, financial news articles, scientific abstracts) and for a particular end use — e.g., assimilation or dissemination of information. Given a specification of the input text type and end use, one or another of the systems might be most appropriate. So, for example, for translating scientific abstracts in a particular domain en masse in order to inform a specific audience of the content of recent papers in the field, example-based translation might be preferable. In order to process short articles on a wide range of topics in order to filter out just those that appear to be of particular interest, a statistics-based approach might be most appropriate. This correspondence between technique, input text type and end use (or output text type) provides yet further motivation for having an adaptive, multiple-engine system.

Within the so-called "rationalist" camp of machine translation workers, a central point of scientific debate has for a long time been the question of the level of complexity of source text analysis. The battle lines were drawn as follows. One group of researchers maintains that unless a large number of linguistic phenomena widely occurring in natural texts are analyzed and overtly represented, high quality translation is unattainable. The adherents of this meaning-oriented approach acknowledge that such phenomena can be described only with the help of abundant knowledge stored in detailed (and usually hand-crafted) grammars and lexicons of each language involved.

The other group of researchers argues that this acquisition task is not realistic. Taking heart in numerous observations that deep analysis is not always needed for translation (for instance, that the polysemous Spanish noun *centro* is translated into German as *zentrum* no matter which of the senses of *centro* was used in the source language text), they opt for a simpler analysis and use of more direct source-target language substitutions in place of involved meaning analysis.

A typical formulation of this position is given by Ben Ari et al. (1988, 2): "It must be kept in mind that the translation process does not necessarily require full understanding of the text. Many ambiguities may be preserved during translation ..., and thus should not be presented to the user (human translator) for resolution."

Similarly, Isabelle and Bourbeau (1985, 21) contend that, "Sometimes, it is possible to ignore certain ambiguities, in the hope that the same ambiguities will carry over in translation. This is particularly true in systems like TAUM-AVIATION that deal with only one pair of closely related languages within a severely restricted subdomain. The difficult problem of prepositional phrase attachment, for example, is frequently bypassed in this way. Generally speaking, however, analysis is aimed at producing an unambiguous intermediate representation."

A more fruitful approach to adaptivity would have as a goal the capability of performing "only" as much work in translation as is absolutely required, optimizing the overall system by the types of work which each of the modules does best. Thus, if an input passage is identical (or very similar) to a previously translated passage, stored in a database of past translations, the use of the EBMT engine will be indicated, as it may involve essentially only the retrieval of the stored translation. Putting in motion the entire mechanism of, say, KBMT, in this case will be a waste of resources.

An adaptive system, which would allow direct source — target substitutions when feasible but would activate engines with progressively more complex analysis components when necessary, is a preferred solution to the double jeopardy of either spending resources and processing while running the risk of failure due to lack of decision knowledge in a high-level system or, alternatively, of building a system which is *a priori* unable to treat an important area of natural language meaning.

2.   The Dispatcher

Input passages in a multi-engine system will be assigned to particular MT engines by a "dispatcher" module. The function of the dispatcher is to break up the input text into segments and assign each segment to one or another of the MT engines: (e.g., SBMT, EBMT, TBMT, and KBMT). It takes text (and certain general specifications) as input, executes a set of diagnostics for establishing the translation units and selecting the best engine and, finally, distributes the appropriate text segments to the "best" engine. The dispatcher will evaluate each passage according to the following general criteria:

- Type of translation — whether the result of translation is intended for dissemination (that is, to be read by people outside an organization) or for assimilation (that is, to be read by those internal to an organization); whether a complete translation is needed or an abstract or even a simple categorization of a text (e.g., as a text that is important enough to be translated in its entirety). We would hope that the best quality translation is produced by KBMT and TBMT systems, unless the input matches a text from a bilingual aligned corpus very closely, but this would be an empirical matter. Translations for assimilation can be less "polished" and therefore some errors and omissions can be tolerated, which may tip the decision balance toward the lower-sophistication engines, such as EBMT and SBMT.

- Availability of parallel text in a particular domain and on a particular topic. This is the crucial enabling condition for EBMT and SBMT. The quality of SBMT output can only obliquely be determined before actually performing a full run on a passage (Peter Brown, personal communication). The SBMT module is the most experimental of all the engines in our adaptive architecture and, as with all the modules we postulate, the dispatcher decisions

with respect to SBMT will have to be derived empirically in the evaluation and testing phase of our system.

- Amount of ambiguity in the source passage, both in the source language itself and vis-a-vis a target language (including the vagueness phenomena, computed as a function of the sense range of the component words with respect to the lexicon). In the former case the general rule is that the greater the amount of ambiguity, the more indicated is the KBMT approach. However, if the target language realizes the ambiguous source meanings identically (recall the *zentrum — centro* example), the transfer approach offers a simpler solution.   In the case of vagueness, belief-pragmatic solutions would require the use of KBMT, whereas if simpler resolution criteria are found in particular cases of vagueness (such as, for instance, collocation techniques), transfer can be used once again.

- Size and quality of available KBMT resources (ontology, lexicons, etc.). If it is determined that, for a given input passage, the coverage of the available static knowledge sources is seriously incomplete, the quality of expected output will plunge below a threshold. The dispatcher can judge that the amount of augmentor interaction which would be required to salvage the situation may be excessive and favor the example-based engine instead, even if its results would also be of lower quality.

An additional important parameter in the operation of the dispatcher is determining the most appropriate size of input passage. Since an entire input text can be processed by a combination of MT engines, it is necessary to maximize the expected quality of output over a variety of possible ways of "chunking" the input text for processing. Thus, the dispatcher will be entrusted with finding the optimum length of each passage to be assigned to a particular engine.

References

Ben Ari, D. M. Rimon and D. Berry. 1988. Translational Ambiguity Rephrased. Proceedings of the Second International Conference on Theoretical and methodological Issues in Machine Translation. Pittsburgh, June 1988.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, E, Mercer, R.L. and Roossin P.S. 1990. A statistical approach to language translation, *Computational Linguistics,* vol 16, 79-85.

Furuse, O and H. Iida. 1992. An Example-Based Method for Transfer-Driven Machine Translation. Proceedings of TMI-92. Montreal. 139-50.

Grishman, R. and M. Kosaka. 1992. Combining rationalist and empiricist approaches to machine translation. Proceedings of TMI-92. Montreal. 263-74

Isabelle, P. and L. Bourbeau. 1985. TAUM-AVIATION: Its Technical Features and Some Experimental Results. *Computational Linguistics,* 11: 18-27.

Jones, D. 1992. Non-hybrid example-based machine translation architectures. Proceedings of TMI-92. Montreal. 163-71.

Lehmann, H. and N. Ott. 1992. Translation relations and the combination of analytical and statistical methods in machine translation. Proceedings of TMI-92. Montreal. 237-48.

Maruyama, H. and H. Watanabe. 1992. Tree cover search algorithm for example-based translation. Proceedings of TMI-92. Montreal. 173-84.

McLean, I. 1992. Example-based machine translation using connectionist matching. Proceedings of TMI-92. Montreal. 35-43.

Nirenburg, S., J. Carbonell, M. Tomita and K. Goodman. 1992. **Machine Translation: A Knowledge-Based Approach.** San Mateo, CA: Morgan Kaufmann.

Su, K-Y. and J-S. Chang. 1992. Why corpus-based statistics-oriented machine translation. Proceedings of TMI-92. Montreal. 249-62.

Thompson, H. 1992. Presentation at the AMTA Meeting, San Diego, CA. November.

Wilks, Y., Fass, D., Guo, C-M., McDonald, J., Plate, T., & Slator, B. 1990. Providing Machine Tractable Dictionary Tools. *Machine Translation, 5:2,* 99-151.