SESSION 3: MACHINE TRANSLATION

Jaime Carbonell

Carnegie Mellon University
Center for Machine Translation
Pittsburgh, PA 15213

Machine Translation (MT) technology has progressed significantly since the days of the ALPAC* report. In particular, multiple paradigms are being investigated ranging from statistical methods to full knowledge-based interlingual MT systems. Much of the recent work is based on advances in natural language processing since ALPAC in the 1960s, including:

- **Semantic analysis** to resolve lexical and syntactic ambiguities during parsing, and thus reduce translation errors very significantly.
- **Unification grammars** allowing syntactic and semantic constraints to be checked in a unified manner while parsing, and permitting reversible grammars—i.e., the same grammars to be used for generation as well as for analysis.
- **Advanced parsing methodologies**, including augmented-LR compilation where knowledge sources (syntactic grammars, lexicons, and semantic ontologies) can be defined and maintained separately but are jointly compiled to apply simultaneously at run time, both in parsing and in generation.
- **Natural language generation**, focusing on how to structure fluent target-language output, an activity not truly investigated in the pre-ALPAC days.
- **Automated corpus analysis tools**, statistical and other means of extracting useful information from large bi- or multi-lingual corpora, including collocations, transfers, and contextual cues for disambiguation.
- **MRDs => MTDs**, use of electronic machine-readable dictionaries (MRDs) to partially automate the creation of machine-tractable dictionaries (MTDs) in processable internal form for parsers and generators, permitting principled scaling up in MT configurations.

APPROACHES TO MODERN MT

In light of these advances, several major MT paradigms have evolved to supplant the early hand-coded direct-transfer methods. One approach is purely statistical, as practiced at IBM, in which the direct-transfer paradigm is still king and translation is viewed as transduction between two character (or word) streams—essentially two encodings of the same message. However, the direct transfer rules are totally learned by statistical analysis of large bi-lingual corpora, rather than laboriously and incompletely hand-coded. A drawback of the statistical approach, of course, is that it cannot guarantee the accuracy of any textual passage being translated, but rather strives to minimize the total number of errors over time.

Another MT approach is to provide a measure of analysis of the source language prior to transfer. At minimum, morphological and syntactic analysis is performed, then the transfer component transforms the parse trees into corresponding parse trees in the target language with appropriate lexical substitution. These transformed parse trees are then used to generate the desired target texts. Performing analysis and generation reduces the size of the transfer component, which is a major benefit, considering that translation across N languages requires $O(N^2)$ transfer grammars. Transfer at the syntactic level represents the classical approach on which most commercial attempts at MT are based. The problem with classical transfer is that it too makes a significant number of errors in the translation, primarily through its inability to reduce much of the lexical and syntactic ambiguity of the source language texts.

A deeper analysis, including semantic restrictions to produce case frames (rather than parse trees), reduces both the number of errors (as some ambiguity has been resolved) and the size of the transfer component (case frame representations in different languages will be much more similar to each other than syntactic parse trees). Some Japanese firms working in machine translation, for instance, have adopted this approach, which they call *semantic transfer*. Since Japanese and English are more different than two Indo-European languages, there is more justification for the deeper level of analysis and the desire to minimize the size of the transfer component.

The deepest level of analysis produces an underlying non-ambiguous semantic representation, independent of both source and target language. This is called the ***interlingua*** or ***pivot*** approach, and it trades off much more work at analysis and generation for no work at all in the transfer phase. Benefits include much lower errors (as ambiguities must be resolved to produce interlingua), and no N^2 problem, as there is no transfer component. Thus the interlingua approach is particularly well-suited for multi-lingual translation. The most serious problem with the interlingual approach is its requirements for vast knowledge bases if one desires general-purpose, highly accurate translation in any domain. Specialized-domain interlingual systems are far more practical.

[*] In 1965 the United States Academy of Science commissoned a study of he state of the art in Machine Translation, whose findings were published the following year and become popularly known as the ALPAC report. In essence, ALPAC argued that there was insufficient scientific basis in natural langauge processing to perform reliable machine translation, and the large expensive computers of the time would make MT economically infeasible. Both situations have since changed drastically, invalidating the ALPAC conclusions. In fact, DARPA has played a major role in fostering the development of the NLP scientific infrastructure in the post-ALPAC years.

EVALUATION METHODS

Yorick Wilks, one of America's foremost MT researchers, states that "Machine translation evaluations methods are better developed than machine translation itself." In a sense, he is right. Since MT is a complete throughput process from source to target text that corresponds precisely to the task of human translations, it is not difficult to compare the two and provide relative assessments. MT has been evaluated with respect to *semantic accuracy* of the translation and *intelligibility* of the final output. But, other factors such as *degree of automation* are equally relevant. The more human intervention (e.g., pre and post editing) required to produce a good translation, the less useful the MT system.

MT evaluations must always be made task-relative. On the one hand, MT for the sole purpose of scanning translated texts in order to establish their relevance to a given topic must be fast, with little if any human assistance, but can be rough, partially inaccurate and of low-legibility output. On the other hand, technical or legal texts translated for publication must be accurate and legible, although slower processing and additional human assistance may be tolerated. Therefore, one challenge for the DARPA MT research community is to develop more appropriate, task-sensitive and comprehensive evaluation criteria.

PREVIEW OF MT PAPERS

The three papers on machine translation in this section cover only part of the DARPA MT effort—and this should be interpreted as a sign of the breadth of coverage and DARPA interest in the field, pursuing different technologies at different sites. In particular, the knowledge-based interlingua approaches of CMU, CRL and ISI are not represented, but form a major component of the overall program. Those areas represented in this volume include translation as abductive reasoning at SRI and two papers on the role of statistics in machine translation. More precisely, statistical word-sense disambiguation at IBM and establishing lexical-transfer correspondences for MT at AT&T are discussed in some detail.

Jerry Hobbs and Megumi Kameyama at SRI extend their existing TACITUS architecture for abductive natural language interpretation and apply it to the task of translating a few examples into Japanese. The work is exciting in the sense of showing how an existing system and underlying theory are sufficiently general to address the new task: machine translation across radically different languages. Other issues, such as scaling up, are not yet addressed in this work.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra and Robert Mercer at IBM developed a statistically-based word-sense disambiguation method as an integral component of a statistical machine translation system. In fact, statistical help in word disambiguation may prove to be of major help in more traditional MT approaches (transfer and interlingua) when definitive semantic and syntactic knowledge do not narrow down word senses to a single candidate. Therefore this work should be followed closely by those researchers of the non-statistical-MT persuasion as well.

William Gale and Ken Church develop a mathematical infrastructure for determining lexical correspondences across words in parallel texts. Parallel text means that the same text is available in two (or more) languages. Establishing lexical correspondences is crucial for building knowledge bases for symbolic translation methods (whether transfer or interlingual) and for automated training of statistical translation methods such as that advocated at IBM by Brown, Mercer, *et al.* Gale and Church argue that their methods are statistically more reliable than the earlier IBM methods for establishing word correspondences.