Towards Speech Translation Systems

Masaru Tomita
Center for Machine Translation
and
Computer Science Department
Carnegie-Mellon University
Pittsburgh, PA 15213

Position Paper for the panel: "Real-time Interpretive MT"

Development of a machine translation system that can take spoken utterances as input via a microphone is certainly one of the most challenging tasks for the coming decades. Such systems, which we shall call speech translation systems in this paper, have different applications and performance requirements from conventional machine translation (MT) systems, and we need a totally different design philosophy to approach speech translation systems.

First of all, unlike conventional MT systems, translation produced by a speech translation system is always final. A conventional MT system usually requires human *post-editors* to fix mistakes made by the system. It is considered acceptable for a MT system to have humans post-edit its output, as long as the cost of the post-editing is significantly less than the cost of manual translation processed by human translators without the MT system. On the other hand, it is clearly unreasonable for a speech translation system to expect to have a human post-editor to fix the system's mistranslation, as such post-editing seems to cost no less than what a human interpreter would cost to interpret a speech conversation. This means that the output of a speech translation system must be highly accurate (if not fluent) so that two different language speakers can at least communicate without intervention of a third person. Accurate translation requires a degree of comprehension, and the *knowledge-based* approach seems to be the unanimous choice for development of a speech translation system. The knowledge-based approach is particularly suitable for translation in a limited task domain, and speech translation systems can find much more limited task applications (such as translating conversations about airline reservations at airport counters, doctor-patient conversations for medical diagnosis at hospital emergency rooms, conversations with international telephone operators, etc) than can conventional off-line translation systems.

Secondly, the kind of sentences speech translation systems must handle is very different from that of MT systems for translating off-line texts. Conversational utterances include elliptical and anaphoric expressions much more often than do off-line documents. They also tend to include many more syntactically ill-formed expressions than typed texts do. Resolving these utterances requires serious contextual analysis and substantial knowledge of the subject domain. This fact supports our choice of the knowledge-based approach to speech translation.

Thirdly, a speech translation system must respond in real time. For conventional MT systems, most of which are intended to translate documents such as technical papers and manuals, it is quite acceptable for the system to spend an entire weekend producing a translation as a batch job. A speech translation system, however, cannot afford to spend any more than a few seconds (perhaps 20 seconds at longest) to translate a sentence, as the users are right there waiting for the system's response. What this means is that speech translation systems must be very fast, and thus efficient algorithms are crucial.

Finally, besides the extragrammatical nature of spoken utterances previously mentioned, a speech translation system has to cope with recognition errors made by its speech recognition device. Speech translation is not as easy as simply connecting a speech recognition system and a machine translation system; serious study is needed to develop the integration of these two components. Until recently, speech recognition and natural language understanding (including knowledge-based machine translation) have been regarded as completely different areas, and two different groups of people have been working on each problem rather independently. Researchers in the speech recognition area have been trying to recognize a spoken sequence of words (i.e., a sentence) as accurately as possible. Unfortunately, speech recognition systems will never be perfect, no matter how much improvement will be made. On the other hand, researchers in the natural language understanding area, even those who work on spoken dialogues, usually begin with a perfectly recognized sequence of words. Tolerating the inevitable recognition errors thus presents a big problem. Parsing noisy sentences involves much a larger search space; a substantial amount of semantic knowledge, as well as an efficient parsing algorithm, is required to constrain the search. This fact also supports the choice of the knowledge-based approach, and the need for efficient parsing algorithms.

Given the considerations outlined above, it is logical to conclude that the knowledge-based approach, paired with efficient algorithms, is the key building successful speech translation systems. There appears to be a conflict between implementing a knowledge-based system and implementing an efficient system; a knowledge-based system must access, manipulate, and maintain a large number of frames (or objects) with complex relations among them, and is usually inefficient in terms of speed. On the other hand, an efficient parsing system usually requires much simpler data structures, such as context-free grammars. Representing a large body of knowledge within such a simple framework, even if it is ever possible, would likely result in an unmodularized knowledge base which is difficult to develop, debug and maintain. We believe that it is possible to overcome this conflict, using an approach which transforms modular, human-readable knowledge sources into efficient, machine-readable parsing grammars.

Our solution to this problem is the Universal Parser Architecture. It enables us to develop linguistic and domain semantic knowledge bases in a highly modularized and perspicuous manner, which the Universal Parser Compiler then compiles into a lower-level representation which is machine efficient but not necessarily human-readable. The modularity of the knowledge sources in the Universal Parser Architecture also make the system transportable across languages and across task domains. We use the *Pseudo Unification Grammar formalism,* which is in a similar notation to that of the Unification Grammar formalism, for linguistic (syntactic) knowledge representation. For domain semantic knowledge representation, we adopt *FrameKit* which is a frame representation system developed at Carnegie Mellon University. Along with a third knowledge base known as the lexicon/concept mapping rules, the Universal Parser Compiler compiles the semantic and syntactic knowledge sources and produces an Augmented Context-Free Grammar (ACFG) for runtime parsing. An ACFG is essentially a list of context-free phrase structure rules, each of which is paired with an attached LISP program (augmentation)[1]. These LISP programs are further compiled into machine code with the standard LISP compiler. The phrase structure part of the ACFG is also compiled further into a Generalized LR parsing table for the Generalized LR parsing algorithm, which is highly efficient.

We have extended our Generalized LR parser to allow for the errors generated by speech recognition devices, thus producing an integrated system which accepts the output of such a system as its input. Our results to date have been very promising in terms of both accuracy and speed. We find our architecture very suitable for speech translation, as it combines the use of domain semantic knowledge with a highly efficient runtime parsing algorithm, thus accommodating the increased search space necessary for parsing speech input.