Speech Technology

Issue Date: November/December 2004, Posted On: 3/16/2006

Voice Ideas - Automating the Tower of Babel

When we think about how speech recognition is used we generally envision it operating in a human-to-machine interaction. That's not surprising because that's the context in which speech recognition operates today. This human-to-machine paradigm is not, however, how speech technology will be used in the future and one of the most exciting areas of automated human-to-human communication is speech-to-speech machine translation (MT).

Text-to-text MT has been a focal area of artificial intelligence work since the 1950s and has received a great deal of funding from military and intelligence agencies. The fact that, after more than 50 years, text-to-text MT is still the object of active laboratory research indicates how difficult it is. The challenge increases exponentially when the input, output, or both are spoken.

Phrase Translators

One area of strong interest is in the development of hand-held phrasal translators. A few companies, such as Ectaco, offer phrase translators for tourists and businesspeople but most of the development has been done for the military. Hand-held devices, such as Dragon's handheld translator and Marine Acoustics' Phraselator, have accompanied US troops to Bosnia, Afghanistan, and Iraq. They help military and support personnel communicate about injuries and other medical conditions, do debriefings, interact with displaced persons, and perform numerous other activities.

As their name suggests, phrase translators are programmed with a set of standard phrases in a specific domain, such as emergency medical services. "One-way" phrase translators output speech but require individuals to select a pre-programmed text phrase in their language (the "source" language). "Two way" systems accept spoken input (using commercial ASR technology) of comparable pre-defined phrases as well as generating spoken output.

Dialogue Systems

Projects, such as DARPA's CAST (Compact Aids for Speech Translation), hope to push the technology beyond phrase translation, primarily for dialogues. Dialogue MT systems also demand far greater flexibility than can be provided by phrase translators, including the ability to "go out of domain." For example, it would not be unusual for a victim involved in an automobile accident to need to communicate about the circumstances surrounding the accident and about his injuries.

I spoke with Dr. Yuqing Gao, about her research on MT for dialogue interactions. Dr. Gao heads IBM's two-way research for the US Government and that work is also part of IBM's Super Human computing effort.

Gao's approach to MT deviates from established practices. "Normally, speech translation is treated as two components: first you do speech-to-text ASR and then you do MT on the text. The problem is that each of those components is far from perfect. So, we believe that just putting the two together won't produce a satisfactory result. That's why we treat the entire process as a single task."

The approach is a blend of the acoustic patterns in the stream of speech, N-gram language modeling, and various aspects of meaning. "When we started work in the Super Human project we focused on the acoustic component. Now, we try to decipher not only the

acoustic and phonetic information but also the semantic information and the intention behind the speech - the goal of the communication." The clues that reveal the intention come from a variety of sources, including the history of the dialogue and various levels of semantics. "We don't just look at the meanings of the words - lexical meaning. We do multiple levels of syntactic and semantic parsing that include analysis of the whole sentence and the dialogue structure."

Sentences tell the system what the person is trying to accomplish at each point in the dialogue: are they trying to tell the other person something, ask for help, give instruction, or something else. The dialogue provides the context within which the sentence is understood which is why the system keeps a record of what has been said in the dialogue. "This is critical because if you don't have the history you don't always know how to translate a sentence. The same sentence can be translated different ways depending on how it fits into a dialogue." Prosody and emotion are also considerations that affect meaning although, as Gao admits, they are extremely difficult to interpret and "sometimes people can be so skillful that what they say doesn't change the prosody."

All these elements are embedded in a cultural context that is critical for determining an appropriate reconstruction in the target language. "Keep in mind that none of this uses grammatical rules. We do statistical modeling as we have been doing for a couple of decades. Statistical modeling is a core element of our work in the Super Human technology program."

This work is a far cry from phrase translators but, like commercial ASR dialogue systems, they are not yet able to go out of domain. According to Gao, "That is the future and that is our goal."