System Description: A Highly Interactive Speech-to-Speech Translation System

Mike Dillinger and Mark Seligman
Spoken Translation, Inc.

Spoken Translation, Inc. (STI) of Berkeley, CA has developed a commercial system for interactive speech-to-speech machine translation designed for both high accuracy and broad linguistic and topical coverage. Planned use is in situations requiring both of these features, for example in helping Spanish-speaking patients to communicate with English-speaking doctors, nurses, and other health-care staff.

The twin goals of accuracy and broad coverage have until now been in opposition: speech translation systems have gained tolerable accuracy only by sharply restricting both the range of topics which can be discussed and the sets of vocabulary and structures which can be used to discuss them. The essential problem is that both speech recognition and translation technologies are still quite error-prone. While the error rates may be tolerable when each technology is used separately, the errors combine and even compound when they are used together. The resulting translation output is generally below the threshold of usability -- unless restriction to a very narrow domain supplies sufficient constraints to significantly lower the error rates of both components.

STI's approach has been to concentrate on interactive monitoring and correction of both technologies.

First, users can monitor and correct the speaker-dependent speech recognition system to ensure that the text which will be passed to the machine translation component is completely correct. Voice commands (e.g. **Scratch That** or **Correct <incorrect text>**) can be used to repair speech recognition errors. While these commands are similar in appearance to those of IBM's ViaVoice or ScanSoft's Dragon NaturallySpeaking dictation systems, they are unique in that they remain usable even when speech recognition operates at a server. Thus they provide for the first time the capability to interactively confirm or correct wide-ranging text which is dictated from anywhere.

Next, during the MT stage, users can monitor, and if necessary correct, one especially important aspect of the translation -- lexical disambiguation.

The problem of determining the correct sense of input words has plagued the machine translation field since its inception. In many cases, the correct sense of a given term is in fact available in the system with an appropriate translation, but for one reason or another it does not appear in the output. Word-sense disambiguation algorithms being developed by research groups have made significant progress, but still often fail; and the most successful still have not been integrated into commercial MT systems. Thus no really reliable solution for automatic word-sense disambiguation is on the horizon for the short and medium term.

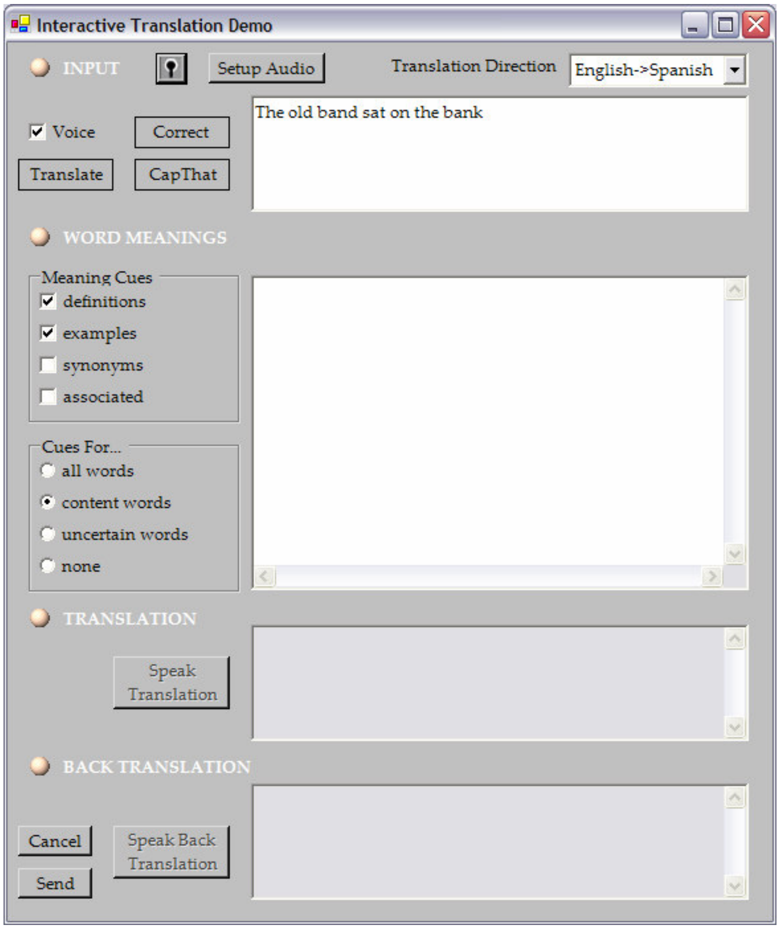
STI's approach to lexical disambiguation is twofold: first, we supply a specially controlled *back translation*, or translation of the translation. Using this paraphrase of the initial input, even a monolingual user can make an initial judgment concerning the quality of the preliminary machine translation output. To make this technique effective, we use proprietary facilities to ensure that the lexical senses used during back translation are appropriate.

In addition, in case uncertainty remains about the correctness of a given word sense, we supply a proprietary set of Meaning Cues™ – synonyms, definitions, etc. – which have been drawn from various resources, collated in a unique database (called SELECT™), and aligned with the respective lexica of the relevant machine translation systems. With these cues as guides, the user can select the preferred meaning from among those available. Automatic updates of translation and back translation then follow.

The result is an utterance which has been monitored and perhaps repaired by the user at two levels – those of speech recognition and translation. By employing these interactive techniques while integrating state-of-the-art dictation and machine translation programs – we work with Philips Speech Processing for speech recognition; with Word Magic and Lingenio for MT (for Spanish and German, respectively); and with ScanSoft for text-to-speech – we have been able to build the first commercial-grade speech-to-speech translation system which can achieve broad coverage without sacrificing accuracy.

A usage example

When run on a Motion Computing Tablet PC, the system has four input modes: speech, typing, handwriting, and touchscreen. To illustrate the use of interactive correction for speech recognition, we will assume that the user has clicked on the microphone icon onscreen to begin entering text by speaking. The image below shows the preliminary results after pronunciation of the sentence "The old man sat on the bank".



Translation Direction: English->Spanish

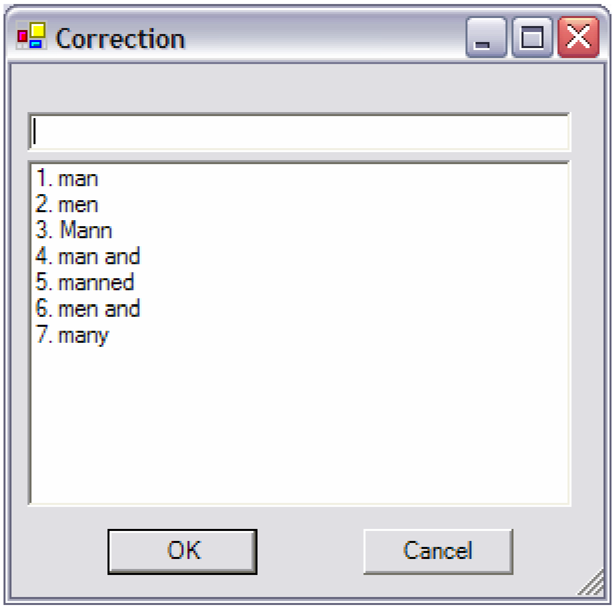
☑ Voice

The old band sat on the bank

Meaning Cues
☑ definitions
☑ examples
☐ synonyms
☐ associated

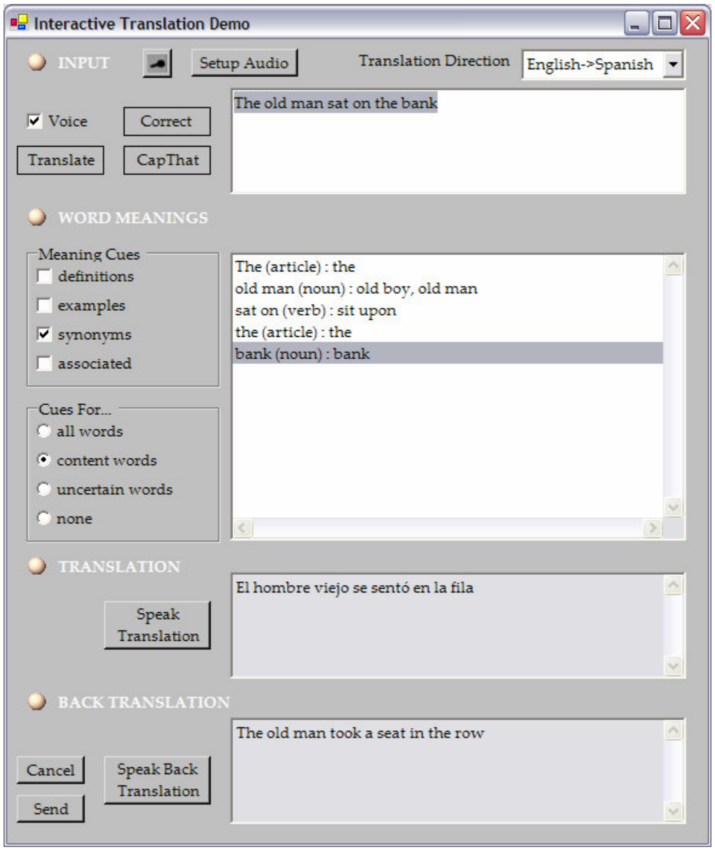
Cues For...
○ all words
● content words
○ uncertain words
○ none

The results of automatic speech recognition are good, but often imperfect. In this example "man" was incorrectly transcribed as "band". Accordingly, the user can perform voice-activated correction by saying "Correct band". A list of alternative speech recognition candidates then appears, seen in the image below. The user can select the correct alternative in this case by saying "Choose one", yielding a corrected sentence. (If the intended alternative is not among the candidates, the user can supply it manually – by typing on a standard keyboard, by using a touchscreen keyboard, or by writing with a stylus for high-accuracy handwriting recognition.)



1. man
2. men
3. Mann
4. man and
5. manned
6. men and
7. many

The spoken (or clicked) “Translate” command provides a translation of the corrected input, seen below in the Translation window. Also provided are a Back Translation (the translated sentence re-translated back into the original, as explained above) and an array of Meaning Cues giving information about the word meanings that were used to perform the translation, seen in the Word Meanings list. The user can use these cues to verify that the system has interpreted the input as intended.



Translation Direction: English->Spanish

☑ Voice

The old man sat on the bank

Meaning Cues
☐ definitions
☐ examples
☑ synonyms
☐ associated

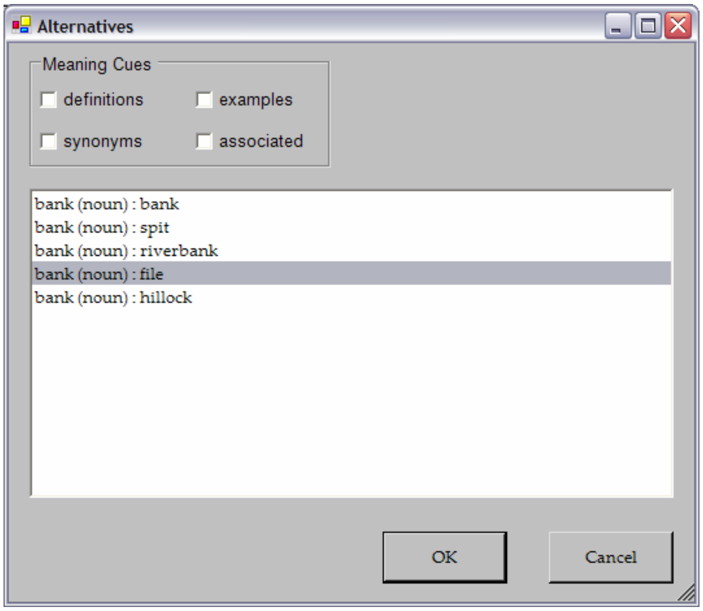
Cues For...
○ all words
● content words
○ uncertain words
○ none

The (article) : the
old man (noun) : old boy, old man
sat on (verb) : sit upon
the (article) : the
bank (noun) : bank

El hombre viejo se sentó en la fila

The old man took a seat in the row

In this example, synonyms are used as Meaning Cues, but definitions, examples, and associated words can also be shown. Here the back-translation ("The old man took a seat in the row") indicates that the system has understood "bank" as meaning "row". Presumably, this is not what the user intended. By clicking on the word in the Word Meanings window, he or she can bring up a list of alternative word meanings, as in the image below.



bank (noun) : bank
bank (noun) : spit
bank (noun) : riverbank
bank (noun) : file
bank (noun) : hillock

When a new word meaning has been chosen from this list, e.g. the "riverbank" meaning in this case, the system updates the display in all windows to reflect that change. In this example, the updated Spanish translation becomes "El hombre viejo se sentó en la orilla del río". The corresponding back translation is now "The old man took a seat at the bank of the river" – close enough, we can assume, to the intended meaning.

When the user is satisfied that the intended meaning has been correctly understood and translated by the system, the system's Send button can be used to transmit the translation to the foreign-language speaker via instant messaging, chat, or on-screen display for face-to-face interaction. At the same time, synthesized speech can be generated, and if necessary transmitted, thus completing the speech-to-speech cycle.

Languages

The current version of the system is for English <> Spanish, and a German <> English version is in development. Discussion is also in progress with several vendors of Japanese MT.

Implementation

The system runs as a stand-alone application on a variety of Windows computers: tablet PC's provide the greatest range of input modes, but the system will run stand-alone on standard current laptops, PCs, etc. With future ports to client-server implementations in mind, it has been programmed with .NET technology.