Behind the Scenes in an Interactive Speech Translation System

Mark Seligman, Ph.D.
Spoken Translation, Inc.
1100 West View Drive
Berkeley, CA 94705
mark.seligman@spokentranslation.com

Mike Dillinger, Ph.D.
Spoken Translation, Inc.
1100 West View Drive
Berkeley, CA 94705
mike@mikedillinger.com

Abstract

This paper describes the facilities of Converser for Healthcare 4.0, a highly interactive speech translation system which enables users to verify and correct speech recognition and machine translation. Corrections are presently useful for real-time reliability, and in the future should prove applicable to offline machine learning. We provide examples of interactive tools in action, emphasizing semantically controlled back-translation and lexical disambiguation, and explain for the first time the techniques employed in the tools' creation, focusing upon compilation of a database of semantic cues and its connection to third-party MT engines. Planned extensions of our techniques to statistical MT are also discussed.

1 Introduction

Multiple applications for spoken language translation (SLT) or automatic interpreting are now in use – SpeechTrans, Jibbigo, iTranslate, and others. SLT projects are in operation at several large communications companies, including Google and Facebook. However, widespread use remains in the future for serious use cases like healthcare, business, emergency relief, and law enforcement, despite demonstrably high demand.

The essential problem is that, despite dramatic advances during the last decade, both speech recognition and translation technologies are still error-prone. While the error rates may be tolerable when the technologies are used separately, the errors combine and even compound when they are used together. The resulting translation output is often below the threshold of usability when accuracy is essential. As a result, present use is still largely restricted to use cases – social networking, travel – in which no representation concerning accuracy is demanded or given.

The speech translation system discussed here, Converser for Healthcare 4.0, applies interactive verification and correction techniques to this essential problem of overall reliability.

First, users can monitor and correct the speech recognition system to ensure that the text which will be passed to the machine translation component is completely correct. Typing or handwriting can be used to repair speech recognition errors.

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

The system's approach to lexical disambiguation is twofold: first, we supply a *back-translation*, or re-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. Other systems, e.g. IBM's MASTOR (Gao, Liang, et al., 2006), have also employed re-translation. Converser, however, exploits proprietary technologies, outlined below, to ensure that the lexical senses used during back-translation accurately reflect those used in forward translation.

In addition, if uncertainty remains about the correctness of a given word sense, the system supplies a proprietary set of Meaning Cues™ – synonyms, definitions, etc. – which have been

Proceedings of the Workshop on Interactive and Adaptive Machine Translation, pages 42–50
AMTA Workshop. Vancouver, Canada. September 22, 2014

drawn from various resources, collated in a database (called SELECT™), and aligned with the respective lexica of the relevant MT systems. With these cues as guides, the user can monitor the current, proposed meaning and when necessary select a different, preferred meaning from among those available. Automatic updates of translation and back-translation then follow.

The initial purpose of these techniques is to increase reliability during real-time speech translation sessions. Equally important, however, they can also enable even monolingual users to supply feedback for off-line machine learning to improve the system. Until now, only users with some knowledge of the output language have been able to supply such feedback, e.g. in Google Translate.

Previous papers (Seligman and Dillinger 2013, 2012, 2011, 2008, 2006a, 2006b, Dillinger and Seligman 2004a, 2004b) have reported on the user-facing design and use of the facilities just described. Here we provide updated examples of interactive facilities and explain for the first time how they were constructed.

For orientation, Section 2 of this paper will review Converser's current interactive facilities. Section 3 explains the implementation of the system's back-translation and Section 4 does the same for its lexical disambiguation facilities. We conclude in a final section.

Converser has been pilot tested successfully at a San Francisco medical center, part of a very large healthcare organization (Seligman and Dillinger, 2011). Evaluation results concerning system accuracy and usability are discussed below.

Negotiations concerning continued use are ongoing with the host of the pilot and with another large Bay Area hospital system.

2 The Converser System

We now briefly illustrate Converser's approach to interactive automatic interpretation. We describe Version 4.0, *italicizing* new interface elements.

Converser adopts rather than creates its speech and translation components, adding value through the interactive interface elements to be explained. Nuance, Inc. supplies cloud-based speech recognition, modified by a third party for access via desktops; rule-based English↔Spanish machine translation is supplied by Word Magic of Costa Rica; and text-to-speech is again provided by Nuance.

Depending on the platform, the system can offer up to four input modes: speech, typing, handwriting, and touchscreen. Since we want to illustrate the use of interactive correction for speech recognition as well as machine translation, we assume that the user has clicked on the round red ***Mic Button*** to activate the microphone (Figure 1). (Starting with the 4.0 release, *no voice training or profile creation is required for either language.*)

Still in Figure 1, notice the ***Traffic Light Icon*** and two ***Earring Icons***. These are used to switch *Verification Mode* on and off for translation and speech recognition, respectively. Both icons are currently green, indicating "Full speed ahead!" That is, verification has been temporarily switched off: the user has indicated that it is unnecessary to pre-check either ASR or MT before transmitting the next utterance, preferring speed to accuracy.

Just prior to the figure's snapshot, the user said, "San Jose is a pleasant city." Since verification had been switched off for both ASR and MT, these functioned without interruption. The speech recognition result appeared briefly (and in this case correctly) in the **Input Window**. Immediately thereafter the Spanish translation result (also correct in this case) appeared in the right-hand section of the **Transcript Window**, and was immediately pronounced via text-to-speech. Meanwhile, the original English input was recorded in the left-hand section of the transcript.

Also on the English side of the transcript and just below the original English input is a specially prepared back-translation:[1] the original input was translated into Spanish, and then retranslated back into English. Techniques to be explained in Section 3 ensure that the back-translation means the same as the Spanish. Thus, even though *pre*-verification was bypassed for this utterance in the interest of speed, *post*-verification via the transcript was still enabled. (The **Transcript Window**, containing inputs from both English and Spanish sides and the associated back-translations, can be saved for record-keeping. *Inclusion of back-translation is new to Version 4.0.* Participant identities can optionally be masked for confidentiality.)

Using this back-translation, the user might

[1] Proprietary, and branded as Reliable Retranslation™.

conclude that the translation just transmitted was inadequate. In that case, or if the user simply wants to rephrase this or some previous utterance, she can click the ***Rewind Button*** (round, with chevrons). A menu of previous inputs then appears (not shown). Once a previous input is selected, it will be brought back into the **Input Window**, where it can be modified using any available input mode – voice, typing, or handwriting. In our example sentence, for instance, *pleasant* could be changed to *boring*; clicking the **Translate Button** would then trigger translation of the modified input, accompanied by a new back-translation.

In Figure 2, the user has selected the *yellow* ***Earring Icon***, specifying that the speech recognition should "proceed with caution." As a result, spoken input remains in the **Input Window** until the user explicitly orders translation. Thus there's an opportunity to make any necessary or desired corrections of the ASR results. In this case, the user has said "This morning, I received an email from my colleague Igor Boguslavsky." The name, however, has been misrecognized as "Igor bogus Lovsky." Typed or handwritten correction can fix the mistake, and the **Translate Button** can then be clicked to proceed.

Just prior to Figure 3, the ***Traffic Light Icon*** was also switched to yellow, indicating that translation (as opposed to speech recognition) should also "proceed with caution": it should be pre-checked before transmission and pronunciation. This time the user said "This is a cool program." Since the ***Earring Icon*** is still yellow, ASR results were pre-checked and approved. Then the **Translation Verification Panel** appeared, as shown in the figure. At the bottom, we see the preliminary Spanish translation, "Éste es un programa frío." Despite the best efforts of the translation program to determine the intended meaning in context, "cool" has been mistranslated – as shown by the back-translation, "This is a cold program."

Another indication of the error appears in the **Meaning Cues Window** (third from the top), which indicates the meaning of each input word or expression as currently understood by the MT engine. Converser 4.0 employs synonyms as Meaning Cues, compiled using techniques to be explained in Section 4. (In the future, pictures, definitions, and examples may also be used.) In the present case, we see that the word "cool" as been wrongly translated as "cold, fresh, chilly, …".

To rectify the problem, the user double clicks on the offending word or expression. The **Change Meaning Window** then appears (Figure 4), with a list of all available meanings for the relevant expression. Here the third meaning for "cool" is "great, fun, tremendous, …". When this meaning has been selected, the entire input is retranslated. This time the Spanish translation will be "Es un programa estupendo" and the translation back into English is "Is an awesome program." The user may accept this rendering, despite the minor grammatical error, or may decide to try again.

The ***Traffic Light*** and ***Earring Icons*** help to balance a conversation's reliability with its speed. Reliability is indispensible for serious applications like healthcare, but some time is required to interactively enhance it. The icons let users proceed carefully when accuracy is paramount or a misunderstanding must be resolved, but more quickly when throughput is judged more important. This flexibility, we anticipate, will be useful in future applications featuring automatic detection of start-of-speech: in Green Light Mode, ASR and translation will proceed automatically without start or end signals and thus without demanding the user's attention, but can be interrupted for interactive verification or correction as appropriate. Currently, in the same mode, for inputs of typical length (ten words or less), the time from end of input speech to start of translation pronunciation is normally less than five seconds on a 2.30 GHz Windows 7 desktop with 4.00 GB RAM, and faster in a pending cloud-based version.

Statistics have not yet been compiled to determine how many corrections are typically needed to obtain translations which users consider satisfactory. However, in a survey performed by an independent third party during the abovementioned pilot project at a national healthcare organization, when 61 users (staff and patients) were asked whether the system met their needs, 93% responded either Completely or Mostly. Translation was judged accurate by 90%; and the system (Version 3.0, in which verification was still mandatory) was found easy to use by 57%. Unfortunately, these results have been published only in internal reports marked confidential.

Translation Shortcuts. The Converser system includes Translation Shortcuts™ – pre-packaged translations, providing a kind of translation memory. When they're used, re-verification of a given utterance is unnecessary, since Shortcuts are pre-translated by professionals (or, in future versions of the system, verified using the system's feedback and correction tools). Access to stored Shortcuts is very quick, with little or no need for text entry. *Shortcut Search* can retrieve a set of relevant phrases given only keywords or the first few characters or words of a string. (If no Shortcut is found to match the input text, the system seamlessly gives access to broad-coverage, interactive speech translation.) A **Translation Shortcuts Browser** is provided (on the left in Figure 1), so that users can find needed phrases by traversing a tree of Shortcut categories, and then execute them by tapping or clicking. Shortcuts are fully discussed in (Seligman and Dillinger, 2006a).

Symmetry. Identical facilities are available for Spanish speakers: when the Spanish flag is clicked, all interface elements – buttons and menus, onscreen messages, Translation Shortcuts, handwriting recognition, etc. – change to Spanish.

Having surveyed the Converser interface, we now go on to look behind the scenes, discussing the system's specially controlled back-translation and its lexical disambiguation facilities.

3 Back-translation

Back-translation – translation from the target language back into the source language – suggests itself as a way to show users how accurately an input has been translated. However, the technique has until now been of limited use because mistakes can occur during backward translation which bear no relation to any errors made during the original, forward translation. The forward and backward translations, in other words, are normally separate and unrelated processes. For this reason, automatic back-translation has remained more a source of amusement than a useful indicator of translation accuracy. Converser aims to make back-translation more useful for verification by forging a closer relationship between the forward and backward translation processes.

To illustrate, assume that the user wants to translate the ambiguous English word *bank* into Spanish. Of course, the word can mean "bank as in money," "bank as in river," "bank of switches," etc. (Figure 5). However, the worse problem for back-translation is that the respective Spanish translations for some of these meanings are themselves ambiguous. For example, the word *banco*, which would be appropriate for the "money bank" meaning, also has the meaning "bench." Accordingly, semantically uncontrolled back-translation can fail as follows: the user says "bank," intending the "money bank" meaning; the translation system gives the correct translation *banco* (whether through skill or luck); the system is asked for a revealing back-translation; and it brightly and misleadingly responds, *bench*. No good: the translation was in fact what the user wanted, but the back-translation erroneously indicated otherwise, since the uncontrolled system had forgotten the forward translation by the time the back-translation was requested.

Converser addresses the problem by remembering which meaning of *bank* was used during the forward translation and forcing reuse of the *same* meaning during backward translation. If the "money bank" meaning was used, leading to a translation of *banco*, then that meaning – right or wrong – will be used during back-translation as well, leading to such translations as *financial institution*, *cash repository* … or to the original input, *bank*. In the latter case, uncertainty about the translation accuracy would remain; but two recourses are on hand. First, the system can be directed to avoid the original input during back-translation if any synonyms are available. However, when this strategy was experimentally applied to whole utterances, wordy or unnatural paraphrases often resulted. The second remedy is to make use of Meaning Cues for lexical disambiguation. By examining the synonyms of *bank*, the user can determine which meaning has actually been translated. Back-translation thus provides an initial check on translation meaning, sufficient in many cases; and when ambiguity remains, the Meaning Cues remain as a fallback. We have found this second solution to be the more helpful till now. Further experiments with the synonym-based solution may be resumed in the future.

But how is "same meaning" represented in the system, whether it is used to synchronize forward and backward translation or to find synonyms that can be substituted for original terms during backward translation? We now make use of an MT engine whose lexicon elements are *Meaning IDs* (*MIDs*) – semantic elements comparable to the synsets (synonym sets) of WordNet (Miller, 1995; Fellbaum et al, 1998). Thus during back-translation or synonym substitution, the system can enforce re-use of specific MIDs. (MIDs are illustrated in Figure 6 as e.g. MID#567122-567127.) Later in the paper, we'll comment on comparable techniques for statistical machine translation (SMT) systems.

4 Lexical Disambiguation

In Section 2, we illustrated Converser's facility for lexical disambiguation using Meaning Cues. The cues are not part of the third-party MT engine itself, but are added by Converser as a bridge between that engine and the user. This section explains how the addition is accomplished.

The explanation will refer to the rule-based machine translation engine presently in use; but again, we'll sketch below how the procedures can be extended to statistical engines.

As mentioned, the main lexicon of our engine is composed of Meaning IDs or MIDs, semantic elements comparable to WordNet's synsets. However, while these unique identifiers are suitable for programming, they remain opaque to human readers, as seen on the left of Figure 6. Hence there is a need to elucidate their meanings for those readers. Many suitable cues are available in the public domain – synonyms, pictures, examples, definitions, and others. The problem is to associate these with the lexicon's opaque symbols.

The first step toward this link-up is to collect relevant cues. We then sort collected cues into semantic groups, using techniques described in (Seligman et al 2004). In essence, we define and exploit best-match metrics for grouping purposes, for instance in terms of maximum intersection among such elements of interest as synonym sets or definitions. The result is a proprietary database called SELECT, a collection of Meaning Cue Groups, as seen on the right of Figure 6.

The remaining task is to map or align every group of semantic cues with an appropriate MID, if one can be found, in the current machine translation lexicon. A successful mapping, as portrayed in Figure 6, will for instance associate MID#567123 with the Meaning Cue Group containing cues for *bank* in the money sense, while MID#567124 maps to *bank* in the river sense, and so on. Three such associations are represented by arrows in the figure. The techniques for automating MID-to-cue-group mappings are described in (Seligman et al 2004). Groupings and mappings are checked by linguists, so the overall process of adding Meaning Cues to the native MT engine can be described as semi-automatic.

Statistical machine translation. When extending these techniques to statistical MT engines, we plan to proceed as follows:

Begin with a standard SMT phrase table, in which each line represents a source language term and a possible translation. Employ a *paraphrase extraction tool* to create a secondary table in which each line is a source language term and one possible synonym. Consolidate such synonym lines to compose synsets, or synonym sets. Finally, collect synsets related to a given word or expression to yield *sets* of synsets, equivalent to sets of Meaning Cues seen in the **Change Meaning Window** of Figure 4. These can be presented to users as described above to enable word meaning choices. Once a preferred meaning has been selected, e.g. for *cool*, a new translation can be generated by modifying translation probabilities during decoding, or by re-ranking candidate translations following decoding. (Temporary data structures can be used to avoid premature alteration of permanent ones. However, temporary results can eventually be integrated into master structures to improve translation results.)

To create meaning-preserving back-translations of an input sentence in an SMT context, we first identify, for each word or expression in the input, the meaning (represented as a synset) used in the forward translation. To make this identification, we observe the currently proposed translation of the current word. For example, it might be English *cool*, provisionally translated as *frio*. We compose a synset containing English synonyms for *cool* which according to the translation table can likewise be translated as *frio*. Then, armed with the meaning (synset) of every expression in the input sentence, we exploit the techniques just explained to force a new meaning-preserving translation.

5 Conclusions

The first purpose here has been to give an updated view of the toolset for highly interactive speech translation in Converser for Healthcare 4.0, with emphasis upon lexical disambiguation. We've illustrated the new interface's handling of several examples, involving the ***Rewind Button***, icons for switching *Verification Mode* on and off for speech recognition and translation, the **Verification Panel**, and the **Change Meaning Window**.

The second goal has been to give a look backstage: we've explained in outline how semantically controlled back-translation and Meaning Cues have been implemented, with a look-ahead toward their extension into statistical machine translation.

Future work will feature actual implementation of interactive SMT, enabling interactive spoken language translation among many more languages.

Acknowledgments

The authors thank the many participants in the development of Converser for Healthcare and look forward to thanking by name the organization which sponsored a pilot project for Converser.

References

Mike Dillinger and Mark Seligman. 2004a. “System Description: A Highly Interactive Speech-to-speech Translation System.” Association for Machine Translation in the Americas (AMTA-04). Washington, DC, September 28 – October 2, 2004.

Mike Dillinger and Mark Seligman. 2004b. “A highly interactive speech-to-speech translation system.” In *Proceedings of the VI Conference of the Association for Machine Translation in the Americas*. Washington, D.C., September-October, 2004.

Christiane Fellbaum, ed. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Yuqing Gao, Gu Liang, Bowen Zhou, Ruhi Sarikaya, Mohamed Afify, Hong-Kwang Kuo, Wei-zhong Zhu, Yonggang Deng, Charles Prosser, Wei Zhang, and Laurent Besacier. 2006. “IBM MASTOR system: multilingual automatic speech-to-speech translator.” In *HLT-NAACL 2006: Proceedings of the Workshop on Medical Speech Translation*. New York, NY, June, 2006.

George Miller. 1995. “WordNet: A Lexical Database for English.” Communications of the AMC, Vol. 38, No. 11:39-41.

Mark Seligman, Mike Dillinger, Barton Friedland, and Gerald Richard Cain. 2014. “Method and Application for Cross-lingual Communication.” US Patent 7, 539, 619, Application for Continuation 13797628.130326.

Mark Seligman and Mike Dillinger. 2013. “Automatic Speech Translation for Healthcare:: Some Internet and Interface Aspects.” TIA (Terminology and Artificial Intelligence) 2013: Proceedings of the Workshop on Optimizing Understanding in Multilingual Hospital Encounters. Paris, France, October 30, 2013.

Mark Seligman and Mike Dillinger. 2012. “Spoken Language Translation: Three Business Opportunities.” Association for Machine Translation in the Americas (AMTA-12). San Diego, CA, October 28 – November 1, 2012.

Mark Seligman and Mike Dillinger. 2011. “Real-time Multi-media Translation for Healthcare: a Usability Study.” Proceedings of the 13th Machine Translation Summit. Xiamen, China, September 19-23, 2011.

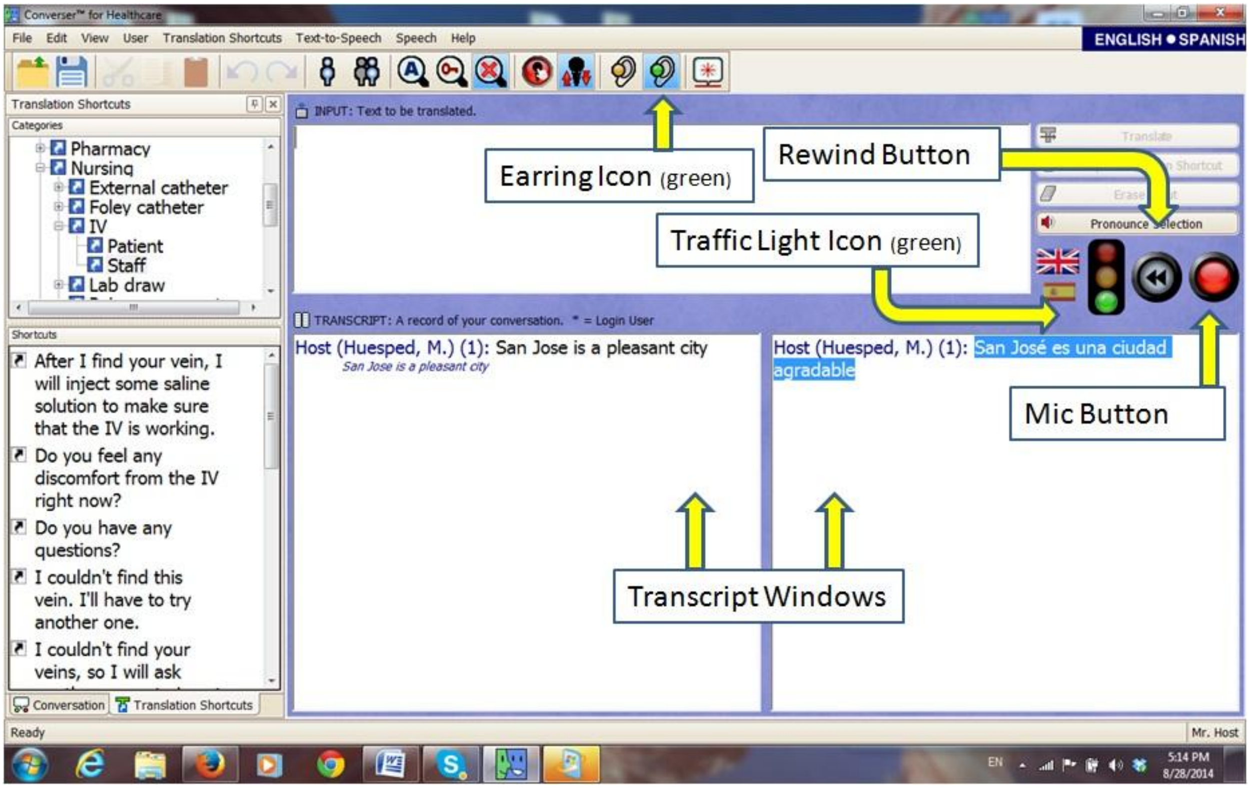
Mark Seligman and Mike Dillinger. 2008. “Rapid Portability among Domains in an Interactive Spoken Language Translation System.” *COLING 2008: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*. Manchester, UK, August 23, 2008, pages 40-47.

Mark Seligman and Mike Dillinger. 2006a. “Usability Issues in an Interactive Speech-to-Speech Translation System for Healthcare.” HLT/NAACL-06: Proceedings of the Workshop on Medical Speech Translation. NYC, NY, June 9, 2006.

Mark Seligman and Mike Dillinger. 2006b. “Converser: Highly Interactive Speech-to-speech Translation for Healthcare.” HLT/NAACL-06: Proceedings of the Workshop on Medical Speech Translation. NYC, NY, June 9, 2006.

Mark Seligman, Mike Dillinger, Barton Friedland, and Gerald Richard Cain. 2004. “Method and Application for Cross-lingual Communication.” US Patent 7, 539, 619.

Alexander Waibel. 2012. http://innovation.mfg.de/en/news-and-features/simultaneous-translation-university-without-language-barriers-1.11379



After I find your vein, I will inject some saline solution to make sure that the IV is working.

Do you feel any discomfort from the IV right now?

Do you have any questions?

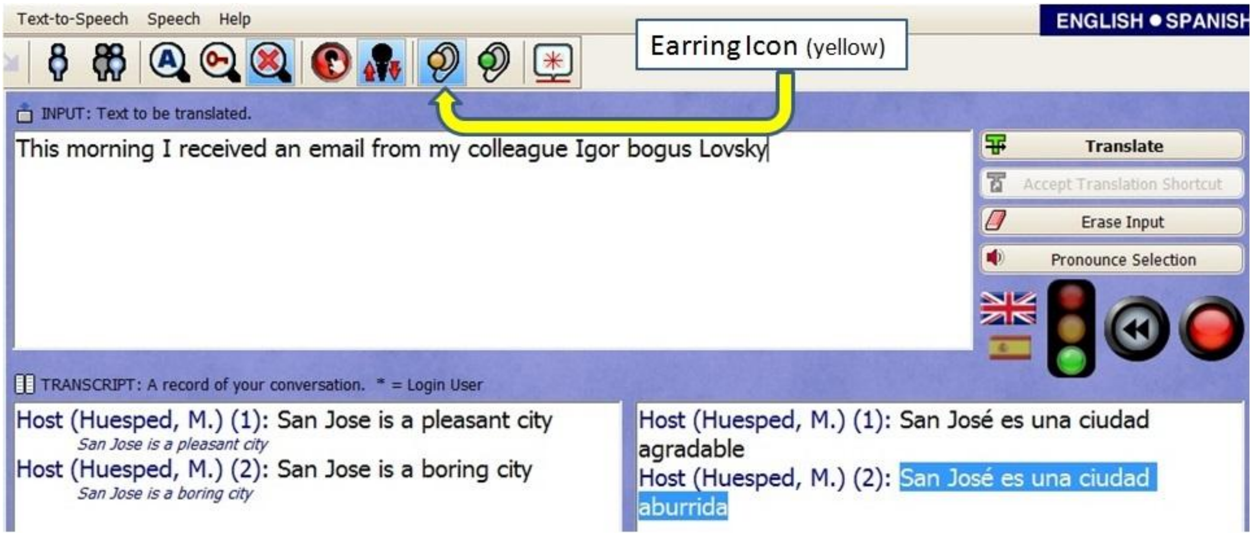
I couldn't find this vein. I'll have to try another one.

I couldn't find your veins, so I will ask

Host (Huesped, M.) (1): San Jose is a pleasant city
San Jose is a pleasant city

Host (Huesped, M.) (1): San José es una ciudad agradable

Figure 1: Earring and Traffic Light Icons are green: "Full speed ahead!"

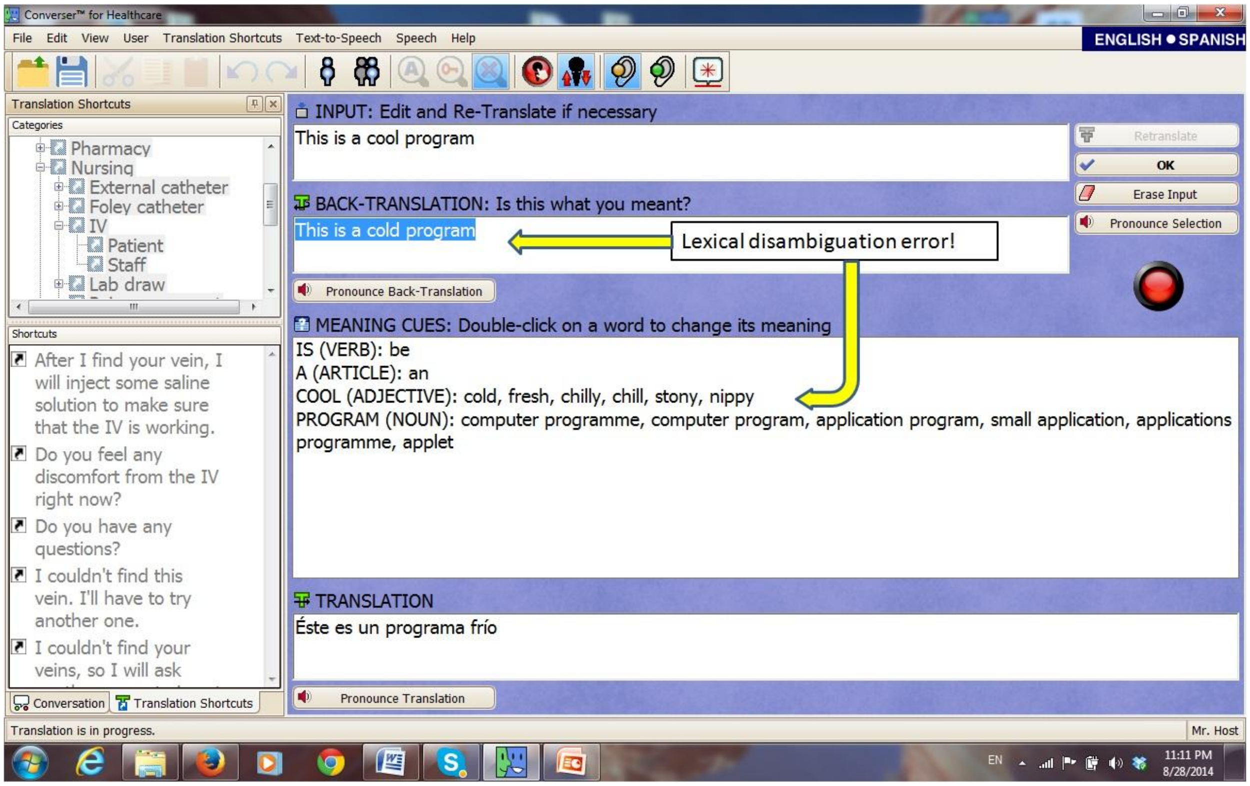

Text-to-Speech Speech Help
ENGLISH ● SPANISH
Earring Icon (yellow)
INPUT: Text to be translated.
Translate
Accept Translation Shortcut
Erase Input
Pronounce Selection
TRANSCRIPT: A record of your conversation. * = Login User

This morning I received an email from my colleague Igor bogus Lovsky

Host (Huesped, M.) (1): San Jose is a pleasant city
San Jose is a pleasant city
Host (Huesped, M.) (2): San Jose is a boring city
San Jose is a boring city

Host (Huesped, M.) (1): San José es una ciudad agradable
Host (Huesped, M.) (2): San José es una ciudad aburrida

Figure 2: Earring Icon is yellow: "Proceed with caution!"



After I find your vein, I will inject some saline solution to make sure that the IV is working.

Do you feel any discomfort from the IV right now?

Do you have any questions?

I couldn't find this vein. I'll have to try another one.

I couldn't find your veins, so I will ask

INPUT: Edit and Re-Translate if necessary

This is a cool program

BACK-TRANSLATION: Is this what you meant?

This is a cold program

Lexical disambiguation error!

MEANING CUES: Double-click on a word to change its meaning

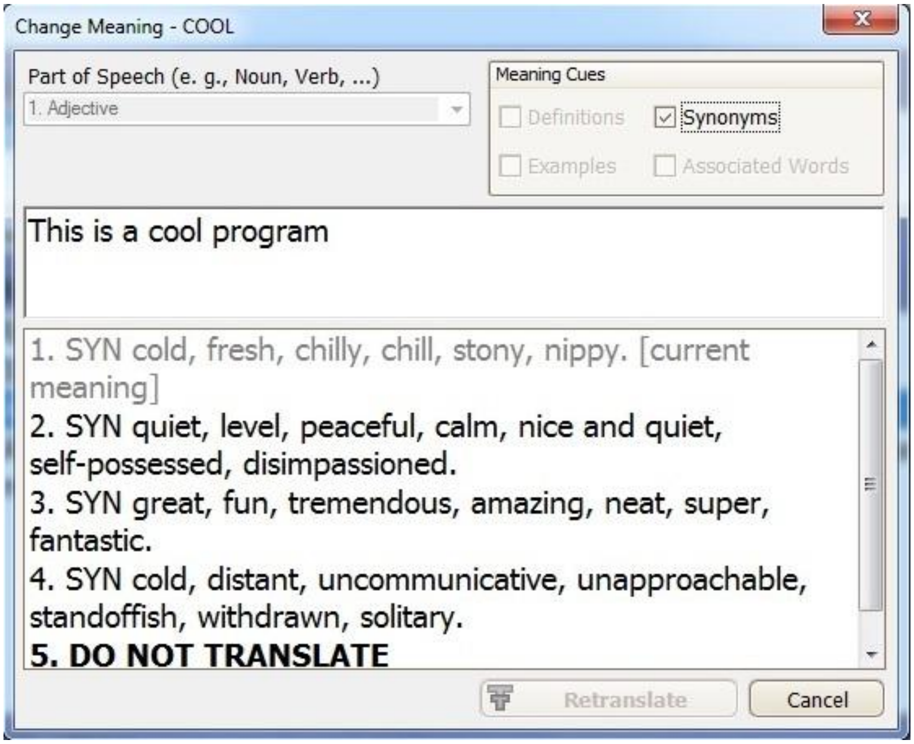
IS (VERB): be
A (ARTICLE): an
COOL (ADJECTIVE): cold, fresh, chilly, chill, stony, nippy
PROGRAM (NOUN): computer programme, computer program, application program, small application, applications programme, applet

TRANSLATION

Éste es un programa frío

Translation is in progress.

Figure 3: Verification Panel, with a lexical disambiguation error in *This is a <u>cool</u> program*.



1. Adjective

□ Definitions ☑ Synonyms
□ Examples □ Associated Words

This is a cool program

1. SYN cold, fresh, chilly, chill, stony, nippy. [current meaning]
2. SYN quiet, level, peaceful, calm, nice and quiet, self-possessed, disimpassioned.
3. SYN great, fun, tremendous, amazing, neat, super, fantastic.
4. SYN cold, distant, uncommunicative, unapproachable, standoffish, withdrawn, solitary.
5. DO NOT TRANSLATE

Figure 4: The Change Meaning Window, with four meanings of *cool*.

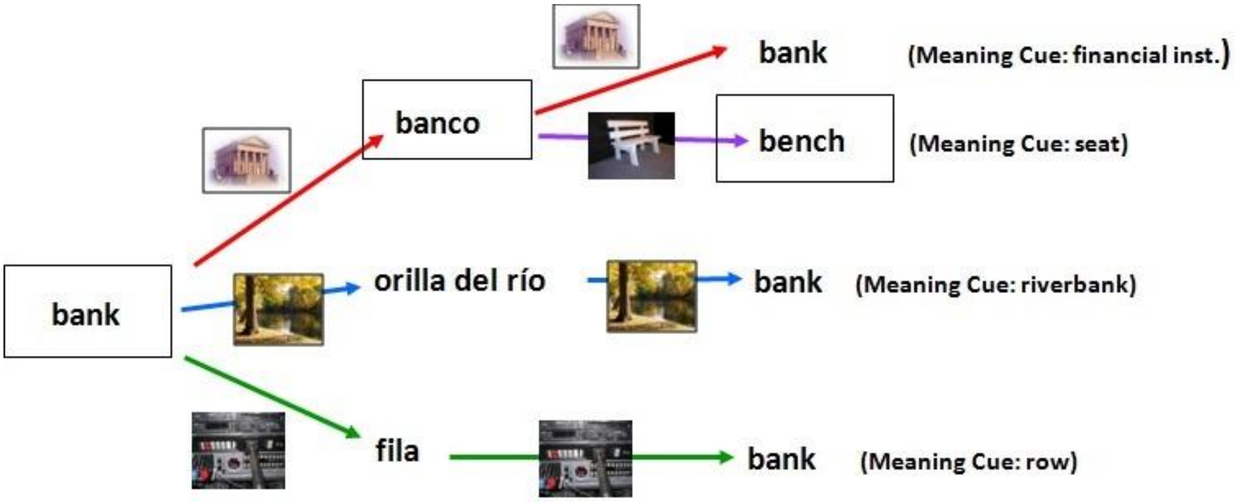


Figure 5: Translation and erroneous back-translation of *bank*.

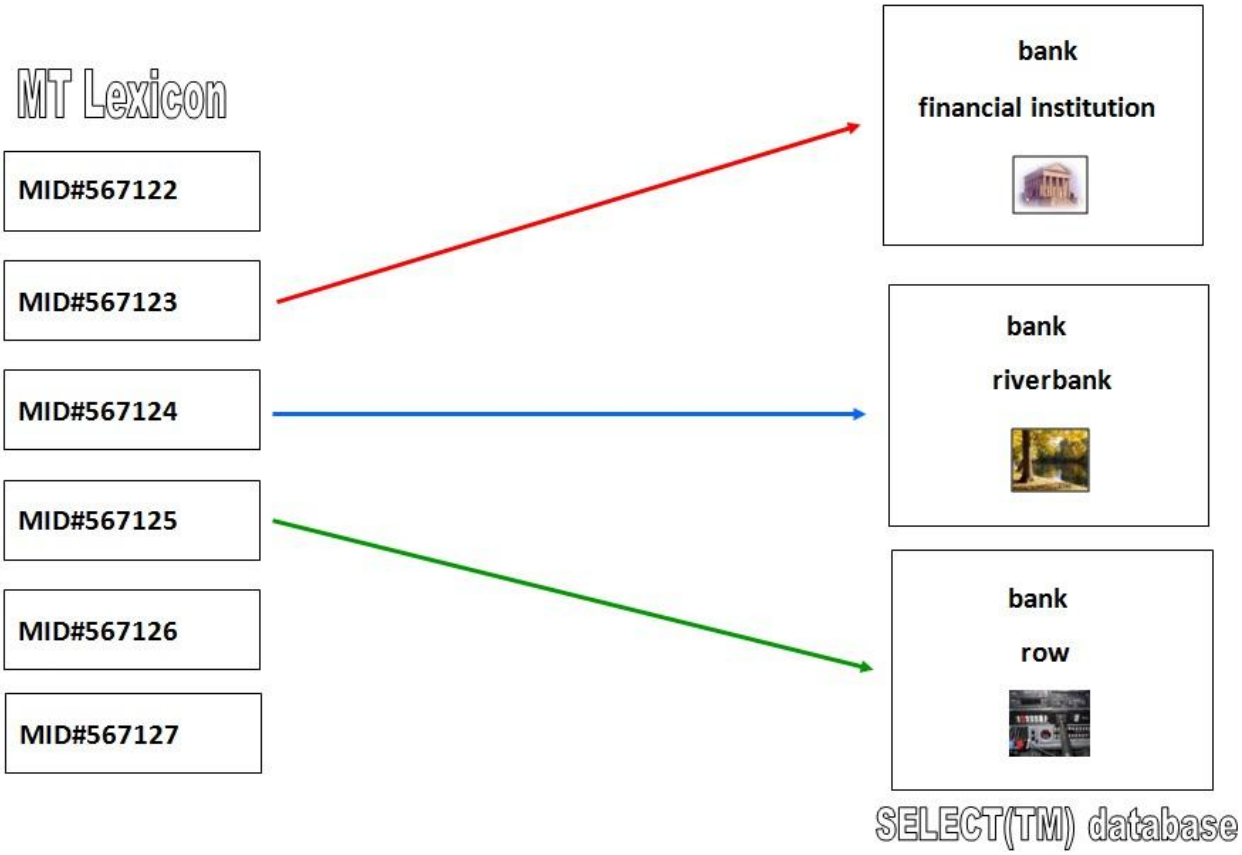


Figure 6: Mapping between MIDs in an MT lexicon and Meaning Cues in the SELECT database.