Interactive Real-time Translation via the Internet

Mark Seligman
Université Joseph Fourier
GETA, CLIPS, IMAG-campus, BP53
150, rue de la Chimie
38041 Grenoble Cedex 9, France
seligman@cerf.net

Abstract

The Internet offers a tremendous opportunity for experiments with real-time machine translation (MT) of dialogues. Interchanges may be typed, or speech recognition (SR) techniques might be used to accept spoken input. In either case, given the number of ambiguities still escaping automatic resolution, interactive disambiguation for both components seems likely to remain important in many Internet applications for some time to come. In fact, simple concatenation of interactive SR and interactive MT may offer a "low road" to "quick and dirty" speech translation, as a pragmatic near-term alternative to tightly integrated systems aiming for more automatic operation. Disambiguation at the speech recognition stage is particularly important as a foundation for speech translation, in view of the need to "kill ambiguity before it multiplies". I briefly discuss interactive disambiguation for MT. Interactive disambiguation for speech recognition is the next topic: isolated-word speech recognition is seen as an especially interactive input mode with several advantages which are usually overlooked. Discussion then turns to ways of transmitting translated text via the Internet, with consideration of ftp and Internet Relay Chat. A pioneering demo of speech translation over the Internet organized by Kowalski, Rosenberg, and others is described, and possible alternative architectures are examined.

Introduction

Text can now be transmitted inexpensively over the Internet in real time, and a range of relatively inexpensive speech recognition and translation components are becoming available. As a result, there is now a tremendous opportunity for experiments with real-time machine translation (MT) of dialogues over the Net. Interchanges could be typed, or speech recognition (SR) techniques might be used to accept spoken input in order to create speech translation systems.

The central theme of this paper is that interactivity will be crucial for exploiting this opportunity. That is, given the number of ambiguities still escaping automatic resolution in current SR and MT components, *interactive disambiguation* for both components seems likely to remain important for many Internet translation applications for some time to come, as Section 1 will argue. For speech translation, disambiguation at the speech recognition stage will be described as particularly important in laying a firm foundation for translation, in view of the need to "kill ambiguity before it multiplies".

Section 2 will propose that in fact simple concatenation of interactive SR and interactive MT may offer a "low road" to "quick and dirty" speech translation, as a pragmatic near-term alternative to tightly integrated systems aiming for more automatic translation (though integrated systems appear more theoretically satisfying and promising in the longer term).

Section 3 will briefly discuss interactive disambiguation for MT, demonstrating some possible modes of interaction by reference to the LIDIA-1.2 project (Boitet 1996, Blanchon 1996). This dialogue-based MT system was designed for off-line use, but might be adapted for on-line use as well.

Interactive disambiguation for speech recognition is next discussed (in Section 4). In particular, isolated-word (as opposed to continuous) speech recognition will be viewed as an especially interactive input mode with several advantages which are usually overlooked.

Finally, discussion turns in Section 5 to transmission of translated text via the Internet, with consideration of ftp and Internet Relay Chat. A pioneering Internet demo of speech translation organized by (Kowalski, Rosenberg, & Krause 1995) will be described, and alternative dialogue translation architectures for the future will be discussed.

1. The Need for Interaction in Translation

The argument for interaction during translation, especially speech translation, is straightforward: at the present state of the art, attempts to automatically translate even simple texts often leave many ambiguities which cannot reliably be automatically resolved. Programs cannot yet resolve these ambiguities; humans (usually) can; so humans should help the programs, at least until they can better help themselves.

The ambiguities at issue may be structural (such as uncertain attachments for prepositional phrases, as in *The man saw the woman in the park with the telescope*) or lexical (such as uncertainty whether *bank* means "edge of river" or "lending institution"). Further, information necessary for smooth translation into the target language is often missing in the source language input: for instance, a Japanese noun will normally lack information about definiteness and number which would be needed for translation into natural English. This underspecification, too, can be viewed as a kind of ambiguity, since it leaves open many possible interpretations.

To the extent that humans must intervene, they should do so early. Translation ambiguities tend to multiply exponentially as several sources of uncertainty permute within translation stages — within analysis, or transfer, or generation — and as earlier stages pass unresolved problems to later ones. A basic motivation for human intervention during text translation is to resolve ambiguities before they can permute or propagate. The hope is that by intervening within or between translation stages to select and prune, users can reduce the uncertainty which is passed along. To nip propagating ambiguities in the bud, intervention should take place as early as possible. A slogan for this aspect of the case for interactive translation: "Kill it (ambiguity) before it multiplies".

But if this argument for early human intervention is persuasive for text translation, it should be even more persuasive for translation based on spoken input. Speech recognition introduces very considerable uncertainty of its own, and does so at the earliest stage in the overall speech translation process, where there is a maximum possibility that ambiguity will be amplified by later stages. If much uncertainty remains in the speech recognition results, the whole speech translation process stands on a very shaky foundation. Any efforts to eliminate that uncertainty will certainly be worthwhile.

2. The High Road Vs. the Low Road to Speech Translation

I have been arguing that as long as many translation ambiguities escape automatic resolution, interactive disambiguation will remain desirable; and that, when disambiguation is necessary, it should be done as early as possible. In systems with speech input, early intervention — that is, disambiguation of speech recognition candidates — will be especially important.

Since speech translation provides the greatest scope for the interaction which is our central interest here, and moreover presents an especially exciting prospect for communication over the Internet, we can concentrate on this ambitious translation mode for the balance of the discussion. I now want to suggest that simple concatenation of interactive speech recognition with interactive machine translation may suffice for a "quick and dirty" speech translation system suitable for widespread use on the Internet.

Such a pragmatic approach contrasts with that of most current speech translation research, where the aim is tight integration among system component in order to achieve maximally automatic speech translation. (An example of the integrated approach is the Verbmobil speech translation system (Wahlster 1993). (Seligman & Boitet 1993) propose one possible architecture for such integration.)

The purpose of tight integration between speech and translation is to allow two-way feedback, so that each component can help the other to resolve its ambiguities. In the "quick and dirty" or concatenative approach, however, much of the need for integration between speech and translation is removed, since its burden is passed to a human operator: because the operator is allowed to completely disambiguate the speech input before it is passed to the translation component, feedback from translation to speech becomes redundant; while in the reverse direction, ideal transmission of segmental information from speech to translation is achieved, since a correct outcome is guaranteed. (Non-segmental, or prosodic, information is absent, however. The aid it might give to analysis must be supplied by the operator during interactive MT disambiguation.) Following this "low road" to speech translation, the goal of relatively automatic or transparent dialogue translation is postponed in hopes of obtaining a usable system sooner.

As a matter of fact, the "quick and dirty" approach to speech translation, in which a human intervenes between speech and translation, has been applied for practical reasons even in systems which ideally aim for maximally automatic operation. For example, in the ASURA speech translation system (Morimoto 1993) used in a widely-reported international demo, speech recognition provided an n-best list of utterance candidates. If the correct text string was among the candidates, the operator selected it, and it became the input to translation — now guaranteed to be correct. If the correct string was absent, the entire utterance had to be repeated.

As the goal of the system was maximally automatic operation, the lack of real integration between speech and translation might be considered a weakness of the system. However, the arrangement did increase its reliability to the point of usability, and this seems crucial. Such interactive

or interventionist speech-to-translation interfaces may offer the quickest path to speech translation systems which are reliable enough for actual use. And actual use is vital: actual experience with working systems may well lead to rapid progress, even if the systems are initially relatively primitive.

It should be clear that the ultimate goal remains relatively automatic speech translation based upon tight integration between speech recognition and translation. Greater automaticity is intrinsically desirable; there can be little doubt that it will depend on knowledge source integration; and such integration is in any case a fundamental issue in cognitive science and computational engineering. But one possible path to this goal is to begin less ambitiously in order to gather momentum more quickly. As the Scottish song goes, "You take the high road, and I'll take the low road, and I'll be in Scotland beforre ye." While work continues on fully automatic, highly integrated speech translation systems as a long-term goal, it is prudent to explore highly interactive, less integrated approaches for the near term. Both high and low paths can, and I think should, be pursued in tandem.

3. Interactive Disambiguation for Machine Translation

I have been arguing for interactive disambiguation during machine translation, especially speech translation. Some description of such interaction will now be helpful, first for interactive MT, then for interactive speech recognition.

In interactive MT systems, the system first tries to resolve by itself as many ambiguities as possible. Among the problems which remain, it may select those which it judges most crucial. It then presents the user with questions, usually multiple choice, whose answers should resolve the structural or lexical ambiguity at issue.

Description of a recent dialogue-based MT system, LIDIA-1.2, and full references to earlier work can be found in (Boitet 1996), to appear in the proceedings of MIDDIM-96 (the International Seminar on Multimodal Interactive Disambiguation, Col de Porte, France, August 11 - 15, 1996). A good idea of the range of current work in this area can be gained by browsing this volume.

Boitet's description of the program's operation:

> ... the document is first interactively "cleaned" and "tagged" (for special terms or idioms, utterance styles, etc.). The units of translation are then sent to a state-of-the-art, all-paths analyzer written in Ariane-G5 and running on a server, which returns a structure factorizing all ambiguities which cannot be reliably solved with the knowledge available. A question mark then appears next to each ambiguous unit of translation. The user clicks on it to activate a disambiguation dialogue which doesn't suppose any particular expertise in linguistics, computer science, or any of the target languages.

To resolve structural ambiguities, the disambiguation dialogue presents the user with a menu of possible interpretations in the source language. Each interpretation is based on canonical patterns which have been chosen for the type of ambiguity found. For example, to resolve two senses of *L'ouvrier creuse la chaussée et l'asphalte* (which may mean "the worker digs the road and the asphalt" or "the worker digs the road and asphalts it"), the program asks the user to choose between *L'ouvrier creuse l'asphalte* ("the worker digs the asphalt") and *L'ouvrier asphalte la chaussée* ("the worker asphalts the road"). Similarly, to resolve lexical ambiguities, the user is asked to choose among several definitions in the source language.

LIDIA-1.2 and its predecessors have been designed for off-line translation of text rather than for real-time dialogue translation. LIDIA does, however, demonstrate the sort of interaction likely to be needed for on-line translation where quality is important, and adaptation for real-time use may be possible. (See further below.)

4. Interactive Disambiguation for Speech Recognition

I have argued the near-term advantages of interactive disambiguation for speech translation. However, many questions remain about the style of interaction, the points of interaction, and so on.

For example, it is unlikely that the procedure described above for ASURA, selection from an n-best list for the entire utterance, is at all optimal from an ergonomic point of view. The user had to choose the right candidate from among ten or so similar competitors. While this style of choice was manageable in a carefully planned demo, it would have been tiring in a more spontaneous dialogue.

An improved interface for a connected speech recognizer might collapse the common elements of several candidates, thus presenting a word or phrase lattice to the user, who might then indicate the correct path using a pointing device. Or it might present only the best candidate, and provide a means to quickly correct erroneous portions.

But there is another approach to selection of speech recognition candidates which is particularly interactive, since it intervenes particularly early and particularly often to nip uncertainty and error in the bud before they can propagate. This is the use of isolated-word speech recognition, as seen in several dictation programs currently available on the market. When dictating word by word, the user corrects each word before going on to the next one. The prefix word string is thus known to be correct, and the language model can take full advantage of this knowledge in guessing what comes next. Further, the boundaries of the current word are known exactly, and this knowledge, too, greatly constrains the recognition task. When ambiguity remains, the dictation program presents a menu of candidates for the next word, and the user can choose the right one, using a pointing device or saying "Choose one" or "Choose two".

The disadvantages of isolated-word speech input are obvious enough: this entry mode is much slower than natural speech, and considerably less natural, so that some learning is required. However, the advantages of this input mode are also rather striking, and may cover a multitude of sins at the current stage of research.

The first advantage has already been stressed: reliability or robustness. Good results can now be obtained even for user-independent dictation systems, those requiring no registration for new users.

Perhaps more surprising, though, is the breadth of coverage which is immediately available — the capacity to recognize words across many domains, even when they are quite rare. I was impressed, for instance, by one dictation system's ability to recognize my rendering of Abraham Lincoln's Gettysburg Address: infrequent words like continent, conceived, liberty, dedicated, proposition, etc. were correctly recognized immediately and without ambiguity, although there had been no prior registration of my voice. To a researcher accustomed to the need for extensive domain-specific training in order to recognize common words in connected speech using lexicons of a few hundred words, this success was striking indeed. But then, when word boundaries are known, the frequency of the target word and the size of the lexicon become less important. What matters much more under these conditions is confusability with similar words. A word which sounds like no other word in the lexicon can be recognized easily, even if it is quite arcane, and even if the lexicon is quite large. (The word "arcane", for instance, would probably cause little difficulty for the dictation program under discussion.)

Since it would not require users to remain within a specified and extensively trained domain of discourse, a speech translation system based on isolated word dictation could potentially be used for many purposes with a minimum of domain adaptation for speech recognition. (Of course, the need to customize the translation component would remain as an obstacle to completely unrestricted use.)

The advantages of fully disambiguated speech for a speech translation system have already been stressed: barring editing errors, the segmental information passed to the translation component is guaranteed to be correct. But if the speech recognition interface employs isolated-word entry in particular, there are several additional advantages from the translation component's viewpoint. Dictated entry is likely to eliminate many of the disfluencies of spontaneous speech, since it forces users to speak relatively slowly and carefully and gives continuing opportunities for revision and editing. Thus the syntax of whole utterances is likely to be much cleaner than in spontaneous speech, and accordingly easier to analyze. Further, since dictation is slower than connected speech, whole utterances will probably be shorter on the average. Overall, since input will be (more nearly) correct, shorter, and syntactically simpler, in all likelihood less interaction will be needed in the MT stage to produce sufficient translations.

5. Architectures for Dialogue MT on the Internet

Having argued for the extensive use of interactive disambiguation in building "quick and dirty" dialogue translation systems for near-term use on the Internet, we come now to the question of architecture.

The Information Transcript Project

One "quick and dirty" speech translation system for the Internet has already been demonstrated. This is the MIT-Lyon Information Transcript Project (Kowalski, Rosenberg, & Krause 1995). (See also *http://sap.mit.edu/projects/mit-lyon/project.html.*). The project provides a proof of concept, and shows what can be done with existing commercial components.

Sites at MIT and at the Biennale d'Art Contemporain in Lyon, France were linked for four hours a day for several days between December 20, 1995 and February 18, 1996. Visitors at each site dictated in their own languages to IBM's Voice Type isolated-word speech recognition software. The recognized source-language text files were

transmitted transatlantically by ftp, via a dedicated page on the World Wide Web. Meanwhile, observers of the page could see a transcript of the multi-party dialogue. (Hence the project's name.) The Web observers could also type into a window on the page; the resulting text was treated as source text to be translated, just as if it had been dictated. At the receiving end, the source-language text, whether dictated or typed, was translated via Global-Link commercial translation software. English target text was spoken by standard Macintosh speech synthesis software. French target text was pronounced by a dedicated device. To complete the conference call effect, video and the original audio were also transmitted via CU-SeeMe.

The aim was largely artistic and inspirational: "...the intention ... is to reveal some hidden aspects of communication at the time of its globalization." (Web site, above) Thus translation quality was not crucial. All the same, the project demonstrates that technology is now widely available which permits the creation of "quick and dirty" speech translation systems, in which interactive disambiguation is exploited to make system integration less necessary in the interest of getting something usable up and running quickly.

Some further details of the Information Transcript system may be appreciated. It involved nine machines in four locations, and permitted:

(1) *Translation between English and French, in both directions, of text input from any source.* Translation was handled on a pair of Macintoshes via Apple Script. In the script, a loop checked periodically for the presence of a text input file. When input text was found, it was translated using Global-Link commercial software, and the translated text was sent overseas via ftp. (As the ftp connection was handled through a Web server in Miami, text traveling in both directions could be captured in order to build up a transcript, visible via a Web page.) On arrival, translated text was deposited in a text file to await pronunciation.

(2) *Pronunciation of text.* On the English-French translation Macintosh, a script for SimpleText received English text in a buffer, selected the text, and ran Speak Text to pronounce it. On the French-English translation Macintosh, no such software was available; so a separate French speech synthesis device was attached to the Macintosh's serial out. In either case, the text to be pronounced might be the output of translation arriving from overseas, or it might be text typed to a Web page and forwarded by the Web server.

(3) *Speech recognition in English and French.* Speech recognition was carried out by IBM's Voice Type isolated-word dictation software, running on OS/2 on IBM machines (one for English, one for French). Perl scripts handled input of speech and output of text. Output was sent via ftp to the translation Macintoshes for translation.

(4) *Typed participation in the translated dialogue via the World Wide Web.* As mentioned, a Web server in Miami handled the ftp traffic between translating Macintoshes (via a Common Gateway Interface script) and captured a transcript of text going both ways. The Web server also maintained a Web page which could accept typed input. Viewers of the page could thus participate in the interchange by typing text to the page, right along with participants using speech recognition: the translating Macintoshes didn't care whether the text they translated came through the Web page or via speech recognition.

(5) *A 2-way video hookup.* Speakers talking to the speech recognition software were on camera. Two Macintoshes running Player, in addition to those used for translation, were used. A CU-SeeMe connection was established via an (unsuspecting) reflector, or relay, in New York City.

Other Architectures

The Information Transcript project exploited the World Wide Web and ftp for transmission of files. It also distributed processing across several machines, with one running SR, one running MT, etc. We can take a moment to consider other possibilities.

Where transmission is concerned, another natural direction would be to use Internet Relay Chat, or IRC. Chat requires an IRC server which links numerous users running client software. A user, who may be anywhere in the world, selects a Chat channel from among hundreds. He or she types into a special input window, and hits Return when his/her utterance is finished. The finished utterance instantly appears in a larger window visible to all who have selected the same channel. This common window thus comes to resemble the transcript of a cocktail party, with numerous conversations interleaved. Any user can easily record all of the utterances in a session from his or her vantage point.

Chatters can use prefixes or other tricks to direct conversation to particular fellow chatters. Some current Chat client software even includes voice synthesis, so that a given user not only sees but hears the incoming utterances — differentiated, if desired, by different synthesized voices. By mutual agreement, chatters can "move" into "private rooms" to hide their words from

others. It is also straightforward to establish a new channel, in which the establisher controls participation.

With respect to distribution, one could assemble all of the processing for a single conversant on a single sufficiently roomy and fast computer; or a client/server arrangement could be arranged, with translation or speech recognition software on a large server accessed over the Internet. The first plan has simplicity to recommend it; the second allows for much greater computational power, but incurs the difficulties of more complex communication protocols and network delays and breakdowns.

Possible combinations are many. Combining the IRC and all-in-one-machine options, we could arrive at a the following particularly simple setup for experiments in interactive translation: interactive MT running on a PC or workstation would serve as a front end for IRC — a user would first interactively disambiguate translations in the source language, and when satisfied would press return or give an equivalent verbal signal; the MT component would in turn pass its output, the translated text string, to Chat for transmission. Experiments could be conducted with or without interactive speech recognition (for greatest simplicity, also on the same PC or workstation as the translation program) as an additional front-end layer; and with or without speech synthesis (ditto) as a back end.

Possible components for "quick and dirty" systems are many, too. Now available commercially or as shareware are a variety of isolated-word dictation systems; text translation systems which are sufficiently fast (though quality is likely to be fair to poor); operating systems supporting adequate scripting languages; IRC clients with sufficiently open architectures to permit the easy addition of a front end; etc.

Two desirable components for "quick and dirty" MT or speech translation systems are still difficult to obtain, however.

First, interactive disambiguation tools for MT are generally unavailable. The LIDIA-1.2 system described in Section 3, while having many of the capabilities needed for dialogue disambiguation, presently accepts input only from the Ariane-G system; and this system, developed for off-line text translation, has until now been too slow for real-time use. To use LIDIA-1.2 for dialogue translation, one might modify other translation systems to deliver output in the proper format; or one might speed up the original translation system, either by recoding or by running on a very powerful machine. Both options are now under consideration.

A second difficulty is related: that of finding an appropriate translation server if a client/server architecture is desired. Early efforts to use Ariane-G as a translation server are reported by (LaFourcade 1996). Slow but robust communication has been established by email, and faster but less reliable connections have been arranged via HTTP, the protocol used in the World Wide Web.

Conclusions

Now that inexpensive real-time transmission of text over the Internet is a reality and a range of speech recognition and translation components adaptable for interactive use are becoming available, it has become practical to build and experiment with "quick and dirty" dialogue translation systems, including speech translation systems. I have suggested that in order to take the "low road" to systems usable in the near term, such systems should substitute interactive disambiguation for tight integration among components. However, maximum automaticity via tight system integration remains the long-term aim.

I have argued the advantages of interactivity for near-term MT in general: as long as many translation ambiguities escape automatic resolution, interactive disambiguation will remain important. Disambiguation, when needed, should be done as early as possible. In speech translation systems, early disambiguation of speech recognition candidates will be especially worthwhile. Viewed in this light, isolated-word speech recognition has a number of unexpected advantages as an especially interactive input mode.

Procedures for interactive MT and SR have been discussed, and a working demo of a "quick and dirty" speech translation system, implemented by Kowalski, Rosenberg, and others, has been reported. Possibilities for alternative designs have also been briefly presented.

While the Information Transcript project demonstrates the technical feasibility of "quick and dirty" systems for machine translation — even speech translation — over the Internet, many questions remain about the practicality of such systems. While this project's use of dictated input effectively supplied interactive disambiguation for speech recognition, there was no interactive disambiguation of translation, and in fact the translation software provided no such capability. The resulting translation quality was uninspiring (Burton Rosenberg, personal communication), but given the principally social and artistic aims of the project, this was no great cause for concern.

For the future, experiments should stress questions of usability: For specific communication tasks, what is the most practical architecture? For given architectural candidates, how difficult will it prove to add interactive disambiguation capabilities for MT? How much interaction is tolerable? How much is needed for what degree of quality, and how much quality is enough?

While these questions remain open, it is clear that "quick and dirty" systems for Internet translation are beginning to be built. If natural language researchers neglect the low road to dialogue translation offered by such systems, great opportunities for study could be missed. And there is the danger that we will arrive in Scotland to find other enterprising folk already there.

Acknowledgements

Warm thanks to Burton Rosenberg for hospitality and discussion of the Information Transcript project.

References

Boitet, Christian. 1996. "Dialogue-based Machine Translation for Monolinguals and Future Self-explaining Documents". In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*, Col de Porte, France, August 11 - 15, 1996.

Blanchon, Hervé. 1996. "A Customizable Interactive Disambiguation Methodology and Two Implementations to Disambiguate French and English Input." In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*, Col de Porte, France, August 11 - 15, 1996.

Kowalski, Piotr, Burton Rosenberg, and Jeffery Krause. 1995. *Information Transcript.* Biennale de Lyon d'Art Contemporain. December 20, 1995 to February 18, 1996. Lyon, France.

LaFourcade, Mathieu. 1996. "Software Engineering in the LIDIA Project: Distribution of Interactive Disambiguation and Other Components between Various Processes and Machines." In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*, Col de Porte, France, August 11 - 15, 1996.

Morimoto, Tsuyoshi, and T. Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata, et al. 1993. "ATR's Speech Translation System: ASURA." In *Proceedings of Eurospeech-93*. Berlin. Sept. 21-23, 1993.

Seligman, Mark and Christian Boitet. 1993. A "Whiteboard" Architecture for Automatic Speech Translation." In *Proceedings of ISSD-93 (International Symposium on Spoken Dialogue)*, Waseda University, Tokyo, November 10 - 12, 1993.

Wahlster, Wolfgang. 1993. "Verbmobil: Translation of Face-to-face Dialogues." In *Proceedings of the European Conference on Speech Communication and Technology*, Vol. "Opening and Plenary Sessions", pages 29-38, 1993.