Practical Speech Translation Systems will Integrate Human Expertise, Multimodal Communication, and Interactive Disambiguation

Ch. Boitet[1]
ATR Interpreting Telecommunications Research Laboratories
Hikari-dai 2-2, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
boitet@itl.atr.co.jp

Contribution to the panel on "Future MT Technology"
Machine Translation Summit IV, Kobe, 19 - 22 July 1993

Summary

It has always been remarkably difficult to build really practical Machine Translation (MT) and Speech Processing (SP) systems. As Speech Translation (ST) combines the difficulties of both endeavours, it should come as no surprise that the first prototypes, although well-researched and brilliantly demonstrated, cannot be extended towards practical systems.

Dramatic progress in both MT and SP technology is not being likely to be witnessed in the near future. Besides necessary but inherently limited improvements in the component technologies, the construction of practical ST systems will require better user-friendliness, achievable through the introduction of a human expert (interpreter), multimodal communication facilities between the expert and the speakers, and various control and feed-back facilities.

Because of the quality and coverage required of the Speech Recognition (SR) and Natural Language Analysis (NLA) components in realistic applications and their inherent difficulty, however, it will also be necessary to involve the end users (the speakers) in these processes, by encouraging them to control their own voice, and asking them to help through multimodal active and passive disambiguation.

Introduction

Research in Speech Translation has been initiated by ATR in Japan at the beginning of 1986. Less than seven years later, it has been possible to give brilliant demonstrations, with good mediatic success. The goal was to perform consecutive interpretation of bilingual conversations (Japanese-English-German) prepared (and read) between the secretary and a participant of a hypothetical international conference. Interlocutors spoke from Kyoto (ATR), Pittsburgh (CMU), and Munich (Siemens) using normal international telephone. Speech analysis and translation were performed at the emitting site and speech synthesis at the receiving site. Visual feed-back through video communication proved ergonomically very useful.

Applications envisaged for Speech Translation include assistance to professional or personal telephone dialogues (car rental, medical consultation, scheduling of meetings, greetings, explanation of itinerary ...), teleconference, and multilingual dissemination of information.

[1] Visiting researcher from GETA, IMAG, UJF&CNRS, France. This paper has benefitted from the constant support of Dr. A. Kurematsu and Dr. Y. Yamazaki, presidents of ATR-ITL, and from fruitful discussions with K. H. Loken-Kim, T. Morimoto, M. Seligman, H. Singer, T. Tashiro, M. Tomokiyo, N. Uratani, and F. Yato.

The current black box sequential speech-only architecture cannot be extended towards practical systems.

There are no miracles: the present prototype [3] has not been developed with practical usability in mind, and cannot provide it. Recall that it operates by successively executing 5 basic components (SR, analysis, transfer, generation, speech synthesis), each fully automatic. Its most notable positive points are the original SR part, the new treatment of some important communicative aspects (politeness, honorifics, illocutionary force type), and the implementation of a full setup for transcontinental demonstrations, including video for passive control. Moreover, the study and separate prototyping of discourse & dialogue structure recognition are promising.

However, this architecture is inherently too limited for practical systems: (1) the MT part is too slow to be used in real communication, and its architecture (unification-based, lexicon not distinguished from grammar) makes it impractical and perhaps intractable (because of at least exponential growth in time complexity) to extend it to larger sublanguages, (2) the coverage is too small for any realistic application, as the MT part comprises about 500 lexical elements, and rules handling only the structures of the 250 sentences of the test corpus, and (3) the all-or-nothing approach in NLA makes it impossible to degrade gracefully in presence of an error or of an unknown (or misrecognized) word, so that the user interface cannot be sufficiently user-friendly.

It would be possible to improve the current black box speech-only sequential architecture, speeding it up to quasi real-time by modifying the MT engine and using mixed computational strategies (exactly which techniques to choose in the existing proven repertory is a matter of taste), scaling up to the size required by the most restricted but still realistic applications (≈ 3000 terms/4000 wordforms in English), and introducing simple user control facilities. However, that would not suffice to produce acceptable systems, because all envisageable applications would require far more robustness of their SR and NLA components (hesitations, errors, self-corrections...), and impose more constraints on their SR component (noisy environment, variety of speakers & accents...).

Introduction of an expert (interpreter) and of support for multimodal HMH interaction and system control will be necessary, but not sufficient.

Integrating a human interpreter ("warm body") in the overall architecture would guarantee that the system works (because the human can do all the work if the system fails completely!), and provide a smooth transition in existing operational environments. This is certainly an essential factor in future practical systems.

Also, *providing support for multimodal Human-Machine-Human (HMH) interaction and system control* would make it reasonably user-friendly [6]. For example, it should be possible to tune parameters controlling the perception of other agents (interlocutor, system, interpreter...), to monitor the progression of the translation process, and even to interrupt it (because the meaning has already been understood, or to correct the previous utterance), thereby reducing waiting time and associated frustration.

This last point suggests the interesting possibility of building *"progressive" MT systems,* which would output successive states of the translation (on appropriate media), beginning with isolated words, then phrases, then complete raw translation, to finish with polished translation if the speakers help the system through interactive disambiguation (see below).

Equipping the system with knowledge about the generic task (partial ontology) is also a possibility, but is likely to be an overshot, because the speakers will in any case be far better at understanding the task at hand. But, even with a complete ontology, an automatic system can not fully disambiguate clean, typed sentences (this is why interactive disambiguation through the "augmentor" was introduced in KBMT-89 [5]).

Hence, if the SR and NLA components function as black boxes, there is a high risk that the proportion of utterances successfully handled automatically will remain quite small, making the resulting system unacceptable by the users, the interpreter, and the investors alike.

End users should help the system in its internal working.

First, users should be encouraged to *make speech recognition easier* by controlling their own voice. The SR component could indicate (visually or acoustically) its level of difficulty of recognition, perhaps with an appropriate diagnostic (e.g. too quick, too slurred...). A very important feature would be to propose ways to "clean" the input, perhaps by editing its written form.

Second, users could *guide the system through "active" (user-initiated) multimodal disambiguation.* One example is to press a button to indicate the end of a sentence within a speech period. Another is to navigate through a graphic representation of the task domain while speaking, in order to dynamically restrict the expected vocabulary. It would also be possible to indicate the communicative type of the current utterance (assertion, question, request, advice...) to facilitate semantic and pragmatic interpretation.

Third, *"passive" (system-initiated) disambiguation* could be used. The user would be asked questions, in various possible ways [1]. Previous studies [6] report that up to 33% of utterances are of that type in bilingual telephone conversations (using a human interpreter), so that questions spoken by the system are a distinct possibility. In a multimodal context, it would also be possible to ask the user to select items in menus, or even to correct an intuitive graphic representation of the utterance by direct manipulation.

Conclusion

The construction of practical Speech Translation systems will require more consideration of human factors at the external level. Introduction of a human expert ("warm body") and of multimodal facilities for system control and HMH interaction is necessary, but not sufficient. Involvement of the end users at the internal level, in the translation process itself, through speech control and interactive disambiguation, will be required. Previous work on integration of multimodality [4] and interactive disambiguation [2] in NLP are encouraging as far as the feasibility of the approach outlined here in the middle term future is concerned.

These technological requirements trigger numerous interesting research problems, most notably (1) the construction of "progressive" MT systems, which would be a first step towards simultaneous MT (an ambitious research theme proposed by T. Morimoto), (2) the integration of several modalities in MT systems and NLP systems in general, (3) the production of diagnostics by the SR component, and (4) the search of adequate computational methods for Multimodal Interactive Disambiguation.

References

[1] **Boitet C. (1989)** *Speech Synthesis and Dialogue Based Machine Translation.* Proc. ATR Symp. on Basic Research for Telephone Interpretation, Kyoto, December 1989,6-5-1-6-5-22.

[2] **Maruyama H., Watanabe H. & Ogino S. (1990)** *An Interactive Japanese Parser for Machine Translation.* Proc. COLING-90, Helsinki, 20-25/8/90, H. Karlgren, ed., ACL, vol. 2/3, 257-262.

[3] **Morimoto T., Suzuki M., Takezawa T., Kikui G.-L, Nagata M. & Tomokiyo M. (1992)** *A Spoken Language Translation System: SL-TRANS2.* Proc. COLING-92, ACL, vol. 3/4,1048-1052.

[4] **Neal J. G. & Shapiro S. C. (1991)** *Intelligent Multimedia Interface Technology.* In "Intelligent User Interfaces", ACM Press & Addison-Wesley, New-York, 11-44.

[5] **Nirenburg S & al. (1989)** *KBMT-89 Project Report.* CMT, CMU, Pittsburg, April 1989, 286 p.

[6] **Oviatt S. L. (1993)** *Toward multimodal support for interpreted telephone dialogues.* In "Structure of Multimodal Dialogue", M. M. Taylor, F. Néel & D. G. Bouwhuis, ed., Elsevier, Amsterdam, in press.

[7] **Oviatt S. L. & Cohen P. R. (1991)** *Discourse structure and performance efficiency in interactive and noninteractive spoken modalities.* Comp. Speech & Lang., 5/4, 297-326.

[8] **Sullivan J. W. & Tyler S. W., ed. (1991)** *Intelligent User Interfaces.* Addison-Wesley, N. Y., 472 p.