Six Issues in Speech Translation

Mark Seligman
Université Joseph Fourier
GETA, CLIPS, IMAG-campus, BP 53
150, rue de la Chimie
38041 Grenoble Cedex 9, France
seligman@cerf.net

Abstract

This position paper sketches the author's research in six areas related to speech translation: interactive disambiguation; system architecture; the interface between speech recognition and analysis; the use of natural pauses for segmenting utterances; dialogue acts; and the tracking of lexical co-occurrences.

Introduction

This position paper reviews some aspects of my research in speech translation since 1992. Since the purpose is to prompt discussion, the treatment is informal and speculative, with frequent reference to work in progress.

The paper sketches work in six areas: interactive disambiguation; system architecture; the interface between speech recognition and analysis; the use of natural pauses for segmenting utterances; dialogue acts; and the tracking of lexical co-occurrences. There is no attempt to provide a balanced survey of the speech translation scene. However, I have touched upon a number of research areas which seem to me of particular interest.

1 Interactive Disambiguation

At the present state of the art, several stages of speech translation leave ambiguities which current techniques cannot yet resolve correctly and automatically. Such residual ambiguity plagues speech recognition, analysis, transfer, and generation alike.

Since users generally can resolve these ambiguities, it seems reasonable to incorporate facilities for interactive disambiguation into speech translation systems, especially those aiming for broad coverage. A good idea of the range of work in this area can be gained by browsing the proceedings of MIDDIM-96 (the International Seminar on Multimodal Interactive Disambiguation, Col de Porte, France, August 11 - 15, 1996).

In fact, (Seligman 1997) suggests that, by stressing such interactive disambiguation — for instance, by using isolated-word dictation rather than connected speech for input, and by adapting existing techniques for interactive disambiguation of text translation (Boitet 1996, Blanchon 1996) — practically usable speech translation systems may be constructable in the near term. The paper also reports on an Internet-based demo along these lines (Kowalski, Rosenberg, and Krause 1995). In such "quick and dirty" or "low road" speech translation systems, user interaction is substituted for system integration. For example, the interface between speech recognition and analysis can be supplied entirely by the user, who can correct SR results before passing them to translation components, thus bypassing any attempt at effective communication or feedback between SR and MT.

The argument, however, is not that the "high road" toward integrated and maximally automatic systems should be abandoned. Rather, it is that the "low road" of forgoing integration and embracing interaction may offer the quickest route to widespread usability, and that experience with real use is vital for progress. Clearly, the "high road" is the most desirable for the longer term: integration of knowledge sources is a fundamental issue for both cognitive and computer science, and maximally automatic use is intrinsically desirable. The suggestion, then, is that the low and high roads be traveled in tandem; and that even systems aiming for full automaticity recognize the need for interactive resolution when automatic resolution is insufficient.

2 System Architecture

An ideal architecture for "high road" or integration-intensive speech translation systems would allow global coordination of, cooperation between, and feedback among, components (speech recognition, analysis, transfer, etc.), thus moving away from linear or pipeline arrangements. For instance, speech recognition, as it moves through an utterance, should be able to benefit from preliminary analysis results for segments earlier in the utterance. The architecture should also be modular, so that a variety of configurations can be tried: it should be possible, for instance, to exchange competing speech recognition components; and it should be possible to combine components not explicitly intended for work together, even if these are written in different languages or running on different machines.

Blackboard architectures have been proposed (Erman and Lesser, 1980) to permit cooperation among components. In such systems, all participating components read from and write to a central set of data structures — the blackboard. To share this common area, however, the components must all "speak a common (software) language". Modularity thus suffers, since it is difficult to assemble a system from components developed separately. Further, blackboard systems are widely seen as difficult to debug, since control is typically distributed, with each component determining independently when to act and what actions to take.

In order to maintain the cooperative benefits of a blackboard system while enhancing modularity and facilitating central coordination or control of components, (Seligman and Boitet 1994 and Boitet and Seligman 1994) proposed and demonstrated a "whiteboard" architecture for speech translation. As in the blackboard architecture, a central data structure is maintained which contains selected results of all components. However, the components do not access this "whiteboard" directly. Instead, only a privileged program called the Coordinator can read from it and write to it. Each component communicates with the Coordinator and the whiteboard via a go-between program called a *manager*, which handles messages to and from the Coordinator in a set of mailbox files. Because files are used as data holding areas in this way, components (and their managers) can be freely distributed across many machines. Managers are not only mailmen, but interpreters: they translate between the reserved language of the whiteboard and the native languages of the components, which are thus free to differ. In our demo, the whiteboard was maintained in a commercial Lisp-based object-oriented language, while components included independently-developed speech recognition, analysis, and word-lookup components written in C. Overall, the whiteboard architecture can be seen as an adaptation of blackboard architectures for client-server operations: the Coordinator becomes the main client for several components behaving as servers.

Since the Coordinator surveys the whiteboard, in which are assembled the selected results of all components, all represented in a single software interlingua, it is indeed well situated to provide central or global coordination. However, any degree of distributed control can also be achieved by providing appropriate programs alongside the Coordinator which represent the components from the whiteboard side. That is, to dilute the Coordinator's omnipotence, a number of demi-gods can be created. In one possible partly-distributed control structure, the Coordinator would oversee a set of agendas, one or more for each component.

A closely-related effort to create a modular "agent-based" (client-server-style) architecture with a central data structure, usable for many sorts of systems including speech translation, is described in (Julia et al 1997). Lacking a central board but still aiming in a similar spirit for modularity in various sorts of translation applications is the project described in (Zajac and Casper 1997).

3 Interface between Speech Recognition and MT Analysis

In a certain sense, speech recognition and analysis for MT are comparable problems. Both require the recognition of the most probable sequences of elements. In speech recognition, sequences of short speech segments must be recognized as phones, and sequences of phones must be recognized as words. In analysis, sequences of words must be recognized as phrases, sentences, and utterances.

Despite this similarity, current speech translation systems use quite different techniques for phone, word, and syntactic recognition. Phone recognition is generally handled using hidden Markov models (HMMs); word recognition is often handled using Viturbi-style search for the best paths in phone lattices; and sentence recognition is handled through a variety of parsing techniques.

It can be argued that these differences are justified by differences of scale, perplexity, and meaningfulness. On the other hand, they introduce the need for interfaces between processing levels. The processors may thus become black boxes to each other, when seamless connection and easy communication might well be preferable. In particular, word recognition and syntactic analysis (of phrases, sentences, and utterances) should have a lot to say to each other: the probability of a word should depend on its place in the top-down context of surrounding words, just as the probability of a phrase or larger syntactic unit should depend on the bottom-up information of the words which it contains.

To integrate speech recognition and analysis more tightly, it is possible to employ a single grammar for both processes, one whose terminals are phones and whose non-terminals are words, phrases, sentences, etc.[1] This phone-grounded strategy was used to good effect e.g. in the HMM-LR speech recognition component of the ASURA speech translation system (Morimotô et al 1993), in which an LR parser extended a parse phone by phone left to right while building a full syntactic tree.[2] The technique worked well for scripted examples. For spontaneous examples, however, performance was unsatisfactory, because of the gaps, repairs, and other noise common in spontaneous speech. To deal with such structural problems, an island-driven parsing style might well be preferable. An island-based chart parser, like that of (Stock et al 1989), would be a good candidate.

However, chart initialization presents some technical problems. There is no difficulty in computing a lattice from spotted phones, given information regarding the maximum gap and overlap of phones. But it is not trivial to convert that lattice into a "chart" (i.e. multi-path finite state automaton) without introducing spurious extra paths. The author has implemented a Common Lisp program which does so correctly, based on an algorithm by Christian Boitet. Experiments with bottom-up island-driven chart parsing from charts initialized with phones are anticipated.

4 Use of Pauses for Segmentation

It is widely believed that prosody can prove crucial for speech recognition and analysis of spontaneous speech, but effective demonstrations have been few. Several aspects of prosody might be exploited: pitch contours, rhythm, volume modulation, etc. However, (Seligman, Hosaka, and Singer 1996) propose focusing on natural pauses as an aspect of prosody which is both important and relatively easy to detect automatically.

Given the frequency of utterances in spontaneous speech which are not fully well-formed — which contain repairs, hesitations, and fragments — strategies for dividing and conquering utterances would be quite useful. The suggestion is that natural pauses can play a part in such a strategy: that *pause units*, or segments within utterances bounded by natural pauses, can provide chunks which (1) are reliably shorter and less variable in length than entire utterances and (2) are relatively well-behaved internally from the syntactic viewpoint, though analysis of the relationships among them appears more problematic.

Four specific questions are addressed: (1) Are pause units reliably shorter than whole utterances? If they were not, they could hardly be useful in simplifying analysis. It was found however, that in the corpus investigated (Loken-Kim, Yato, Kurihara, Fais, and Furukawa 1993; Furukawa, Yato, and Loken-Kim 1993), pause units are in fact about 60% the length of entire utterances, on the average, when measured in Japanese morphemes. The average length of pause units was 5.89 morphemes, as compared with 9.39 for whole utterances. Further, pause units are less variable in length than entire utterances: the standard deviation is 5.79 as compared with 12.97. (2) Would hesitations give even shorter, and thus perhaps even more manageable, segments if used as alternate or additional boundaries? The answer seems to be that because hesitations so often coincide with pause boundaries, the segments they mark out are nearly the same as the segments marked by pauses alone. No combination of expressions was found which gave segments as much as one morpheme shorter than pause units on average. (3) Is the syntax within pause units relatively manageable? A manual survey showed that, once hesitation expressions are filtered from them, some 90% of the pause units studied can be parsed using standard Japanese grammars; a variety of special problems appear in the remaining 10%. (4) Is translation of isolated pause units a possibility? We found that a majority of the pause units in four dialogues gave understandable translations into English when translated by hand.

The study provided encouragement for a "divide and conquer" analysis strategy, in which parsing and perhaps translation of pause units is carried out before, or even without, attempts to create coherent analyses of entire utterances.

As mentioned, parsability of spontaneous utterances can be enhanced by filtering hesitation expressions from them in preprocessing. Research on spotting techniques for such expressions would thus seem to be worthwhile. Researchers can exploit speakers' tendency to lengthen hesitations, and to use them just before or after natural pauses.

---

[1] Inclusion of other levels is also possible. At the lower limit, assuming the grammar were stochastic, one could even use sub-phone speech segments as grammar terminals, thus subsuming even HMM-based phone recognition in the parsing regime. At an intermediate level between phones and words, syllables could be used.

[2] The parse tree was not used for analysis, however. Instead, it was discarded, and a unification-based parser began a new parse for MT purposes on a text string passed from speech recognition.

5 Communicative Acts

Speech act analysis (Searle 1969) — analysis in terms of illocutionary acts like INFORM, WH-QUESTION, REQUEST, etc. — can be useful for speech translation in numerous ways. Six uses, three related to translation and three to speech processing, will be mentioned here. Concerning translation, it is necessary to:

- *Identify the speech acts of the current utterance.* Speech act analysis of the current utterance is necessary for translation. For instance, the English pattern "can you (VP, bare infinitive)?" may express either an ACTION-REQUEST or a YN-QUESTION (yes/no-question). Resolution of this ambiguity will be crucial for translation.
- *Identify related utterances.* Utterances in dialogues are often closely related: for instance, one utterance may be a prompt and another utterance may be its response; and the proper translation of a response often depends on identification and analysis of its prompt. For example, Japanese *hai* can be translated as *yes* if it is the response to a YN-QUESTION, but as *all right* if it is the response to an ACTION-REQUEST. Further, the syntax of a prompt may become a factor in the final translation. Thus, in a responding utterance *hai, sou desu* (meaning literally "yes, that's right"), the segment *sou desu* may be most naturally translated as *he can*, *you will*, *she does*, etc., depending on the structure and content of the prompting question. The recognition of such prompt-response relationships will require analysis of typical speech act sequences.
- *Analyze relationships among segments and fragments.* Early processing of utterances may yield fragments which must later be assembled to form the global interpretation for an utterance. Speech act sequence analysis should help fit fragments together, since we hope to learn about typical act groupings.

Concerning speech processing, it is necessary to:

- *Predict speech acts to aid speech recognition.* If we can predict the coming speech acts, we can partly predict their surface patterns. This prediction can be used to constrain speech recognition. For example, in recognizing spoken Japanese, if we can predict the relative probability that the current utterance is a YN-QUESTION as opposed to an INFORM, we may be able to differentiate utterance-final *ka* (a question particle) and utterance-final *ga* (a conjunction or politeness particle), which are often very similar phonetically.
- *Provide conventions for prosody recognition.* Once spontaneous data is labeled, speech recognition researchers can try to recognize prosodic cues to aid in speech act recognition and disambiguation. For instance, they can try to distinguish segments expressing INFORMs and YN-QUESTIONs according to the F0 curves associated with them — a distinction which would be especially useful for recognizing YN-QUESTIONs with no morphological or syntactic markings.
- *Provide conventions for speech synthesis.* Similarly, speech synthesis researchers can try to provide more natural prosody by exploiting speech act information. Once relations between prosody and speech acts have been extracted from corpora labeled with speech act information, researchers can attempt to supply natural prosody for synthesized utterances according to the specified speech acts. For instance, more natural pronunciations can be attempted for YN-QUESTIONs, or for CONFIRMATION-QUESTIONs (including tag questions in English, as in *The train goes east, doesn't it?*).

While a well-founded set of speech act labels would be useful, it has not been clear what the theoretical foundation should be. As a result, no speech act set has yet become standard. Labels are proposed intuitively or by trial and error.

Speakers' goals can certainly be analyzed in many ways. However, (Seligman, Fais, and Tomokiyo 1995) hypothesize that only a limited set of goals is conventionally expressed in a given language. For just these goals, relatively fixed expressive patterns are learned by speakers when they learn the language. In English, for instance, it is conventional to express certain invitations using the patterns "Lets *" or "Shall we *?" In Japanese, one conventionally expresses similar goals via the patterns "(V, combining stem)mashou" or "(V, combining stem)masen ka?"

The proposal is to focus on discovery and exploitation of these conventionally-expressible speech acts, or *Communicative Acts*. The relevant expressive patterns and the contexts within which they are found have the great virtue of being objectively observable; and assuming the use of these patterns is common to all native speakers, it should be possible to reach a consensus classification of the patterns according to their contextualized meaning and use. This functional classification should yield a set of language-specific speech act labels which can help to put speech act analysis for speech translation on a firmer foundation.

The first reason to analyze speech acts in terms of observable linguistic patterns, then, is the measure of objectivity thus gained: the discovery process is to some degree empirical, data-driven, or corpus-based. A second reason is that on-line analysis, being shallow or surface-bound, should be relatively quick as opposed to plan-based analysis. Plan-based analysis may well prove necessary for certain purposes, but it is quite expensive. For applications like speech translation which must be carried out in nearly real time, it seems wise to exploit shallow analysis as far as possible.

With these advantages of cue-based processing — empirical grounding and speed — come certain limitations. When analyzing in terms of CAs, we cannot expect to recognize all communicative goals. Instead, we restrict our attention to communicative goals which can be expressed using conventional linguistic cue patterns. Communicative goals which cannot be described as Communicative Acts include utterance goals which are expressed non-conventionally (compare the non-conventional warning *May I call your attention to a potentially dangerous dog* to the conventional WARNING *Look out for the dog!*); or goals which are expressed only implicitly (*It's cold outside* as an implicit request to shut the window); or goals which can only be defined in terms of relations between utterances. (While speakers often repeat an interlocutor's utterance to confirm it, we do not use a REPEAT-TO-CONFIRM CA, since it is apparently signaled by no cue patterns, and thus could only be recognized by noting inter-utterance repetition.)

Given that the aim is to classify expressive patterns according to their meaning and function, how should this be done? The paper describes a paraphrase-based approach: native speakers are polled as to the essential equivalence of expressive patterns in specified discourse contexts. If by consensus several patterns can yield paraphrases which are judged equivalent in context, and if the resulting pattern set is not identical to any competing pattern set, then it can be considered to define a Communicative Act.

Communicative Acts are defined in terms of monolingual conventions for expressing certain communicative goals using certain cue patterns. For translation purposes, however, it will be necessary to compare the conventions in language A with those in language B. With this goal in mind, the discovery procedure was applied to twin corpora of Japanese-Japanese and English-English spontaneous dialogues concerning transportation directions and hotel accommodations (Loken-Kim et al. 1993). CAs were first identified according to monolingual criteria. Then, by observing translation relations among the English and Japanese cue patterns, the resulting English and Japanese CAs were compared. Interestingly, it was found that most of the proposed CAs seem valid for both English and Japanese: only two out of 27 CAs seem to be monolingual for the corpus in question.

6 Tracking Lexical Co-occurrences

In the processing of spontaneous language, the need for predictions at the morphological or lexical level is clear. For bottom-up parsing based on phones or syllables, the number of lexical candidates is explosive. It is crucial to predict which morphological or lexical items are likely so that candidates can be weighted appropriately. (Compare such lexical prediction with the Communicative Act-based predictions discussed above. In general, it is hoped that by predicting CAs we can in turn predict the structural elements of their cue patterns. We are now shifting the discussion to the prediction of open-class elements instead. The hope is that the two sorts of prediction will prove complementary.)

N-grams provide such predictions only at very short ranges. To support bottom-up parsing of noisy material containing gaps and fragments, longer-range predictions are needed as well. Some researchers have proposed investigation of associations beyond the n-gram range, but the proposed associations remain relatively short-range (about five words). While stochastic grammars can provide somewhat longer-range predictions than n-grams, they predict only within utterances. Our interest, however, extends to predictions on the scale of several utterances.

Thus (Seligman 1994) proposes to permit the definition of windows in a transcribed corpus within which co-occurrences of morphological or lexical elements can be examined. A window is defined as a sequence of minimal segments, where a segment is typically a turn, but can also be a block delimited by suitable markers in the transcript. A flexible set of facilities (CO-OC) has been implemented in Common Lisp to aid collection of such discourse-range co-occurrence information and to provide quick access to the statistics for on-line use.

Sparse data is somewhat less problematic for long-range than for short-range predictions, since it is in general easier to predict what is coming "soon" than what is coming next. Even so, there is never quite enough data; so smoothing will remain important. CO-OC can support various statistical smoothing measures. However, since these techniques are likely to remain insufficient, a new technique for semantic smoothing is proposed and supported: researchers can track co-occurrences of semantic tokens associated with words or morphs in addition to co-occurrences of the words or morphs themselves. The semantic tokens are obtained from standard on-line thesaura. The benefits of such semantic smoothing appear especially in the possibility of retrieving reasonable semantically-mediated associations for morphs which are rare or absent in a training corpus.

A weighted co-occurrence between morphemes or lexemes can be viewed as an association between these items; so the set of co-occurrences which CO-OC discovers can be viewed as an associative or semantic network. Spreading activation within such networks is often proposed as a method of lexical disambiguation. (For example, if the concept MONEY has been observed, then the lexical item bank has the meaning closest to MONEY in the network: "savings institution" rather than "edge of river", etc.) Thus disambiguation becomes a second possible application of CO-OC's results, beyond the abovementioned primary use for constraining speech recognition. A third possible use is in the discovery of topic transitions: we can hypothesize that a span within a dialogue where few co-occurrence predictions are fulfilled is a topic boundary. Once the new topic is determined, appropriate constraints can be exploited, e.g. by selecting a relevant sub-grammar.

Preliminary tests of CO-OC were carried out on a corpus of Japanese-Japanese dialogues concerning street directions and hotel arrangements at ATR Interpreting Telecommunications Laboratories. However, further testing is necessary to demonstrate the reliability and usefulness of the approach. A principle aim would be to determine how large the corpus must be before consistent co-occurrence predictions are obtained.

Conclusions

The six areas of research just examined suggest a six-item wish list for an experimental speech translation system. (1) The system would include facilities for interactive disambiguation of both speech and translation candidates. (2) Its architecture would allow modular reconfiguration and global coordination of components. (3) The system would employ a grammar whose terminals were phones, recognizing both words and syntactic structures in a uniform and integrated manner, e.g. via island-driven chart parsing. (4) Natural pauses and other aspects of prosody would be used to segment utterances and otherwise aid analysis. (5) Speech or dialogue acts would be defined in terms of their cue patterns, and analyses based upon them would be exploited for speech recognition and analysis. (6) Semantically-smoothed tracking of lexical co-occurrences would provide a network of associations useful for speech recognition, lexical disambiguation, and topic boundary recognition.

Acknowledgements

Work on all six of the issues discussed here began at ATR Interpreting Telecommunications Laboratories in Kyoto, Japan. I am very grateful for the support and stimulation I received there.

References

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.

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.

Boitet, Christian and M. Seligman. 1994. "The 'Whiteboard' Architecture: A Way to Integrate Heterogeneous Components of NLP Systems." In *Proceedings of COLING-94*, Kyoto, Aug. 5 - 9, 1994.

Erman, L.D. and V.R. Lesser. 1980. "The Hearsay-II Speech Understanding System: A Tutorial." In *Trends in Speech Recognition*, W.A. Lea, ed., Prentice-Hall, 361-381.

Furukawa, R., F. Yato, and K. Loken-Kim. 1993. *Analysis of Telephone and Multimedia Dialogues.* Technical Report TR-IT-0020, ATR Interpreting Telecommunications Laboratories, Kyoto. (In Japanese)

Julia, L., L. Neumeyer, M. Charafeddine, A. Cheyer, and J. Dowding. 1997. "HTTP:// WWW.SPEECH.SRI.COM/DEMOS/ATIS. HTML." In *Working Notes, Natural Language Pro-cessing for the World Wide Web.* AAAI-97 Spring Symposium, Stanford University, March 24-26, 1997.

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.

Loken-Kim, K., F. Yato, K. Kurihara, L. Fais, and R. Furukawa. 1993. *EMMI-ATR Environment for Multi-modal Interaction.* Technical Report TR-IT-0018, ATR Interpreting Telecommunications Laboratories, Kyoto. (In Japanese).

Morimoto, T., T. Takezawa, F. Yato, et al. 1993. "ATR's Speech Translation System: ASURA." In *Proceedings of Eurospeech-93*, Vol. 2, pages 1291-1294.

Searle, J. 1969. *Speech Acts.* Cambridge: Cambridge University Press, 1969.

Seligman, Mark. 1994. *CO-OC: Semi-automatic Production of Resources for Tracking Morphological and Semantic Co-occurrences in Spontaneous Dialogues.* Technical Report TR-IT-0084, ATR Interpreting Telecommunications Laboratories, Kyoto.

Seligman, Mark. 1997. "Interactive Real-time Translation via the Internet." In *Working Notes, Natural Language Processing for the World Wide Web.* AAAI-97 Spring Symposium, Stanford University. March 24-26, 1997.

Seligman, Mark and C. Boitet. 1994. "A 'Whiteboard' Architecture for Automatic Speech Translation." In *Proceedings of the International Symposium on Spoken Dialogue*, ISSD-93, Waseda University, Tokyo, Nov. 10 - 12, 1993.

Seligman, Mark, Laurel Fais, and Mutsuko Tomokiyo. 1995. *A bilingual set of communicative act labels for spontaneous dialogues.* Technical Report TR-IT-0081, ATR Interpreting Telecommunications Laboratories, Kyoto.

Seligman, Mark, Junko Hosaka, and Harald Singer. 1996. "'Pause Units' and Analysis of Spontaneous Japanese Dialogues: Preliminary Studies." In *Notes of the ECAI-96 Workshop on Dialogue Processing in Spoken Language Systems*, August 12, 1996, Budapest, Hungary. (Also to appear in Springer Series: LNAI - Lecture Notes in Artificial Intelligence.)

Stock, Oliviero, Rino Falcone and Patrizia Insinnamo. 1989. "Bi-directional Charts: A Potential Technique For Parsing Spoken Natural Language Sentences." *Computer Science and Language* (1989) 3, 219-237.

Zajac, Remy and Mark Casper. 1997. "The Temple Web Translator." In *Working Notes, Natural Language Processing for the World Wide Web.* AAAI-97 Spring Symposium, Stanford University, March 24-26, 1997.

Posters