Interactive Speech Translation in the DIPLOMAT Project

Robert Frederking, Alexander Rudnicky, and Christopher Hogan
{ref,air,chogan}@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

The DIPLOMAT rapid-deployment speech translation system is intended to allow naïve users to communicate across a language barrier, without strong domain restrictions, despite the error-prone nature of current speech and translation technologies. Achieving this ambitious goal depends in large part on allowing the users to interactively correct recognition and translation errors. We briefly present the Multi-Engine Machine Translation (MEMT) architecture, describing how it is well-suited for such an application. We then describe our incorporation of interactive error correction throughout the system design. We have already developed a working bidirectional Serbo-Croatian ↔ English system, and are currently developing Haïtian-Creole ↔ English and Korean ↔ English versions.

1 Introduction

The DIPLOMAT project is designed to explore the feasibility of creating rapid-deployment, wearable bi-directional speech translation systems. By "rapid-deployment", we mean being able to develop an MT system that performs initial translations at a useful level of quality between a new language and English within a matter of days or weeks, with continual, graceful improvement to a good level of quality over a period of months. The speech understanding component used is the SPHINX II HMM-based speaker-independent continuous speech recognition system (Huang *et al.*, 1992; Ravishankar, 1996), with techniques for rapidly developing acoustic and language models for new languages (Rudnicky, 1995). The machine translation (MT) technology is the Multi-Engine Machine Translation (MEMT) architecture (Frederking and Nirenburg, 1994), described further below. The speech synthesis component is a newly-developed concatenative system (Lenzo, 1997) based on variable-sized compositional units. This use of subword concatenation is especially important, since it is the only currently available method for rapidly bringing up synthesis for a new language. DIPLOMAT thus involves research in MT, speech understanding and synthesis, interface design, as well as wearable computer systems. While beginning our investigations into new semi-automatic techniques for both speech and MT knowledge-base development, we have already produced an initial bidirectional system for Serbo-Croatian ↔ English speech translation in less than a month, and are currently developing Haïtian-Creole ↔ English and Korean ↔ English systems.

A major concern in the design of the DIPLOMAT system has been to cope with the error-prone nature of both current speech understanding and MT technology, to produce an application that is usable by non-translators with a small amount of training. We attempt to achieve this primarily through user interaction: wherever feasible, the user is presented with intermediate results, and allowed to correct them. In this paper, we will briefly describe the machine translation architecture used in DIPLOMAT (showing how it is well-suited for interactive user correction), describe our approach to rapid-deployment speech recognition and then discuss our approach to interactive user correction of errors in the overall system.

2 Multi-Engine Machine Translation

Different MT technologies exhibit different strengths and weaknesses. Technologies such as Knowledge-Based MT (KBMT) can provide high-quality, fully-automated translations in narrow, well-defined domains (Mitamura *et al.*, 1991; Farwell and Wilks, 1991). Other technologies such as lexical-transfer MT (Nirenburg *et al.*, 1995; Frederking and Brown, 1996; MacDonald, 1963), and Example-Based MT (EBMT) (Brown, 1996; Na-

gao, 1984; Sato and Nagao, 1990) provide lower-quality general-purpose translations, unless they are incorporated into human-assisted MT systems (Frederking *et al.*, 1993; Melby, 1983), but can be used in non-domain-restricted translation applications. Moreover, these technologies differ not just in the quality of their translations, and level of domain-dependence, but also along other dimensions, such as types of errors they make, required development time, cost of development, and ability to easily make use of any available on-line corpora, such as electronic dictionaries or online bilingual parallel texts.

The Multi-Engine Machine Translation (MEMT) architecture (Frederking and Nirenburg, 1994) makes it possible to exploit the differences between MT technologies. As shown in Figure 1, MEMT feeds an input text to several MT engines in parallel, with each engine employing a different MT technology[1]. Each engine attempts to translate the entire input text, segmenting each sentence in whatever manner is most appropriate for its technology, and putting the resulting translated output segments into a shared chart data structure (Kay, 1967; Winograd, 1983) after giving each segment a score indicating the engine's internal assessment of the quality of the output segment. These output (*target language*) segments are indexed in the chart based on the positions of the corresponding input (*source language*) segments. Thus the chart contains multiple, possibly overlapping, alternative translations. Since the scores produced by the engines are estimates of variable accuracy, we use statistical language modelling techniques adapted from speech recognition research to select the best overall set of outputs (Brown and Frederking, 1995; Frederking, 1994). These selection techniques attempt to produce the best overall result, taking the probability of transitions between segments into account as well as modifying the quality scores of individual segments.

Differences in the development times and costs of different technologies can be exploited to enable MT systems to be rapidly deployed for new languages (Frederking and Brown, 1996). If parallel corpora are available for a new language pair, the EBMT engine can provide translations for a new language in a matter of hours. Knowledge-bases for lexical-transfer MT can be developed in a matter of days or weeks; those for structural-transfer MT may take months or years. The higher-quality, higher-investment KBMT-style engine typically requires over a year to bring on-line. The use of the MEMT architecture allows the improvement of initial MT engines and the addition of new engines to occur within an unchanging framework. The only change that the user sees is that the quality of translation improves over time. This allows interfaces to remain stable, preventing any need for retraining of users, or redesign of inter-operating software. The EBMT and Lexical-Transfer-based MT translation engines used in DIPLOMAT are described elsewhere (Frederking and Brown, 1996).

For the purposes of this paper, the most important aspects of the MEMT architecture are:

- the initially deployed versions are quite error-prone, although generally a correct translation is among the available choices, and
- the unchosen alternative translations are still available in the chart structure after scoring by the target language model.

3 Speech recognition for novel languages

Contemporary speech recognition systems derive their power from corpus-based statistical modeling, both at the acoustic and language levels. Statistical modeling, of course, presupposes that sufficiently large corpora are available for training. It is in the nature of the DIPLOMAT system that such corpora, particularly acoustic ones, are not immediately available for processing. As for the MT component, the emphasis is on rapidly acquiring an initial capability in a novel language, then being able to incrementally improve performance as more data and time are available. We have adopted for the speech component a combination of approaches which, although they rely on participation by native informants, also make extensive use of pre-existing acoustic and text resources.

Building a speech recognition system for a target domain or language requires models at three levels (assuming that a basic processing infrastructure for training and decoding is already in place): acoustic, lexical and language.

We have explored two strategies for acoustic modeling. *Assimilation* makes use of existing acoustic models from a language that has a large phonetic overlap with the target language. This allows us to rapidly put a recognition capability in place and was the strategy used for our Serbo-Croatian $\leftrightarrow$ English system. We were able to achieve good recognition performance for vocabularies of up to 733 words using this technique. Of course, such overlaps cannot be relied upon and in any case will not produce recognition performance that approaches that possible with appropriate training. Nevertheless it does suggest that useful recognition performance for a large set of languages can be achieved given a carefully chosen set of core languages that can serve as a source of

[1] Morphological analysis, part-of-speech tagging, and possibly other text enhancements can be shared by the engines.

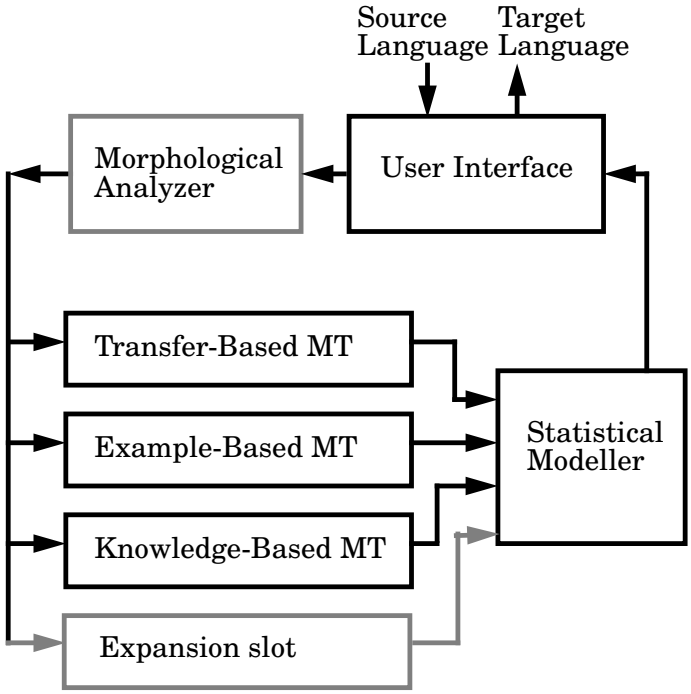

Figure 1: **Structure of MEMT architecture**

acoustic models for a cluster of phonetically similar languages.

The *selective collection* approach presupposes a preparation interval prior to deployment and can be a follow-on to a system based on assimilation. This is being developed in the context of our Haïtian-Creole and Korean systems. The goal is to carry out a limited acoustic data collection effort using materials that have been explicitly constructed to yield a rich phonetic sampling for the target language. We do this by first computing phonetic statistics for the language using available text materials, then designing a recording script that exhaustively samples all diphones observed in the available text sample. Such scripts run from several hundred to around a thousand utterances for the languages we have examined. While the effectiveness of this approach depends on the quality (and quantity) of the text sample that can be obtained, we believe it produces appropriate data for our modeling purposes.

Lexical modeling is based on creating pronunciations from orthography and involves a variety of techniques familiar from speech synthesis, including letter-to-sound rules, phonological rules and exception lists. The goal of our lexical modeling approach is to create an acceptable-quality pronouncing dictionary that can be variously used for acoustic training, decoding and synthesis. We work with an informant to map out the pronunciation system for the target language and make use of supporting published information (though we have found such to be misleading on occasion). System vocabulary is derived from the text materials assembled for acoustic modeling, as well as scenarios from the target domain (for example, interviews focussed on mine field mapping or intelligence screening).

Finally, due to the goals of our project, language modeling is necessarily based on small corpora. We make use of materials derived from domain scenarios and from general sources such as newspapers (scanned and OCRed), text in the target language available on the Internet and translations of select documents. Due to the small amounts of readily available data (on the order of 50k words for the languages we have worked with), standard language modeling tools are difficult to use, as they presuppose the availability of corpora that are several orders of magnitude larger. Nevertheless we have been successful in creating standard backoff trigram models from very small corpora. Our technique involves the use of high discounts and appears to provide useful constraint without corresponding fragility in the face of novel material.

In combination, these techniques allow us to create working recognition systems in very short periods of time and provide a path for evolutionary improvement of recognition capability. They clearly are not of the quality that would be expected if conventional procedures were used, but nevertheless are sufficient for providing cross-language communication capability in limited-domain speech translation.

4 User Interface Design

As indicated above, our approach to coping with error-prone speech translation is to allow user correction wherever feasible. While we would like as much user interaction as possible, it is also important not to overwhelm the user with either information or decisions. This requires a careful balance, which we are trying to achieve through early user testing. We have carried out initial testing using local naïve subjects (e.g., drama majors and construction workers), and intend to test with actual end users once specific ones are identified.

The primary potential use for DIPLOMAT identified so far is to allow English-speaking soldiers on peace-keeping missions to interview local residents. While one could conceivably train the interviewer to use a restricted vocabulary, the interviewee's responses are much more difficult to control or predict. An initial system has been developed to run on a pair of laptop computers, with each speaker using a graphical user interface (GUI) on the laptop's screen (see Figure 2). Feedback from initial demonstrations made it clear that, while we could expect the interviewer to have roughly eight hours of training, we needed to design the system to work with a totally naïve interviewee, who had never used a computer before. We responded to this requirement by developing an asymmetric interface, where any necessary complex operations were moved to the interviewer's side. The interviewee's GUI is now extremely simple, and a touch screen has been added, so that the interviewee is not required to type or use the pointer. In addition, the interviewer's GUI controls the state of the interviewee's GUI. The speech recognition system continuously listens, thus the participants do not need to physically indicate their intention of speaking.

A typical exchange consists of recognizing the interviewer's spoken utterance, translating it to the target language, backtranslating it to English[2], then displaying and synthesizing the (possibly corrected) translation. The interviewee's response is recognized, translated to English, and backtranslated. The (possibly corrected) backtranslation is then shown to the interviewee for confirmation. The interviewer receives a graphic indication of whether the backtranslation was accepted or not. (The actual communication process is quite flexible, but this is a normal scenario.)

In order to achieve such communication, the users currently can interact with DIPLOMAT in the following ways:

- **Speech displayed as text**: After any speech recognition step, the best overall hypothesis is displayed as text on the screen. The user can highlight an incorrect portion using the touch-screen, and respeak or type it.
- **Confirmation requests**: After any speech recognition or machine translation step, the user is offered an accept/reject button to indicate whether this is "what they said". For MT, backtranslations provide the user with an ability to judge whether they were interpreted correctly.
- **Interactive chart editing**: As mentioned above, the MEMT technology produces as output a chart structure, similar to the word hypothesis lattices in speech systems. After any MT step, the **interviewer** is able to edit the best overall hypothesis for either the forward or backward translation using a popup-menu-based editor, as in our earlier Pangloss text MT system (Frederking *et al.*, 1993). The editor allows the interviewer to easily view and select alternative translations for any segment of the translation. Editing the forward translation causes an automatic reworking of the backtranslation. Editing the backtranslation allows the interviewer to recognize correct forward translations despite errors in the backtranslation; if the backtranslation can be edited into correctness, the forward translation was probably correct.

Since a major goal of DIPLOMAT is rapid-deployment to new languages, the GUI uses the UNICODE multilingual character encoding standard. This will not always suffice, however; a major challenge for handling Haïtian-Creole is that 55% of the Haïtian population is illiterate. We will have to develop an all-speech version of the interviewee-side interface. As we have done with previous interface designs, we will carry out user tests early in its development to ascertain whether our intuitions on the usability of this version are correct.

5 Conclusion

We have presented here the DIPLOMAT speech translation system, with particular emphasis on the user interaction mechanisms employed to cope with error-prone speech and MT processes. We expect that, after additional tuning based on further informal user studies, an interviewer with eight hours of training should be able to use the

[2] We realize that backtranslation is also an error-prone process, but it at least provides some evidence as to whether the translation was correct to someone who does not speak the target language at all.

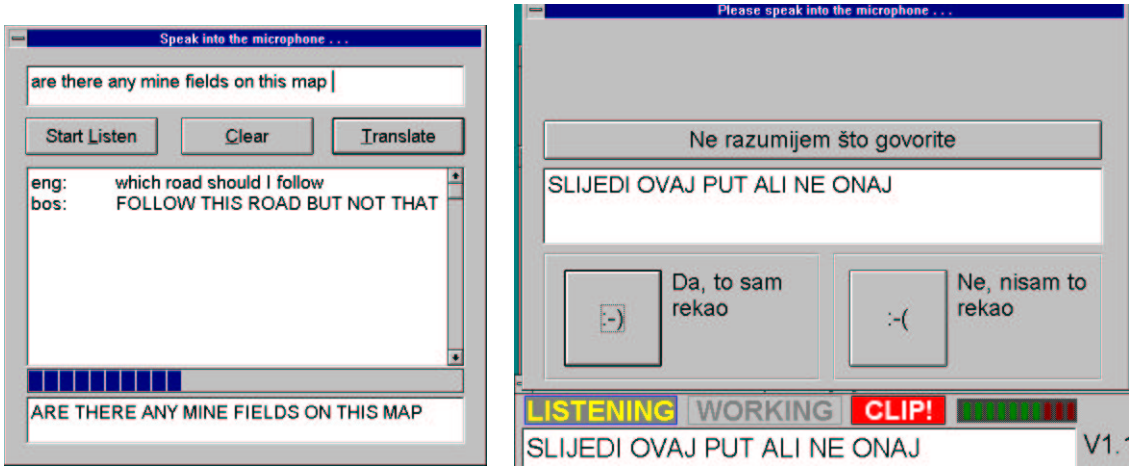

Figure 2: **Screen Shot of User Interfaces: Interviewer (left) and Interviewee (right)**

DIPLOMAT system to successfully interview subjects with no training or previous computer experience. We hope to have actual user trials of either the Serbo-Croatian or the Haïtian-Creole system in the near future, possibly this summer.

References

Ralf Brown. 1996. Example-Based Machine Translation in the Pangloss System. In *Proceedings of the 16th International Conference on Computational Linguistics* (COLING-96).

Ralf Brown and Robert Frederking. 1995. Applying Statistical English Language Modeling to Symbolic Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI-95), pages 221–239.

David Farwell and Yorick Wilks. 1991. Ultra: A Multi-lingual Machine Translator. In *Proceedings of Machine Translation Summit III*, Washington, DC, July.

Robert Frederking. 1994. Statistical Language Models for Symbolic MT. Presented at the *Language Engineering on the Information Highway Workshop*, Santorini, Greece, September. Refereed.

Robert Frederking, D. Grannes, P. Cousseau, and S. Nirenburg. 1993. An MAT Tool and Its Effectiveness. In *Proceedings of the DARPA Human Language Technology Workshop*, Princeton, NJ.

Robert Frederking and Ralf Brown. 1996. The Pangloss-Lite Machine Translation System. In *Proceedings of the Conference of the Association for Machine Translation in the Americas* (AMTA).

Robert Frederking and Sergei Nirenburg. 1994. Three Heads are Better than One. In *Proceedings of the fourth Conference on Applied Natural Language Processing* (ANLP-94), Stuttgart, Germany.

Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, Ronald Rosenfeld. 1992. The SPHINX-II Speech Recognition System: An Overview. Carnegie Mellon University Computer Science Technical Report CMU-CS-92-112.

Martin Kay. 1967. Experiments with a powerful parser. In *Proceedings of the 2nd International COLING*, August.

Kevin Lenzo. 1997. Personal Communication.

R. R. MacDonald. 1963. General report 1952-1963 (Georgetown University Occasional Papers in Machine Translation, no. 30), Washington, DC.

A. K. Melby. 1983. Computer-assisted translation systems: the standard design and a multi-level design. *Conference on Applied Natural Language Processing*, Santa Monica, February.

Teruko Mitamura, Eric Nyberg, Jaime Carbonell. 1991. Interlingua Translation System for Multi-Lingual Document Production. In *Proceedings of Machine Translation Summit III*, Washington, DC, July.

M. Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In: A. Elithorn and R. Banerji (eds.) *Artificial and Human Intelligence*. NATO Publications.

Sergei Nirenburg. 1995. The Pangloss Mark III Machine Translation System. Joint Technical Report, Computing Research Laboratory (New Mexico State University), Center for Machine Translation (Carnegie Mellon University),

Information Sciences Institute (University of Southern California). Issued as CMU technical report CMU-CMT-95-145.

Mosur Ravishankar. 1996. *Efficient Algorithms for Speech Recognition.* Ph.D. Thesis. Carnegie Mellon University.

Alex Rudnicky. 1995. Language modeling with limited domain data. In *Proceedings of the ARPA Workshop on Spoken Language Technology.* San Mateo: Morgan Kaufmann, 66-69.

S. Sato and M. Nagao. 1990. Towards memory based translation. In *Proceedings of COLING-90*, Helsinki, Finland.

Terry Winograd. 1983. *Language as a Cognitive Process. Volume 1: Syntax.* Addison-Wesley.