Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 77-84.
Association for Computational Linguistics.

Speech Translation on a Tight Budget without Enough Data

Robert E. Frederking*, Alan W Black*, Ralf D. Brown*, Alexander Rudnicky*, John Moody†, Eric Steinbrecher†

*Language Technologies Institute
Carnegie Mellon University

†Lockheed Martin Systems Integration
Owego, New York, USA

Abstract

The Tongues speech-to-speech translation system was developed for the US Army chaplains, with fairly stringent constraints on time, budget, and available data. The resulting prototype was required to undergo a quite realistic field test. We describe the development and architecture of the system, the field test, and our analysis of its results. The system performed quite well, especially given its development constraints.

1 Introduction

This paper describes the one-year-long Tongues speech-to-speech translation project. We feel that this project was especially interesting in that we developed this system with fairly stringent constraints on time, budget, and available data, and were required to carry out a field-test of a prototype at the end of the project. Despite all the constraints, the system actually performed quite well in a true field test with real naive users.

We begin this paper by first describing the purpose of, and constraints on, the project. We then present the architecture and development of the speech translation system. We describe the ensuing field test, and finally, present our analysis of the results.

2 Tongues Project Concept: Chaplain support

The Tongues system was funded by the US Army to support the mission of the US Army chaplains, who are increasingly called upon to deal with local populations, usually without the benefit of human translators. It is thus intended to be used by a trained US Army chaplain with a completely naive and untrained non-English speaker. The chaplains' translation problem is compounded by relatively short notice when a new language will be needed, limited funding for any given language, and vaguely defined domains of discourse.

2.1 Limited time/budget

While the initial system was specifically to demonstrate translation in both directions between English and Croatian, the design was also required to allow rapid development for new languages. To ensure rapid development, the entire project was only allowed to take one calendar year, including contractual arrangements, hiring language experts, etc. The total development effort was similarly restricted: six senior research personnel provided an estimated total of about two (2) full-time person-years of effort. In addition to the senior staff, there were also part-time Croatian informants, chaplains, and some student programmers.

In addition to development *time*, development *cost* is also an important consideration for many languages. It is no accident that there are no commercial MT systems available for Haitian Creole, for example. Commercial MT

systems typically have many dozens or even hundreds of person-years of effort invested in them. For many minor languages, there simply is not enough economic activity to justify such an investment. While a major government might invest in developing a system for a minor language if it is of sufficient political interest, in today's volatile international climate there are many possibly interesting languages, so any one language will only receive limited investment. Thus low-cost techniques are the only realistic option, if we wish to make MT systems available for such languages in the foreseeable future.

2.2 Limited data

As another constraint, in addition to rapid development and budget requirements, the system was not permitted to be restricted to a narrowly-limited domain, but had to be wide-coverage. (All of these properties were important for the chaplains' envisioned activities.) Since we were to build a broad-coverage system in a short period of time on a small budget, data-driven approaches were the only reasonable choice.

This raises an important question: since the approaches we are using (described below) rely heavily on parallel corpora, bilingual glossaries, and dictionaries, can we in fact achieve significant savings over more traditional system development in terms of human effort, time and cost? Our experience shows that we can, for several reasons. Firstly, for some languages many of the necessary resources are easily available, and can usually be converted into a desirable format quite easily. Even languages that do not offer much in the way of online text resources usually have a dictionary and/or glossary available. Adapting dictionaries, glossaries and parallel corpora is significantly easier than similarly adapting grammar rules and parsers, should they be available. Secondly, even when it is necessary to *produce* a bilingual parallel corpus and/or glossaries within the project, it is easier to train native speakers to translate into their own language than to find or train native-speaker knowledge engineers to produce grammars, etc., for a minor language. Finally, we have shown that it is possible to significantly shorten (by more than a factor of three) the amount of time required to produce a corpus with a given level of coverage, by carefully choosing the texts to translate (Allen and Hogan, 1998).

3 Tongues System Design and Implementation

The architecture and user interface of the Tongues system were based in large measure on the Diplomat system (Frederking et al., 2000); the main change was changing the synthesis system to be the open-source Festival (Black et al., 1998) system. The speech recognition system used was the open-source Sphinx II (Huang et al., 1992), and the translation system was an EBMT/MEMT (Example-Based MT/Multi-Engine MT) system (Brown, 1996; Frederking and Nirenburg, 1994; Brown and Frederking, 1995) very similar to that in Diplomat.

We have provided a more detailed description of the development of the Tongues system elsewhere (Black et al., 2002b).

3.1 Domain data collection

In order to provide in-domain conversational data, we arranged at the start of the project to record a number of chaplains in role-playing conversations of the type they expected the device to encounter. Fortunately, the chaplains were familiar with role-playing exercises, and all had relevant field experiences to re-enact. Both sides of the conversations were spoken in English. These were digitally recorded with head-mounted microphones at 16KHz in stereo (one speaker on each channel), as this was closest to the intended audio channel characteristics of the eventual system. In all, we recorded 46 conversations, ranging from a few minutes to 20 minutes length. This provided a total of 4.25 hours of actual English speech. The recorded conversations were hand-transcribed at the word level, and translated into Croatian by native Croatian speakers.

3.2 Speech components

Since the speech components of the system were previously-developed, open-source systems, we

will only describe here the development of their training data, emphasizing the small amount of data used. The development of the Tongues speech components is described in greater detail elsewhere (Black et al., 2002a).

English speech components. The English recordings of the role-playing exercises were used directly for training the English acoustic models. That is, we took **only** these 4.25 hours of chaplain speech and directly trained semi-continous HMM models for Sphinx2 (Huang et al., 1992).

For the English language model, we required a larger collection of in-domain text. We used the dialog transcriptions themselves, but also augmented that with text from chaplain handbooks that were made available to us. Although we knew we could provide better recognition accuracy by using more resources, we were interested in limiting what resources were necessary for this work, and also (see below) we found the trained models from this data adequate for the task.

Croatian speech components. Building Croatian acoustic models was harder. As we were aware that our resource of Croatian speakers was limited, and they had less skill in carrying out full word transcription of conversational speech, we wished to find a simpler, less resource-intensive method to build Croatian acoustic models. From the the translated chaplain transcripts, we wished to select example utterances that when recorded would give sufficient acoustic coverage to allow reasonable acoustic models to be trained. To do this, we used a technique originally developed for selecting text to record for speech synthesis (Black and Lenzo, 2001). From a list of several thousand utterances, we used this technique to select groups of 250 utterances that were phonetically rich. These sets were then read by a number of native Croatian speakers. Using read speech avoided the process of hand-transcription of the speech, though it does make it less like the intended conversational speech. Due to the relative scarcity of native Croatian speakers, we recorded only 15 different speakers, of which 13 were female and 2 were male. This resulted in a gender imbalance, which concerned us greatly, but was not observed to affect the system's performance greatly in the field. In all, a total of 4.0 hours of Croatian speech was collected. This data **alone** was then used to train new acoustic models for Croatian.

The Croatian language model was built from the Croatian side of the translation data (see below), which included the Croatian translations of the role-playing exercises.

3.3 Machine Translation component

We will now describe our MT component. Due to the project requirements described above, a translation technology was needed that was well-suited to the problems of rapid, low-cost deployment and wide-domain coverage. While more traditional Knowledge-Based MT (KBMT) can provide high-quality MT in a narrow domain, KBMT systems typically require over a year to bring online, and an order of magnitude more professional staff than we could afford. Statistical MT systems typically require very large amounts of domain data that are not usually available for speech translation applications. In contrast, the EBMT/MEMT translation approach that we had used in the Diplomat system seemed appropriate for this application.

3.3.1 Our EBMT/MEMT design

The Multi-Engine MT (MEMT) system that was used in Diplomat (Frederking et al., 2000) combines a "shallow" EBMT system with any available glossaries or dictionaries.

There are two primary differences between our "shallow" EBMT system and other systems that employ translation by analogy. First, our basic system relies primarily on matching sequences of words (surface strings) in a simple parallel corpus of corresponding sentence pairs in the two languages, rather than matching more complex representations (such as parse trees). Our rationale is that EBMT is being used as the main initial engine in a rapid-deployment MT system; if we will have time to develop deep analysis for a language, we will use the deep analysis as part of a KBMT engine instead of incorporating it into EBMT. Second, our system returns all matching candidates with a minimum level of

goodness, rather than trying to determine the optimal match. The optimal matching problem can require rather extensive computational resources (Horiguchi and Franz, 1997); fortunately we can avoid it within our EBMT, because the MEMT system handles the selection process externally, using a trigram model of the target language, as described elsewhere (Frederking et al., 2000). (This selection process is essentially identical to the stack decoder used in many speech recognizers to combine acoustic hypothesis scores with trigram language model probabilities.)

While the use of bilingual glossaries and dictionaries is a low-quality technique, its simplicity allows us to quickly and semi-automatically develop large databases using native speakers with no special training, allowing an initial rapid-deployment of an MT system even when parallel corpora are unavailable. They are also useful for "filling the cracks" when gaps are discovered in our parallel corpus during system testing.

3.3.2 Translation data

The training corpus for the EBMT engine consisted of the transscripts of the chaplain dialogs and their translations plus pre-existing parallel text from the Diplomat project (Frederking et al., 2000) and newly-acquired parallel text from the web. The dictionary/glossary engine used both statistically-extracted translations and manually-created entries. The English trigram model already existed, and had been generated from newswire and broadcast news transcripts. Finally, the Croatian trigram model was built from the Croatian half of the EBMT corpus, some Croatian text found on the web, and the full text of some sixty novels and other Croatian literary works (in total, approximately six million words).

3.3.3 Effects of spoken input

There appears to be an interesting match between the properties of spoken input and the properties of a rapid-deployment EBMT/MEMT system. Compared to text translation, the input to speech translation is of much lower quality, due both to the word-error-rate of state-of-the-art real-time continuous speech recognition and to the disfluencies present in spontaneous speech. That is, spontaneous speakers often do not utter the complete, grammatical sentences that linguistic analysis typically expects. As noted above, KBMT systems do produce better quality translation than the EBMT and glossary/dictionary engines employed in rapid-deployment EBMT/MEMT. But the degraded quality of the input means that the quality difference between KBMT and rapid-deployment EBMT/MEMT is less important; given a string of words containing word errors and structural anomalies, it appears to us that a rapid-deployment EBMT/MEMT system can do about as well as a (much more costly) KBMT system. This claim of course would require serious testing before it could be asserted as fact. However, at least two other major spoken language translation systems, JANUS (Waibel, 1996; Levin et al., 2000) and SRI Cambridge's SLT (Rayner and Carter, 1997), have adopted some form of Multi-Engine MT.

3.4 System-level issues

Simply stringing together a recognizer, translator, and synthesizer does not make a very useful speech-to-speech translation system. A good interface is necessary to make the parts work together in such a way that a user can actually derive benefit from it. Using our experience from the earlier Diplomat system, we designed the Tongues interface to be asymmetric, with the Croatian side being as simple as possible, and any necessary complexity handled on the English side, since the chaplain would be trained and practiced in using the system. Note that even the English side was not terribly complex.

We included a back-translation capability, to allow a user with no knowledge of the target language to better assess the quality of the translation. (We could not use the approach of generating paraphrases from meaning representations, since the system does not use any meaning representations.) We also included several user-requested features, such as built-in pre-recorded

instructions and explanations for the Croatian (since the Croatian speaker is completely naive regarding the device and the chaplain's intentions), emergency key phrases (such as "Don't move!"), and enhancements such as being able to modify the translation lexicon in the field, so that the system could be tuned to more specific tasks.

The final system ran on a Windows-based Toshiba Libretto, running at 200MHz with 192MB of memory. At the time of the project (2000) this was the best combination of speed and size that was readily available. The system was equipped with a custom touchscreen, so that the Croatian-speaker would not need to type or use a mouse at all. Aware that the system might be used in situations where the non-English participant would be unfamiliar with computer technology, we included a microphone/speaker handset that looks like a conventional telephone handset. This has the advantage of provided a close-talking microphone, thus making speech recognition easier, while coming in a form factor that will be familiar to most people.

Our design provides abundant opportunities for user error correction, in an effort to enable cooperative users to communicate well enough to accomplish significant tasks that they could not accomplish without the system (or a bilingual human interpreter), despite the error-prone nature of current speech recognition, broad-coverage rapid-development machine translation, and speech synthesis. Determining whether we have met such a goal requires task-based evaluation; while error rates of components are useful information, the real system-level issue is whether communication is achieved, and at what level of effort.

4 Tongues Field Test

The US Army ACT-II program under which Tongues was funded is designed to result in field tests of deployable prototypes. Accordingly, in April 2001, representatives of the development team traveled to Zagreb, Croatia, with representatives of the US Army chaplains. We had arranged in advance to have native-Croatian speakers available as conversation partners. This was done by contacting someone at the University of Zagreb, and hiring them as a local organizer. They were instructed to recruit a large number of potential test subjects varying in gender and age, with as little English knowledge as possible.

Since the principal domain of the translation system was interaction with refugees, we prepared a number of refugee scenarios for the Croatian subjects and American chaplains to act out using the translation device. The scenarios were in the intended domain, involving refugees, medical supplies and getting general directions. The refugee side of each scenario was translated into Croatian. We also prepared a questionnaire for each participant, produced translated Croatian questionnaires, and after the test had the Croatian responses translated in to English.

We then travelled to Croatia. Over a three-day period, at the University of Zagreb, naive Croatians were brought into the room knowing only that they were supposed to enact the scenario that they had just been given with a US Army officer, who would be using a translation device. The Croatian only knew the refugee side of the scenario, while the US officer only knew the Army side of the scenario. The actual Croatian subjects consisted of 21 speakers, male and female, of various ages ranging from young teenagers through adults. Each dialog was logged by the system to allow further analysis.

5 Analysis of Field Test Results

As mentioned above, we generated questionnaire responses and system transcript logs in the course of our tests. These are described in detail in another paper (Frederking et al., 2002). The essential result from analysis of the questionnaires was that of 19 questionnaires, communication was described as "good" by 5, "okay" by 11, and "bad" by 3. We feel that this 16% failure rate was clearly overly generous on the participants' parts. The most interesting result from the system transcript logs was that there were 4.67 words per English turn, 3.51 words per

Croatian turn, and 1.37 minutes/turn. Thus the participants used very short sentences, and the system was very slow.

We also directly observed the conversations and took notes. Our subjective impression of the results was that the conversations went reasonably well more than one half of the time. In addition to cases where the parties failed to complete their tasks, the system was often frustrating to use, due to the large amount of user error correction often required, and the corresponding slowness of the dialogue.

Difficulties described by the participants range over all the components; but our subjective impression was that the speech components performed quite acceptably; the translation component was the weakest link. (This was especially surprising to us given that the speech components were not trained on a large amount of data.) In particular, as our rapid-development translation system contains no internal representation of the meaning of the utterance, the only method for feedback of the translation results to the (monolingual) user is (independent) back-translation, as described above. This risks doubling the error rate, and a bilingual team member in fact observed that often an English-to-Croatian translation that was basically correct would be rejected by an English-speaker because the back-translation was seriously garbled.

It is important to note, and immediately obvious when participating in such a conversation, that communication through a translation device is not fast. Each person must speak, check the recognized form and possibly correct it, translate the utterance (possibly checking with back-translation), and then synthesize the result. Such devices thus do not enable truly spontaneous communication, as they deliberately allow the participants to review the translations and decide when they are adequate. It is possible for the component technologies (recognition, translation and synthesis) to become more streamlined, but it would be very difficult to achieve truly spontaneous, simultaneous translation.

In looking over the conversations, it is clear that the translations are often far from ideal, though usually understandable. For example in answer to the question "where are they?" the device produces "twenty minutes of village." The quality in the English to Croatian translations is similar, in our judgment.

Other specific observations we noted were that the users could not easily identify where the problems lay with the system. For example, if speech recognition produced and displayed a correct transcript, and then translation produced an unacceptable result, they would usually *respeak* the same utterance using the same words! Similarly, mistakes in the synthesizer were often erroneously attributed to the translator (and vice versa, despite the output text being visible in the user interface. Thus even if we provided separate user methods to add words to the recognizer, language model, and translation engine, it is clear that the user would not be able to identify which part (or parts) need to be updated. As there is strong user demand for such systems to provide methods of adaptation in the field, it is clear that the interface presented to the user to offer that adaptation needs more work.

A second observation was that the participants continued to use speech and did not resort to the alternative typing interface (although they were clearly aware of it), and only resorted to typing as a last resort. This may have been due to the fact the participants were asked to use the speech-to-speech translation device rather than being given the more abstract goal of achieving successful communication by the best means. The very small keyboard on the (required) small device may also have been a significant factor, in addition to the well-known preference many naive users have for speaking over typing.

We also note an interesting phenomenon with a limitation in the system in dealing with unknown words. Often such out of vocabulary words have direct cognates in the other language, and hence are directly understandable. We could see that some conjugations of the Croatian word for "kilometer" were not recognized by the Croatian speech recognition sys-

tem, and hence failed to translate. When a word fails to translate, the system presents the word in its original language, but capitalized, in the translation output. For example, the recognized phrase "pet gje ometa" is translated as "five GJE OMETA"; given the context, it was clear to the English speaker that the Croatian speaker had said "five kilometers" (in Croatian). A similar example happened with the word "helicopter".

This point is important. We have two people cooperating and actively trying to communicate. Thus where cognates exist, the listener will understand and accommodate mis-recognitions.

We also noted that, as a consequence of the slowness of communication, the participants took more time to think about about they were going to say. Thus their utterances were on the whole more complete sentences than the fragments that one typically encounters in normal conversational speech. This factor almost certainly compensated for the fact that our Croatian speech recognizer was trained on read speech. Conversely, it probably slightly hindered English recognition, as that was trained on more spontaneous conversations.

The conversations took place in a quiet classroom situation, with little external noise. This helped both the speech recognition and the user understanding of the speech output. However, it is also worth noting that synthetic speech is much easier to understand when the written form of what is being spoken also appears on the screen in front of the (literate) listener.

Finally, we also noted that some English questions were answered with simple yes/no answers without using the device to translate them. The effort of translating simple one-word utterances (such as "da"), which can often easily be understood without knowing the language, was unnecessary.

6 Conclusion

We feel that this field test of the Tongues system was unusually rigorous and realistic, in that we tested the system using regular US Army officers speaking with naive Croatians who did not live in an English-speaking country. This was important, since if our system performed well in the field test, it would conceivably have gone into actual use.

Our simple approach appears to have been surprisingly adequate. The official report by the US Army participants was that the system is worth further development, since it is approaching the quality necessary for real use, but still requires further development before actual field use. We believe that this is actually quite a good result, given the current state of speech and MT technology, and especially the time, cost, and broad-coverage constraints of this project.

Acknowledgments

We would like to thank Rita Singh for her help on speech recognition, and Kevin Lenzo for his help in speech synthesis.

References

Jeffrey Allen and Christopher Hogan. 1998. Expanding lexical coverage of parallel corpora for the ebmt approach. In *First Conference on Language Resources and Evaluation (LREC '98)*, pages 747–754, Granada, Spain, May.

A. Black and K. Lenzo. 2001. Optimal data selection for unit selection synthesis. In *4rd ESCA Workshop on Speech Synthesis*, Scotland.

A. Black, P. Taylor, and R. Caley. 1998. The Festival Speech Synthesis System. http://festvox.org/festival.

A. Black, R. Brown, R. Frederking, K. Lenzo, J. Moody, A. Rudnicky, R. Singh, and E. Steinbrecher. 2002a. Rapid devlopement of speech-to-speech translation systems. Submitted to ICSLP2002.

A. Black, R. Brown, R. Frederking, R. Singh, J. Moody, and E. Steinbrecher. 2002b. TONGUES: Rapid Development of a Speech-to-Speech Translation System. In *Proceedings of HLT-2002*, San Diego, CA, USA.

R. Brown and R. 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.

R. Brown. 1996. Example-based machine translation in the Pangloss system. In *Proceedings of COLING-96*, pages 169–174, Copenhagen, Denmark.

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

R. Frederking, A. Rudnicky, C. Hogan, and K. Lenzo. 2000. Interactive Speech Translation in the Diplomat Project. *Machine Translation Journal*, 15(1-2):27–42. Special Issue on Spoken Language Translation.

R. Frederking, A. Black, R. Brown, J. Moody, and E. Steinbrecher. 2002. Field Testing the Tongues Speech-to-Speech Machine Translation System. LREC 2002.

Keiko Horiguchi and Alexander Franz. 1997. A formal basis for spoken language translation by analogy. In Steven Krauwer et al., editors, *Proceedings of the Spoken Language Translation Workshop*, pages 32–39, Madrid, Spain, July. ELSNET.

X. Huang, F. Alleva, H.-W. Hon, K.-F. Hwang, M.-Y. Lee, and R. Rosenfeld. 1992. The SPHINX-II Speech Recognition System: an overview. *Computer Speech and Language*, 7(2):137–148.

Lori Levin, Alon Lavie, Monika Woszczyna, and Alex Waibel. 2000. The Janus III Translation System. *Machine Translation Journal*, 15(1-2). Special Issue on Spoken Language Translation.

Manny Rayner and David Carter. 1997. Hybrid processing in the spoken language translator. In *Proceedings of ICASSP-97*, Munich, Germany.

Alex Waibel. 1996. Interactive translation of conversational speech. *Computer*, 29(7), July.