Speech to Speech Translation for Medical Triage in Korean

Farzad Ehsani, Jim Kimzey, Demitrios Master, Karen Sudre

Engineering Department
Sehda, Inc.
Mountain View, CA 94043
{farzad,jkimzey,dlm,karen}@sehda.com

Hunil Park

Independent Consultant
Seoul, Korea
phunil@hotmail.com

Abstract

S-MINDS is a speech translation engine, which allows an English speaker to communicate with a non-English speaker easily within a question-and-answer, interview-style format. It can handle limited dialogs such as medical triage or hospital admissions. We have built and tested an English-Korean system for doing medical triage with a translation accuracy of 79.8% (for English) and 78.3% (for Korean) for all non-rejected utterances. We will give an overview of the system building process and the quantitative and qualitatively system performance.

1 Introduction

Speech translation technology has the potential to give nurses and other clinicians immediate access to consistent, easy-to-use, and accurate medical interpretation for routine patient encounters. This could improve safety and quality of care for patients who speak a different language from that of the healthcare provider.

This paper describes the building and testing of a speech translation system, S-MINDS (Speaking Multilingual Interactive Natural Dialog System), built in less than 4 months from specification to the test scenario described. Although this paper shows a number of areas for improvement in the S-MINDS system, it does demonstrate that building and deploying a successful speech translation system is becoming possible and perhaps even commercially viable.

2 Background

Sehda is focused on creating speech translation systems to overcome language barriers in healthcare settings in the U.S. The number of people in the U.S. who speak a language other than English is large and growing, and Spanish is the most commonly spoken language next to English. According to the 2000 census, 18% of the U.S. population aged 5 and older (47 million people) did not speak English at home.[1] This represents a 48% increase from the 1990 figure. In 2000, 8% of the population (21 million) was Limited English Proficient (LEP). More than 65% of the LEP population (almost 14 million people) spoke Spanish.

A body of research shows that language barriers impede access to care, compromise quality, and increase the risk of adverse outcomes. Although trained medical interpreters and bilingual healthcare providers are effective in overcoming such language barriers, the use of semi-fluent healthcare professionals and ad hoc interpreters causes more interpreter errors and lower quality of care (Flores 2005).

One study analyzed the problem of language barriers for hospitalized inpatients. The study, which focused on pediatric patients, sought to determine whether patients whose families have a language barrier are more likely to incur serious medical errors than patients without a language barrier (Cohen et al., 2005). The study's conclusion was that patients of LEP families had a twofold increased risk for serious medical incident compared with patients whose families did not have a language barrier. It is important to note that the LEP

1 US Census Bureau, 2000

patients in this study were identified as needing interpreters during their inpatient stay and medical interpreters were available.

Although the evidence favors using trained medical interpreters, there is a gap between best practice and reality. Many patients needing an interpreter do not get one, and many must use ad hoc interpreters. In a study of 4,161 uninsured patients who received care in 23 hospitals in 16 cities, more than 50% who needed an interpreter did not get one (Andrulis et al., 2002).

Another study surveyed 59 residents in a pediatric residency program in an urban children's hospital (O'Leary and Hampers, 2003). Forty of the 59 residents surveyed spoke little or no Spanish. Again, it is important to note that this hospital had in-house medical interpreters. Of this group of nonproficient residents:

- 100% agreed that the hospital interpreters were effective; however, 75% "never" or only "sometimes" used the hospital interpreters.
- 53% used their inadequate language skills in the care of patients "often" or "every day."
- 53% believed the families "never" or only "sometimes" understood their child's diagnosis.
- 43% believed the families "never" or only "sometimes" understood discharge instructions.
- 40% believed the families "never" or only "sometimes" understood the follow-up plan.
- 28% believed the families "never" or only "sometimes" understood the medications.
- 53% reported calling on their Spanish-proficient colleagues "often" or "every day" for help.
- 80% admitted to avoiding communication with non-English-speaking families.

The conclusion of the study was as follows: "Despite a perception that they are providing suboptimal communication, nonproficient residents rarely use professional interpreters. Instead, they tend to rely on their own inadequate language skills, impose on their proficient colleagues, or avoid communication with Spanish-speaking families with LEP."

Virtually every study on language barriers suggests that these residents are not unique. Physicians and staff at several hospitals have told Sehda that they are less likely to use a medical interpreter or telephone-based interpreter because it takes too long and is too inconvenient. Sehda believes that to bridge this gap requires 2-way speech translation solutions that are immediately available, easy to use, accurate, and consistent in interpretation.

The need for speech translation exists in healthcare, and a lot of work has been done in speech translation over the past two decades. Carnegie-Mellon University has been experimenting with spoken language translation in its JANUS project since the late 1980s (Waibel et al., 1996). The University of Karlsruhe, Germany, has also been involved in an expansion of JANUS. In 1992, these groups joined ATR in the C-STAR consortium (Consortium for Speech Translation Advanced Research) and in January 1993 gave a successful public demonstration of telephone translation between English, German and Japanese, within the limited domain of conference registrations (Woszczyna, 1993). A number of other large companies and laboratories including NEC (Isotani, et al., 2003) in Japan, the Verbmobil Consortium (Wahlster, 2000), NESPOLE! Consortium (Florian et al., 2002), AT&T (Bangalore and Riccardi, 2001), and ATR have been making their own research effort (Yasuda et al., 2003). LC-Star and TC-Star are two recent European efforts to gather the data and the industrial requirements to enable pervasive speech-to-speech translation (Zhang, 2003). Most recently, the DARPA TransTac program (previously known as Babylon) has been focusing on developing deployable systems for English to Iraqi Arabic.

3 System Description

Unlike other systems that try to solve the speech translation problem with the assumption that there is a moderate amount of data available, S-MINDS focuses on rapid building and deployment of speech translation systems in languages where little or no data is available. S-MINDS allows the user to communicate easily in a question-and-answer, interview-style conversation across languages in limited domains such as border control,

hospital admissions or medical triage, or other narrow interview fields.

S-MINDS uses a number of voice-independent speech recognition engines with the usage dependent on the languages and the particular domain. These engines include Nuance 8.5[2], SRI EduSpeak 2.0[3], and Entropic's HTK-based engine.[4] There is a dialog/translation creation tool that allows us to compile and run our created dialogs with any of these engines. This allows our developers to be free from the nuances of any particular engine that is deployed. S-MINDS uses a combination of grammars and language models with these engines, depending on the task and the availability of training data. In the case of the system described in this document, we were using Nuance 8.5 for both English and Korean speech recognition.

We use our own semantic parser, which identifies keywords and phrases that are tagged by the user; these in turn are fed into an interpretation engine. Because of the limited context, we can achieve high translation accuracy with the interpretation engine. However, as the name suggests, this engine does not directly translate users' utterances but interprets what they say and paraphrases their statements. Finally, we use a voice generation system (which splices human recordings) along with the Festival TTS engine to output the translations. This has been recently replaced by the Cepstral TTS engine.

Additionally, S-MINDS includes a set of tools to modify and augment the existing system with additional words and phrases in the field in a matter of a few minutes.

The initial task given to us was a medical disaster recovery scenario that might occur near an American military base in Korea. We were given about 270 questions and an additional 90 statements that might occur on the interviewer side. Since our system is an interview-driven system (sometimes referred to as "1.5-way"), the second-language person is not given the option of initiating conversations. The questions and statements given to us covered several domains related to the task above, including medical triage, force protection at the installation gate, and some disaster recovery questions. In addition to the 270 assigned questions, we created 120 of our own in order to make the domains more complete.

3.1 Data Collection

Since we assumed that we could internally generate the English language data used to ask the question but not the language data on the Korean side, our entire focus for the data collection task was on Korean. As such, we collected about 56,000 utterances from 144 people to answer the 390 questions described above. This data collection was conducted over the course of 2 months via a telephone-based computer system that the native Korean speakers could call. The system first introduced the purpose of the data collection and then presented the participants with 12 different scenarios. The participants were then asked a subset of the questions after each of the scenarios. One advantage of the phone-based system – in addition to the savings in administrative costs – was that the participants were free to do the data collection any time during the day or night, from any location. The system also allowed participants to hang up and call back at a later time. The participants were paid only if they completed all the scenarios.

Of this data, roughly 7% was unusable and was thrown away. Another 31% consisted of one-word answers (like "yes"). The rest of the data consisted of utterances 2 to 25 words long. Approximately 85% of the usable data was used for training; the remainder was used for testing.

The transcription of the data started one week after the start of the data collection, and we started building the grammars three weeks later.

3.2 System Development

We have an extensive set of tools that allow non-specialists, with a few days of training, to build complete mission-oriented domains. In this project, we used three bilingual college graduates who had no knowledge of linguistics. We spent the first 10 days training them and the next two weeks closely supervising their work. Their work involved taking the sentences that were produced from the data collection and building grammars for them until the "coverage" of our grammars – that is, the num-

[2] http://www.nuance.com/nuancerecognition/
[3] http://www.speechatsri.com/products/eduspeak.shtml
[4] http://htk.eng.cam.ac.uk/

ber of utterances from the training set that our system would handle – was larger than a set threshold (generally set between 80% and 90%). Because of the scarcity of Korean-language data, we built this system based entirely on grammar language models rather than statistical language models. Grammars are generally more rigid than statistical language models, and as such grammars tend to have higher in-domain accuracy and much lower out-of-domain accuracy[5] than statistical language models. This means that the system performance will depend greatly upon on how well our grammars cover the domains.

The semantic tagging and the paraphrase translations were built simultaneously with the grammars. This involved finding and tagging the semantic classes as well as the key concepts in each utterance. Frame-based translations were performed by doing concept and semantic transfer. Because our tools allowed the developers to see the resulting frame translations right away, they were able to make fixes to the system as they were building it; hence, the system-building time was greatly reduced.

We used about 15% of the collected telephone data for batch testing. Before deployment, our average word accuracy on the batch results was 92.9%. The translation results were harder to measure directly, mostly because of time constraints.

3.3 System Testing

We tested our system with 11 native Korean speakers, gathering 968 utterances from them. The results of the test are shown in Table 1. Most of the valid rejected utterances occurred because participants spoke too softly, too loudly, before the prompt, or in English. Note that there was one utterance with bad translation; that and a number of other problems were fixed before the actual field testing.

5 Note that there are many factors effecting both grammar-based and statistical language model based speech recognition, including noise, word perplexity, acoustic confusability, etc. The statement above has been true with some of the experiments that we have done, but we can not claim that it is universally true.

Category	Percentage
Total Recognized Correctly	82.0%
Total Recognized Incorrectly	5.8%
Total Valid Rejection	8.0%
Total Invalid Rejected	4.1%
Total unclear translations	0.1%

Table 1: Korean-to-English system testing results for the 11 native Korean speakers.

4 Experimental Setup

A military medical group used S-MINDS during a medical training exercise in January 2005 in Carlsbad, California. The testing of speech translation systems was integrated into the exercise to assess the viability of such systems in realistic situations. The scenario involved a medical aid station near the front lines treating badly injured civilians. The medical facilities were designed to quickly triage severely wounded patients, provide life-saving surgery if necessary, and transfer the patients to a safer area as soon as possible.

4.1 User Training

Often the success or failure of these interactive systems is determined by how well the users are trained on the systems' features.

Training and testing on S-MINDS took place from November 2004 through January 2005. The training had three parts: a system demonstration in November, two to three hours of training per person in December, and another three-hour training session in January. About 30 soldiers were exposed to S-MINDS during this period. Because of the tsunami in Southeast Asia, many of the people who attended the November demo and December training were not available for the January training and the exercise. Nine service members used S-MINDS during the exercise. Most of them had attended only the training session in January.

4.2 Test Scenarios

Korean-speaking 'patients' arrived by military ambulance. They were received into one of three tents where they were (notionally) triaged, treated, and prepared for surgery. The tents were about 20 feet wide by 25 feet deep, and each had six to eight cots for patients. The tents had lights and electricity.

The environment was noisy, sandy, and 'bloody.' The patients' makeup coated our handsets by the end of the day. There were many soldiers available to help and watch. Nine service members used S-MINDS during a four-hour period.

All of the 'patients' spoke both English and Korean. A few 'patients' were native Korean speakers, and two were American service members who spoke Korean fairly fluently but with an accent. The 'patients' were all presented as severely injured from burns, explosions, and cuts and in need of immediate trauma care.

The 'patients' were instructed to act as if they were in great pain. Some did, and they sounded quite realistic. In fact, their recorded answers to questions were sometimes hard for a native Korean speaker to understand. The background noise in the tents was quite loud (because of the number of people involved, screaming patients and close quarters). Although we did not directly measure the noise; we estimate it ranged from 65 to 75 decibels.

4.3 Physical and Hardware Setup

S-MINDS is a flexible system that can be configured in different ways depending on the needs of the end user. Because of the limited time available for training, the users were trained on a single hardware setup, tailored to our understanding of how the exercises would be conducted. Diagrams available before the exercises showed that each tent would have a "translation station" where Korean-speaking patients would be brought. The experimenters (two of the authors) had expected that the tents would be positioned at least 40 feet apart. In reality, the tents were positioned about 5 feet apart, and there was no translation station.

Our original intent was to use S-MINDS on a Sony U-50 tablet computer mounted on a computer stand with a keyboard and mouse at the translation station, and for a prototype wireless device – based on a Bluetooth-like technology to eliminate the need for wires between the patient and the system – that we had built previously. However, because of changes in the conduct of the exercise, the experimenters had to step in and quickly set up two of the S-MINDS systems without the wireless system (because of the close proximity of the tents) and without the computer stands. The keyboards and mice were also removed so that the S-MINDS systems could be made portable. The medics worked in teams of two; one medic would hold the computer and headset for the injured patient while the other medic conducted the interview.

5 Results

The nine participants used our system to communicate with 'patients' over a four-hour period. We analyzed qualitative problems with using the system and quantitative results of translation accuracy.

5.1 Problems with System Usage

We observed a number of problems in the test scenarios with our system. These represent some of the more common problems with the S-MINDS system. The authors suspect these may be endemic of all such systems.

5.1.1 Inadequate Training on the System

Users were trained to use the wireless units, which interfered with each other when used in close proximity. For the exercise, we had to set up the units without the wireless devices because the users had not been trained on this type of setup. As a result, service members were forced to use a different system from the one they were trained on.

Also, the users had difficulty navigating to the right domain. S-MINDS has multiple domains each optimized for a particular scenario (medical triage, pediatrics, etc.), but the user training did not include navigation among domains.

5.1.2 User Interface Issues

The user interface and the system's user feedback messages caused unnecessary confusion with the interviewers. The biggest problem was that the system responded with, "I'm sorry, I didn't hear that clearly" whenever a particular utterance wasn't recognized. This made the users think they should just repeat their utterance over and over. In fact, the problem was that they were saying something that were out of domain or did not fit any dialogs in S-MINDS, so no matter how many times

they repeated the phrase, it would not be recognized. This caused the users significant frustration.

5.2. Quantative Analysis

During the system testing, there were 363 recorded interactions for the English speakers. Unfortunately, the system was not set up to record the utterances that had a very low confidence score (as determined by the Nuance engine), and the user was asked to repeat those utterances again. Here is the rough breakdown for all of the English interactions:

- 52.5% were translated correctly into Korean
- 34.2% were rejected by the system
- 13.3% had misrecognition or mistranslation errors

This means that S-MINDS tried to recognize and translate 65.8% of the English utterances and of those 79.8% were correctly translated. A more detailed analysis is presented in Figure 1.

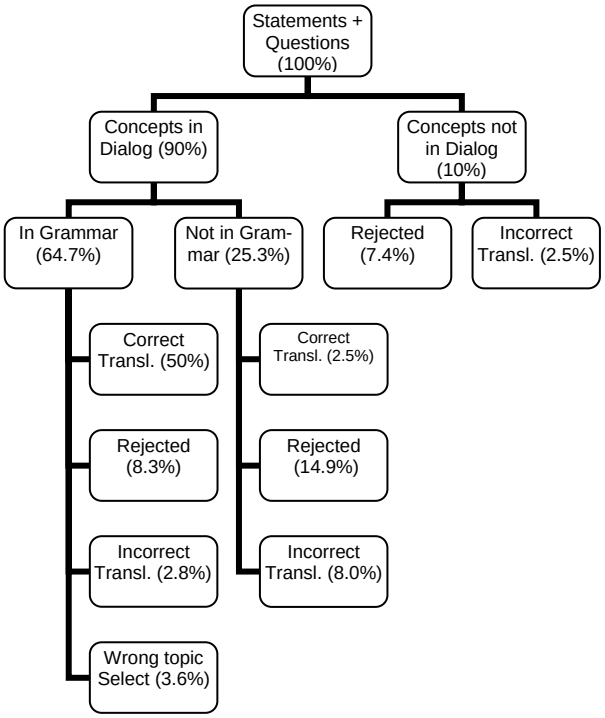

Figure 1: Detailed breakdown for the English utterances and percentage breakdown for each category.

The Korean speakers' responses to each of the questions that were recognized and translated are analyzed in Figure 2. Note that the accuracy for the non-rejected responses is 78.3%.

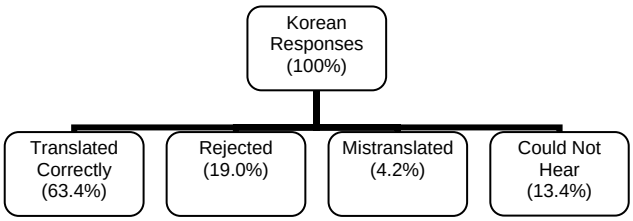

Figure 2: Detailed breakdown of the recognition for the Korean utterances and percentage breakdown for each category.

6 Discussion

Although these results are less than impressive, a close evaluation pointed to three areas where a concentration of effort would significantly improve translation accuracy and reduce mistranslations. These areas were:

1) Data collection with English speakers to increase coverage on the dialogs.
   a) 34% of the things the soldiers said were things S-MINDS was not designed to translate.
   b) We had assumed that our existing English system would have adequate coverage without any additional data collection.
2) User verification on low-confidence results.
3) Improved feedback prompts when a phrase is not recognized; for example:
   a) One user said, "Are you allergic to any allergies?" three times before he caught himself and said, "Are you allergic to any medications?"
   b) Another user said, "How old are you?" seven times before realizing he needed to switch to a different domain, where he was able to have the phrase translated.
   c) Another user repeated, "What is your name?" nine times before giving up on the phrase (this phrase wasn't in the S-MINDS Korean medical mission set).

Beyond improving the coverage, the system's primary problem seemed to be in the voice user interface since even the trained users had a difficult time in using the system.

The attempt at realism in playing out a high-trauma scenario may have detracted from the effectiveness of the event as a test of the systems' abilities under more routine (but still realistic) conditions.

7 New Results

Based on the results of this experiment, we had a secondary deployment in a medical setting for a very similar system.

We applied what we had learned to that setting and achieved better results in a few areas. For example:

1. Data collection in English helped tremendously. S-MINDS recognized about 40% more concepts than it had been able to recognize using only grammars created by subject-matter experts.
2. Verbal verification of the recognized utterance was added to system, and that improved the user confidence, although too much verification tended to frustrate the users.
3. Feedback prompts were designed to give more specific feedback, which seemed to reduce user frustration and the number of mistakes.

Overall, the system performance seemed to improve. We continue to gather data on this task, and we believe that this is going to enable us to identify the next set of problems that need to be solved.

8 Acknowledgement

This research was funded in part by the LASER ACTD. We specially wish to thank Mr. Pete Fisher of ARL for his generous support and his participation in discussions related to this project.

References

Andrulis Dennis, Nanette Goodman, Carol Pryor (2002), "What a Difference an Interpreter Can make" April 2002. Access Project, www.accessproject.org/downloads/c_LEPreport ENG.pdf

Bangalore, S. and G. Riccardi, (2001), "A Finite State Approach to Machine Translation," North American ACL 2001, Pittsburgh.

Cohen, L, F. Rivara, E. K. Marcuse, H. McPhillips, and R. Davis, (2005), "Are Language Barriers Associated With Serious Medical Events in Hospitalized Pediatric Patients?**"**, *Pediatrics*, September 1, 2005; 116(3): 575 - 579

Flores Glenn, (2005), "The Impact of Medical Interpreter Services on the Quality of Health Care: A Systematic Review," *Medical Care Research and Review*, Vol. 62, No. 3, pp. 255-299

Florian M., et. al. (2002), "Enhancing the Usability and Performance of NESPOLE!: a Real-World Speech-to-Speech Translation System", HLT 2002, San Diego, California U.S., March 2002.

Isotani, R., Kiyoshi Yamabana, Shinichi Ando, Ken Hanazawa, Shin-ya Ishikawa and Ken-ichi ISO (2003), "Speech-to-Speech Translation Software on PDAs for Travel Conversation," NEC Research and Development, Apr. 2003, Vol.44, No.2.

O'Leary and Hampers (2003) "**T**he Truth About Language Barriers: One Residency Program's Experience," *Pediatrics*, May 1, 2003; 111(5): pp. 569 - 573.

Keiji Yasuda, Eiichiro Sumita, Seiichi Yamamoto, Genichiro Kikui, Masazo Yanagida, "Real-Time Evaluation Architecture for MT Using Multiple Backward Translations," *Recent Advances in Natural Language Processing*, pp. 518-522, Sep., 2003

Wahlster, W. (2000), *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer.

Waibel, A., (1996), "Interactive Translation of Conversational Speech," *IEEE Computer*, July 1996, 29-7, pp. 41-48.

Woszczyna, et al., (1993), "Recent Advances in JANUS: A Speech Translation System," *DARPA Speech and Natural Language Workshop 1993*, session 6 – MT.

Zhang, Ying, (2003), "Survey of Current Speech Translation Research," Found on Web: http://projectile.is.cs.cmu.edu/research/public/talks/ speechTranslation/sst-survey-joy.pdf