RAPID DEVELOPMENT OF A SPEECH TRANSLATION SYSTEM FOR KOREAN

Farzad Ehsani[1], Jim Kimzey[1], Demitrios Master[1], Hunil Park[2], Karen Sudre[1],
{farzad,jkimzey,karen,dlm}@sehda.com, phunil@hotmail.com

[1] Sehda, Inc., [2] Independent Consulting

ABSTRACT

S-MINDS is a hand held, 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 been able to build an English-to-Korean medical interviews system in 4 months time. The formal system evaluation indicated a translation accuracy of 79.8% (for English) and 78.3% (for Korean) for all non-rejected utterances. In this paper, we will discuss the performance of S-MINDS as well as a complete analysis of our results including the various problems we encountered during deployment.

1. INTRODUCTION

Due to increasing globalization, translation of written and spoken language is becoming more and more important every day. There has been considerable effort in text translation research dating as far back as the ‘50s, but research on speech translation systems began only about 15 years ago. Speech translation still is typically done by human interpreters with varying degrees of competency and fluency. The quality of the resulting translation varies. Near one end of the spectrum are highly trained simultaneous UN translators and Language Line (a service for over-the-telephone interpretation); near the other end are friends and family members who might interpret for a patient during a hospital visit mostly without much training or even fluency.

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 deficiencies 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

There are two major schools of thought on machine translation: knowledge-based and statistical MT (SMT). Since the early 1990s, empirical approaches to MT have sought to produce appropriate translations automatically from parallel data. The development of these MT approaches, and in particular SMT, is consistent with a general trend in natural language processing toward quantitative empirical methods, which have been made possible by the increasing availability of large electronic text corpora. However, the scarcity of bilingual corpora for many language pairs has been a barrier to SMT.

Unlike other systems, that try to solve the speech translation problem with the assumption that there is a moderate amount of data available [1,2,3,4,5], 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.

3. SYSTEM DESCRIPTION

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 [6], SRI EduSpeak 2.0 [7], and Entropic’s HTK-based engine [8]. 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 the 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.

We use our own semantic parser which identifies keywords and phrases that are tagged by the user; these in turn are fed into a slot-filling paraphrase translation engine. Because of the limited context, we can achieve high translation accuracy with the translation engine. Finally we use a Voice Generation system (which splices human recordings) along with the Festival TTS engine to output the translations.

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

1-4244-0469-X/06/$20.00 ©2006 IEEE

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 subtopics 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 mission sets more complete.

3.1 Data Collection

Since we assumed that we could internally generate the English language used to ask the question but not the language 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 2 months via a telephone system which the native speakers could call into. The system first introduced the purpose of the data collection and then presented them with 12 different scenarios. The participants were then asked a subset of the questions after each of the scenarios. The advantage of a phone-based system, other than 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 were 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 useable 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. We have an extensive set of tools that allows non-specialists, with a few days of training, to build complete mission sets. 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 number of utterances from the training set that our system would handle – was larger than a set threshold (generally set between 80 to 90 percent). Due to lack of sufficient data, we built this system based entirely on grammars rather than on a combination of grammars and statistical language models. The semantic tagging and the paraphrase translations were built simultaneously with the grammars. Because our tools allowed the developers to see the resulting 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 percent of the collected telephone data for batch testing. Before deployment, our average word accuracy on the batch results was 92.9%.

3.2 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.

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

Table 1: Korean-to-English system testing results for the eleven 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. Due to 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 (due to 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. Due to 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 the 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 blue-tooth-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 (due to 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. Note that we were using the Nuance 8.5 Recognition Engine for both English and Korean for this evaluation.

5. RESULTS

The nine participants used our system to communicate with 'patients' over a four-hour period. We analyzed both 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 describe 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 interfere 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 subtopic. S-MINDS has multiple subtopics each optimized for a particular scenario (medical triage, pediatrics, etc.), but the user training did not include navigation among subtopics.

5.1.2 User Interface Issues

Our user interface and the user feedback were causing 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 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 setup 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 percent were translated correctly into Korean
- 34.2 percent 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 speaker utterances and of those 79.8% were correctly translated. A more detailed analysis is presented in Figure 1.

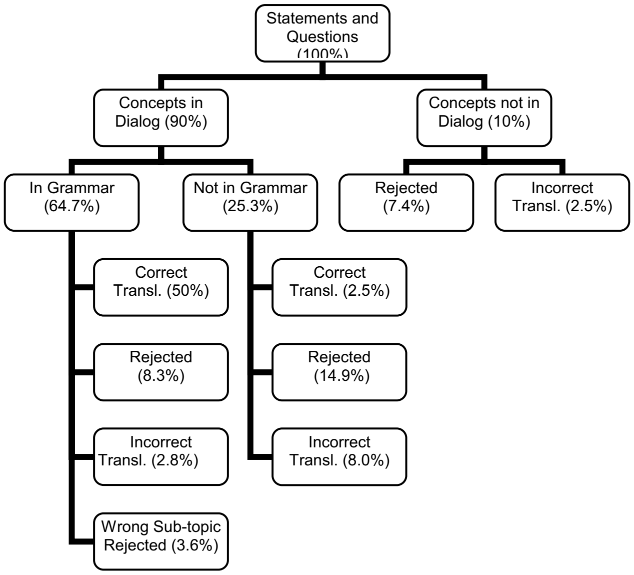

Figure 1: Detailed breakdown for the English utterances and percentage breakdown for each category.

The Koreans' 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%.

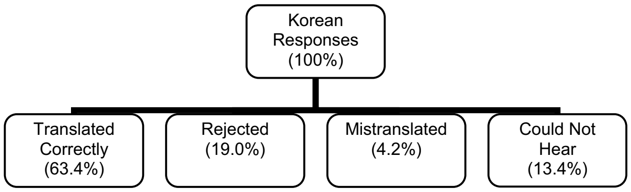

Figure 2: Detailed breakdown of the recognition for the Korean utterances and percent breakdown for each category.

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 are:

1) Data collection with English speakers to increase coverage on the dialogs.
   a) 34 percent 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 subtopic, 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. ACKNOWLEDGEMENT

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

8. REFERENCES

[1] Waibel, et al. “Interactive Translation of Conversational Speech,” *IEEE Computer,* 29-7, P. 41-48, 1996.

[2] Isotani, R., et. Al., “Speech-to-Speech Translation Software on PDAs for Travel Conversation,” *NEC Research and Development,* Vol.44, No.2 Apr. 2003.

[3] Wahlster, W., *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 2000.

[4] Florian M., “Enhancing the Usability and Performance of NESPOLE!: a Real-World Speech-to-Speech Translation System”, *HLT 2002*, San Diego, California, March 2002.

[5] Zhou, B. et al. "Two-way Speech-to-Speech Translation on Handheld Devices", *Int. Conf. of Spoken Language Processing (ICSLP), Korea,* Oct. 2004.

[6] http://www.nuance.com/nuancerecognition/

[7] http://www.speechatsri.com/products/eduspeak.shtml

[8] http://htk.eng.cam.ac.uk/