S-MINDS 2-Way Speech-to-Speech Translation System

Farzad Ehsani, Jim Kimzey, Demitrios Master, Karen Sudre, David Domingo
Engineering Department
Sehda, Inc.
Mountain View, CA 94043
{farzad, jkimzey, dlm, karen, ddomingo}@sehda.com

Hunil Park
Independent Consultant
Seoul, Korea
phunil@hotmail.com

Abstract

Sehda's 2-way speech translation system, S-MINDS, interprets between provider and patient in routine medical interactions with very high accuracy. Optimizing the system for new tasks or languages requires very little data. New developments include a hybrid translation approach that allows participants to say complex or out-of-domain utterances, the expansion of hands-free functionality, and the ability to deliver the most urgent expressions instantaneously.

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.

The most common hospital interactions are interview-style dialogs where the provider's and patient's utterances are simple and relatively predictable. Sehda's speech translation system, S-MINDS, focuses on translating in such situations with extremely high accuracy.

One key difference between S-MINDS and other speech translation systems is the amount of data required in development. Most other systems depend on a moderate amount of domain-specific data being available. If the data is not already available, it is extremely time- and labor-intensive for a developer to collect enough realistic data to effectively model a pure SMT system – even if the developer has direct access to a group of actual users for whom its system is being optimized.

For this and other reasons, Sehda focuses on rapid building and deployment of speech translation systems for tasks or languages where little or no data is available.

2 System Description

This section describes the speech recognition, translation, speech generation, interface and hardware components that make up S-MINDS.

2.1 Speech Recognition

S-MINDS uses a number of voice-independent automated speech recognition (ASR) engines, with the usage dependent on the languages and the particular domain. These engines include Nuance 8.5[i], SRI EduSpeak 2.0[ii], and Entropic's HTK-based engine.[iii]

Sehda's (internal) dialog/translation creation tools allow developers to compile and run new dialogs with any ASR engine so they do not have to be encumbered by the nuances of any particular engine.

2.2 Translation

S-MINDS processes ASR output using a combination of grammars and language models that is selected based on the task and the availability of training data.

S-MINDS first employs a semantic parser to extract the essential words and phrases from the ASR output. This information is then fed into

Sehda's proprietary interpretation engine, which matches the information against a finite set of concepts in the specified domain. The resulting translation is extremely accurate – often more accurate than the ASR output itself. However, as the name suggests, this engine does not directly translate users' utterances but interprets what they say and paraphrases their statements.

2.3 Speech Generation

S-MINDS uses its own voice generation system, which splices human recordings, to output most translations. If recordings do not exist for a word or phrase, S-MINDS generates the speech using a text-to-speech (TTS) engine.

S-MINDS includes a set of tools by which users can modify and augment the existing system with additional words and phrases in the field in a matter of a few minutes.

2.4 Interface

A variety of interface features make S-MINDS particularly easy to use in a hospital environment.

Most S-MINDS functions can be performed hands-free and eyes-free via a voice user interface (VUI) so the provider can focus on the patient and the operation of hospital equipment.

A picture viewer allows digital images to be displayed to aid communication with the patient and add clarity to the log.

Verbal or on-screen verification can be employed (with adjustable upper and lower thresholds) to put an additional check on recognition accuracy.

2.5 Hardware

A complete S-MINDS system contains three main hardware components: a Windows XP computer with S-MINDS software installed; a headset microphone, which the healthcare provider uses to control S-MINDS and communicate with the patient; and a telephone handset, which the patient uses to communicate with the provider.

3 Current Developments

Under a contract with DARPA, Sehda has developed a more interactive system using a combination of SMT and interpretation engines. This allows the users to speak more freely. If an utterance is too complex or too far 'out of domain' to be handled by the interpretation engine, S-MINDS falls back to the SMT engine, which returns a fairly reliable word-for-word translation of the ASR output.

The VUI in S-MINDS is being enhanced to include nearly all system control functions, reducing the need to change settings manually. In addition, users will be able to deliver some urgent expressions (such as "Hold still" or "You can breathe") instantaneously without saying a 'hotword' first.

---

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

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

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