Interactive Error Resolution Strategies for Speech-to-Speech Translation Systems

Rohit Kumar, Matthew Roy, Sankaranarayanan Ananthakrishnan, Sanjika Hewavitharana, Frederick Choi
Speech, Language and Multimedia Business Unit
Raytheon BBN Technologies
Cambridge, MA, USA
{rkumar, mroy, sanantha, shewavit, fchoi}@bbn.com

Abstract[1]

In this demonstration, we will showcase BBN's Speech-to-Speech (S2S) translation system that employs novel interaction strategies to resolve errors through *user-friendly* dialog with the speaker. The system performs a series of analysis on input utterances to detect out-of-vocabulary (OOV) named-entities and terms, sense ambiguities, homophones, idioms and ill-formed inputs. This analysis is used to identify potential errors and select an appropriate resolution strategy. Our evaluation shows a 34% (absolute) improvement in cross-lingual transfer of erroneous concepts in our English to Iraqi-Arabic S2S system.

1 Introduction

Great strides have been made in Speech-to-Speech (S2S) translation systems that facilitate cross-lingual spoken communication (Stallard et. al., 2011). However, in order to achieve broad domain coverage and unrestricted dialog capability, S2S systems need to be transformed from passive conduits of information to active participants in cross-lingual dialogs. These active participants must detect key causes of communication failures and recover from them in an efficient, user-friendly manner.

Disclaimer: This paper is based upon work supported by the DARPA BOLT Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Distribution Statement A (Approved for Public Release, Distribution Unlimited)

Our ongoing work on *eyes-free* S2S systems is focused on detecting three types of errors that affect S2S systems. First, out-of-vocabulary (OOV) words are misrecognized as phonetically similar words that do not convey the intended concept. Second, ambiguous words such as homophones and homographs often lead to recognition and translation errors. Also, unseen idioms produce erroneous literal translations. Third, user errors such as mispronunciations and incomplete utterances lead to ASR errors. We will demonstrate our interactive error resolution strategies to recover from each of these error types.

Section 2 presents our system architecture. Section 3 describes nine interactive error resolution strategies that are the focus of this demonstration. An evaluation of our English to Iraqi-Arabic S2S system is summarized in Section 4.

2 System Architecture

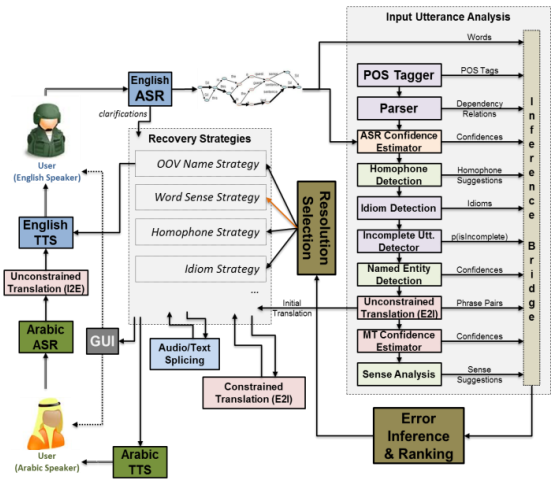

Figure 1: BBN S2S System with Error Recovery in English to Iraqi-Arabic direction

Figure 1 shows the architecture of our two-way

Proceedings of the SIGDIAL 2013 Conference, pages 142–144,
Metz, France, 22-24 August 2013.

English to Iraqi-Arabic S2S translation system. In the English to Iraqi direction, the initial English ASR hypothesis and its corresponding translation are processed through a series of analysis (e.g. parsing, sense disambiguation) and error detection (e.g. ASR/MT confidence, Homophone/Idiom/Named-Entity detection) modules. A detailed discussion on the various error detection modules can be found in Prasad et. al. (2012). A novel *Inference Bridge* data structure supports storage of these analyses in an interconnected and retraceable manner. The potential erroneous spans are identified and ranked in an order of severity using this data structure.

Based on the top ranked error, one of nine error resolution strategies (discussed in Section 3), is selected and executed. Each strategy is composed of a sequence of steps which include actions such as TTS output, user input processing, translation (unconstrained or constrained) and other error type specific operations. This sequence is hand-crafted to efficiently recover from an error. Following a multi-expert design (Turunen and Hakulinen, 2003), each strategy represents an error-specific expert.

3 Error Resolution Strategies

Figure 2 illustrates the sequence of steps for the nine interaction strategies used by our system.

The *OOV Name* and *ASR Error* strategies are designed to interactively resolve errors caused by OOV words (names and non-names) as well as other generic ASR and MT errors. When a span of words is identified as an OOV named-entity, the user is asked to confirm whether the audio segment corresponding to those words is a name. Upon user confirmation, the audio segment is spliced into the output target language utterance. This is based on the principle that audio segments containing names are understandable across languages.

In the case where a generic erroneous span is detected, the user is asked to rephrase the utterance. This strategy is suitable for handling multiple error types including OOVs, mispronunciations, and generic ASR/MT errors. Additionally, the *ASR Errors* strategy has been designed to

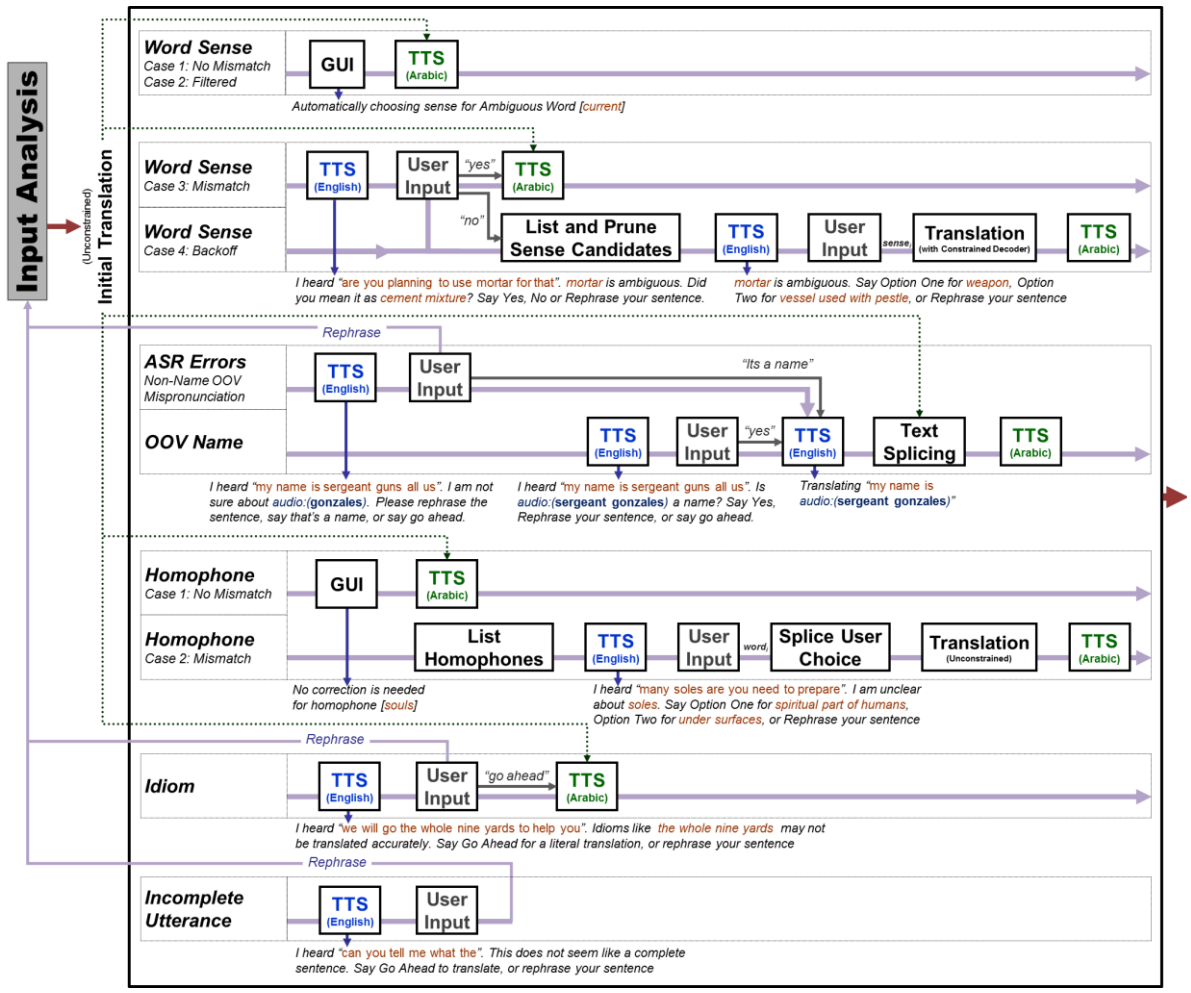

Figure 2. Interaction Strategies for Error Resolution

capture a large fraction of the OOV name false negatives (i.e. missed detections) by allowing the user to indicate if the identified erroneous span is a name. Because of the confusability between the errors handled by these two strategies, we have found it beneficial to maintain reciprocity between them to recover from all the errors handled by each of these strategies.

The four *Word Sense* (WS) disambiguation strategies resolve sense ambiguity errors. The underlying principle behind these strategies is that the sense of an ambiguous word must be confirmed by at least two of four possible independent sources of evidence. These four sources include (a) the translation system (sense lookup corresponding to phrase pair associated with the ambiguous word), (b) a list of source-language contextual keywords that disambiguate a word, (c) the sense predicted by a sense-disambiguation model and (d) sense specified by the user. Besides the objective to minimize user effort, these multiple sources are necessary because not all of them may be available for every ambiguous word. *Case 1: No Mismatch* strategy corresponds to the case where sources (a) and (c) agree. *Case 2: Filtered* strategy corresponds to the case where (a) and (b) agree. In both of these cases, the system proceeds to present the translation to the Arabic speaker without performing any error resolution. If these three sources are unable to resolve the sense of a word, the user is asked to confirm the sense identified by source (a) as illustrated in *Case 3: Mismatch* strategy. If the user rejects that sense, a list of senses is presented to the user (*Case 4: Backoff* strategy). The user-specified sense then drives constrained decoding to obtain an accurate translation.

Albeit simpler, the two homophone resolution strategies mimic the word sense disambiguation strategies in principle and design. The observed homophone variant produced by the ASR must be confirmed either by a homophone disambiguation model (*Case 1: No Mismatch*) or by the user (*Case 2: Mismatch*). The input utterance is modified (if needed) by substituting the resolved homophone variant in the ASR output which is then translated and presented to the Arabic speaker.

Strategies for resolving errors associated with idioms and incomplete utterances primarily rely on informing the user about these errors and eliciting a rephrasal. For idioms, the user is also given the choice to force a literal translation when appropriate.

Following a mixed-initiative design, at all times, the user has the ability to rephrase their utterance as well as to force the system to proceed with the current translation. This allows the user to override system false alarms whenever suitable. The interface also allows the user to repeat the last system message which is helpful for comprehension of some of the synthesized system prompts for unfamiliar users.

4 Summary of Evaluation

Our S2S system equipped with the error resolution strategies discussed in the previous section was evaluated on 103 English utterances (25 unique utterances repeated by multiple speakers). Each utterance was designed to elicit one of the error types listed in Section 1.

The ASR word error rate for these utterances was 23%. The error detection components were able to identify 59% of these errors and the corresponding error resolution strategies were correctly triggered.

The erroneous concepts in 13 of the 103 utterances (12.6%) were translated without any error. Using the error resolution strategies, an additional 34% of the erroneous concepts were accurately translated. This increased precision is achieved at the cost of user effort. On average, the strategies needed 1.4 clarifications turns per utterance.

Besides focusing on improving the error detection and resolution capabilities, we are currently working on extending these capabilities to two-way S2S systems. Specifically, we are designing interactive strategies that engage both users in eyes-free cross-lingual communication.

References

David Stallard, Rohit Prasad, Prem Natarajan, Fred Choi, Shirin Saleem, Ralf Meermeier, Kriste Krstovski, Shankar Ananthakrishnan, and Jacob Devlin. 2011. *The BBN TransTalk Speech-to-Speech Translation System*. Speech and Language Technologies, InTech, 31-52

Rohit Prasad, Rohit Kumar, Sankaranarayanan Ananthakrishnan, Wei Chen, Sanjika Hewavitharana, Matthew Roy, Frederick Choi, Aaron Challenner, Enoch Kan, Arvind Neelakantan, and Premkumar Natarajan. 2012. *Active Error Detection and Resolution for Speech-to-Speech Translation*. Intl. Workshop on Spoken Language Translation (IWSLT), Hong Kong

Markku Turunen, and Jaakko Hakulinen, 2003. *Jaspis - An Architecture for Supporting Distributed Spoken Dialogues*. Proc. of Eurospeech, Geneva, Switzerland