Error-Tolerant Speech-to-Speech Translation

Rohit Kumar rkumar@bbn.com
Sanjika Hewavitharana shewavit@bbn.com
Nina Zinovieva nzinovie@bbn.com
Matthew E. Roy mroy@bbn.com
Edward Pattison-Gordon epgordon@bbn.com
Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA, USA, 02138

Abstract

Recent efforts to improve two-way speech-to-speech translation (S2S) systems have focused on developing error detection and interactive error recovery capabilities. This article describes our current work on developing an eyes-free English-Iraqi Arabic S2S system that detects ASR errors and attempts to resolve them by eliciting user feedback. Here, we report improvements in performance across multiple system components (ASR, MT and error detection). We also present a controlled evaluation of the S2S system that quantifies the effect of error recovery on user effort and conversational goal achievement.

1. Introduction

Over the past decade, considerable progress has been made in developing usable, two-way speech-to-speech (S2S) translation systems that enable real time cross-lingual spoken communication [1][2]. Conventionally, S2S systems comprise a pipeline of three speech and language technology components: automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS).

While each of these components technologies have continued to improve in performance, each is data-driven and its performance will degrade when faced with novel vocabulary. For example, large-vocabulary ASR systems are incapable of recognizing out-of-vocabulary (OOV) words, MT systems cannot translate unseen source words and TTS often mispronounces novel words (often high-value concepts like names and technical terminology). A combination of these deficiencies can render S2S systems unusable, especially on conversational topics not well represented in the training data.

Different approaches to addressing component failures in the context of S2S systems have been explored. There have also been attempts at joint optimization of ASR and MT, as well as MT and TTS [23][24][25]. Most recently, S2S system that actively detect and recover from failures have been built and evaluated [16][26]. In this work, we report our recent efforts to extend and enhance the usability of S2S translation systems through active error detection and recovery.

To address system robustness, we have operationalized recent advances in ASR and statistical MT (SMT) into our real-time S2S system. These advances have dramatically improved ASR and SMT performance, leading to a more robust S2S pipeline, and yet they are fast enough for real-time use. To address the problem of novel vocabulary, we have developed interfaces that enable non-technical users to rapidly enrich our S2S system with new words and phrases that are relevant to emerging use cases. Finally, we detect potential ASR errors before they are amplified by being sent through the SMT and TTS components. A simple in-

teraction strategy then allows users to correct the translation in the event of an error. Section 3 evaluates an end-to-end S2S system augmented with this capability.

2. BBN's Speech-to-Speech Translation System

2.1. Architecture

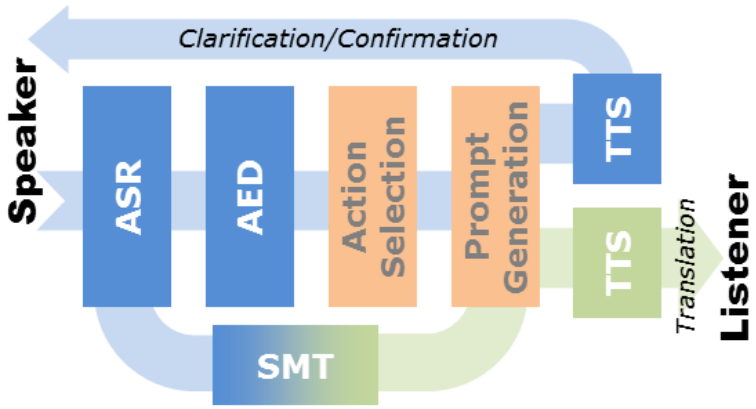

Figure 1. A unidirectional S2S system pipeline

Figure 2. BBN's S2S System with SuperMic

Two-way S2S systems comprise a pair of symmetric unidirectional pipelines of ASR, SMT and TTS components. Each pipeline serves as a communication channel in one direction. Figure 1 depicts one of these pipelines for our S2S system. Spoken input is converted by the ASR into a word lattice and a 1-best, whole-sentence hypothesis. The SMT produces the target language translation of the 1-best ASR hypothesis. The ASR error detector (AED) analyzes the word lattice to identify and rank potentially erroneous spans in the ASR hypothesis. This analysis, along with the current discourse state, drives a rule-based action selection module, which decides whether the translation should be transmitted to the listener or a clarification prompt should be presented to the speaker.

Our English-Iraqi Arabic S2S system uses ASR, AED and SMT components developed at BBN and off-the-shelf English and Iraqi Arabic TTS. The system is deployed on an off-the-shelf mobile computing device shown in Figure 2. The system is self-contained and runs the entire pipeline fully on this device. To enable eyes-free use, the platform is augmented with a pair of proprietary audio I/O devices, also shown in Figure 2. Each device, referred to as a "SuperMic", offers a push-to-talk button, a high-quality close-talking microphone and a speaker.

2.2. System Components

Automatic Speech Recognition: The baseline ASR system is built using data from the DARPA TransTac English-Iraqi Arabic parallel two-way spoken dialogue collection [3]. It is based on the BBN Byblos system, which uses a multi-pass decoding strategy where models of increasing complexity are used in successive passes in order to refine the recognition hypotheses [4]. Speech is modeled as the output of context-dependent phonetic hidden Markov models (HMMs), whose outputs are mixtures of multi-dimensional diagonal Gaussians. Byblos uses various forms of parameter tying, including state tied mixture (STM) triphone models and state clustered tied mixture (SCTM) quinphone models. The models were trained on a set of acoustic features, part of which were obtained by using neural networks as described ahead.

Acoustic features: We use neural networks (NNs) to generate stacked bottleneck (SBN) features [5][6]. The SBN structure contains two NNs. The input features of the first NN are 24 critical-band energies obtained with a Mel filter-bank, with online mean and variance normalization applied. 15 frames of these features are stacked and a Hamming window multiplies the time evolution of each parameter. Finally, DCT is applied, of which 0th to 15th coefficients are retained. The first NN consists of two hidden layers, each with 1,500 nodes, followed by a bottleneck layer. The bottleneck (BN) outputs from the first NN are stacked, down-sampled, and taken as an input vector for the second NN. This second NN is also a bottleneck layer that is roughly the same size as the first NN. Both NNs were trained to classify phoneme states (5 states per phoneme). These targets were generated by forced alignment with baseline, perceptual linear prediction (PLP) models and remained fixed during the training. The final feature stream was built by concatenation of 9 frames of the PLP features together with the bottleneck layer output from the second NN. Finally, region dependent transformation (RDT) [7] is performed to estimate a discriminative feature projection to reduce the dimensionality to 46.

The English acoustic model was trained on approximately 200 hours of transcribed English speech from TransTac data, and the Iraqi Arabic acoustic model was trained on about 600 hours of transcribed speech from TransTac data. We tested ASR performance on held-out development sets, consisting of 13,074 and 3,354 utterances for Iraqi Arabic and English, respectively. Word error rate (WER) on these sets is 16.7% and 7.9%, respectively. For both languages, the WER is better than our previously reported ASR performance in this application [15], attributable to recent implementation of SBN features within Byblos.

Statistical Machine Translation: Our SMT system is based on the state-of-the-art, string-to-string hierarchical decoder described in [8]. In addition to the log-linear combination of generative components such as forward and backward rule probabilities, lexical translation probabilities, n-gram language models, etc., this system further incorporates two multi-layer NN scoring components, viz. the neural network joint model (NNJM) and the neural network lexical translation model (NNLTM). The NNJM estimates the probability of a hypothesized target word t, conditioned on both the n-1 preceding target words and an m-word source context,

centered at the source word s that t is affiliated with. The NNLTM estimates the probability that a source word s translates to a target word t (or NULL if none), given only t's source context.

These models are trained with one hidden layer, and therefore leverage the robust performance of multi-layer neural networks, but, following [8], we achieve look-up speeds on a par with n-gram language models by precomputing the trained hidden layer (avoiding feed-forward computations at run-time) and by training the model to produce approximately normalized output (avoiding costly softmax computations across the whole target vocabulary at run-time). To further speed up the decoder, we do not perform n-best hypothesis re-ranking, and we also tighten up various parameters in the decoder's beam search, leading to average decoding speeds of approximately 41 words/second.

We trained both our English-to-Iraqi Arabic (E2I) and Iraqi Arabic-to-English (I2E) systems on the Iraqi-English parallel text portions of the DARPA TransTac corpus [3], which we word-aligned using GIZA++ [9] and extracted hierarchical rules from using the method of [10]. We trained 4-gram Kneser-Ney language models on the respective target side of the corpus for each system. The NNJM is also limited to a 3-word target history, and both neural network models have an 11-word source context window. The log-linear combination of all components (including neural network models) is tuned using k-best optimization with an expected BLEU objective function [11] on held-out development data.

Table 1 shows model performance using the BLEU [12] and translation edit rate (TER) [13] metrics on a held-out blind set of the TransTac corpus. The addition of the neural network models has improved SMT performance significantly in both directions (+3 BLEU for English-to-Iraqi +2 BLEU for Iraqi-to-English), over our best-performing prior system [17].

	BLEU↑	**TER↓**
E2I	19.1	60.1
I2E	33.6	50.0

Table 1. BLEU and TER scores on held-out TransTac data

ASR Error Detection: The ASR error detector (AED) component performs a word level analysis of the 1-best ASR hypothesis. We trained a conditional random field (CRF) model with the same data used to train our ASR and using automatically generated error annotation by comparing ASR outputs to reference transcriptions. Words in the ASR hypothesis are represented using the following features: (1) ASR confidence; (2) language model perplexity; (3) whether a 1-best word is present in the (pruned) confusion network; (4) density of the corresponding slot in the confusion network; and (5) an indication if the word is commonly misrecognized. The performance of the AED component was evaluated on a collection of utterances designed to be representative of errors encountered by S2S systems [15]. Based on the ASR WERs reported earlier, we choose operating points corresponding to 1% false alarm at word level for English and 2% false alarm for Iraqi Arabic. The corresponding detection rates for AED were 36.4% and 19.8% respectively. These significant improvements over our previous operationalized AED implementation [21] are attributable to both improvements in ASR but also use of new features and classifier formulation for AED [14].

Word-level AED analysis is aggregated to identify error spans (contiguous spans of detected word-level errors). To prevent very large spans from being identified, we apply a heuristic that temporarily increases the operating threshold of the AED to localize only the highest confidence sub-span within large error spans. The results of ASR error detection are used to select an appropriate system action.

Note that the AED approach employed in this work only uses resources available for training and evaluation of the ASR which makes this approach extensible to new languages. Furthermore, features used by AED are based on rich information produced by the ASR. The need for tighter integration of the ASR error detection capability with the ASR makes a case for implementation of AED as a commonly available module within modern ASRs.

Interactive Error Recovery: Error recovery is presented as an optional module which users can choose to enable. When error recovery is enabled, action selection filters out very small error spans (<0.25 seconds long) identified by the AED analysis and ranks the spans by the maximum error confidence of their constituent tokens. Furthermore, action selection is influenced by the discourse state. For example, if three consecutive attempts fail to resolve an error, action selection by-passes error recovery. Table 2 shows three excerpts of an English speaker using the S2S system to communicate with an Iraqi Arabic speaker. Figure 3 summarizes the dialog model of error recovery.

Excerpt I. No error is detected	
Speaker	Hello, My name is Sergeant Jones. I came here to check on a woman that came in here yesterday due to bleeding after childbirth.
ASR	hello my name is sergeant jones i came here to check on a woman that came in here yesterday due to bleeding after childbirth
Translation	مرحبا آني اسمي العريف جونز آني إجيت هنا حتى أتشيك على المرة اللي إجا هنا البارحة بسبب النزيف بعد ولادة
Excerpt II. Error is detected (True Detection)	
Speaker	I'm doing good. My name is Sergeant Edwards. How are you doing today?
ASR	i'm doing good my name is sergeant **at words** how are you doing today
Clarification	*i'm doing good my name is sergeant at words how are you doing today*. May not be able to translate **[Edwards]**. Is that a name? Say *Yes*. Otherwise, say *Go Ahead*.
Speaker	Yes
ASR	yes
Translation	آني زين اسمي العريف **[Edwards]** شلونك اليوم
Excerpt II. Error is detected (False Alarm)	
Speaker	Y e a h about that; I'm gonna have to talk to my superiors and get back to you.
ASR	yeah about that i'm gonna have to talk to my superiors and get back to you
Clarification	*yeah about that i'm gonna have to talk to my superiors and get back to you*. May not be able to translate **[Y e a h]**. Is that a name? Say *Yes*. Otherwise, say *Go Ahead*.
Speaker	Go Ahead
ASR	go ahead
Translation	إي هاي آني رح أحكي ويا المسؤولين مالتي وأرجع لك

Table 2. Excerpts of interaction with the English-Iraqi Arabic S2S System

Excerpt I in Table 2 shows the S2S system behavior when no error is detected in the user input (HT). In this case, a translation is presented to the listener (ST). Excerpts II and III shows the case where the AED finds an error (a true detection in II and a false alarm in III). In each case, the speaker is asked to confirm (SC) if the audio corresponding to the top ranked error span corresponds to a name (or a concept that could be transferred as-is to the other lan-

guage). The user can respond (HR) by saying "yes" to confirm the accurate error detection and localization as shown in excerpt II, or skip the system's attempt to recover from a potential error by issuing the "go ahead" command. If error detection is confirmed, the audio segment corresponding to the error span is spliced into the translation through the source-target alignment produced by SMT. In addition to commands shown in excerpt II and III, the user can also choose to rephrase the input which is analyzed as a new input.

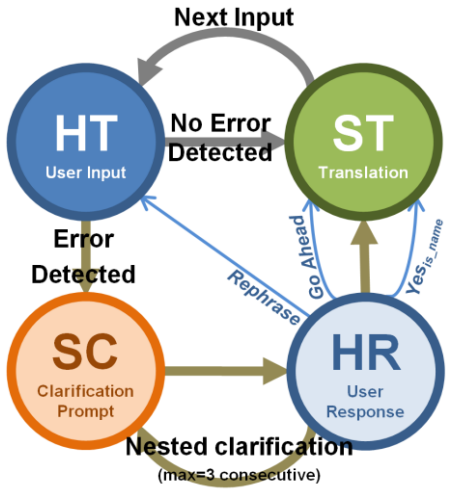

Figure 3. Dialog model of error recovery in S2S system

System behaviors in terms of action selection and prompt generation are symmetric in both directions (i.e. E2I and I2E). We note that, given the user-mediated nature of this error recovery strategy, successful cross-lingual concept transfer depends not only on component performance, but also on the appropriateness of users' responses to clarifications

3. Evaluation

3.1. Experiment Design & Data Collection

DARPA Broad Operational Language Technologies (BOLT) is a three-phase research focused on advancing the state of the art in translation technologies. Under the third phase of this program, an evaluation of error tolerant S2S systems was conducted by NIST over 5 days in January 2015.

Nine speaker pairs, each comprised of one native English speaker and one native Iraqi Arabic speaker, were tasked with communicating with one another using only a S2S system. The speakers were placed in two rooms separated by a see-through glass wall that served as a sound barrier. Each speaker pair interacted with each system over two 120-minute-long sessions. The S2S system was configured to perform ASR error detection in only one of the two sessions. The ordering of the two sessions (with and without error detection) was balanced across speaker pairs.

Each speaker pair was provided 16 conversational scenarios to be accomplished in each session. The scenarios assigned one of the two speakers a conversational driver role and the other a respondent role. The driver of the scenario was provided a conversational objective, such as ascertaining the extent of damage due to a hypothetical natural disaster. Speakers

were asked to achieve the conversational objective for each scenario in less than 8 minutes, and then to move on to the next scenario until either all scenarios were completed or the 120 minutes had expired (whichever came first). The scenarios cover a range of conversational domains relevant to military and humanitarian operations as well as everyday conversational topics like sports, family and pets.

Before using the S2S system, speakers received 30 minutes of training which included a 4-minute-long video demonstrating SuperMic features and the error recovery functionality, as well as several minutes of free-form interaction for additional practice.

Each conversational trial collected was scored by 6 independent judges along three-metrics. These metrics, listed below, used a 7 point scale ranging from -3 (strongly disagree) to +3 (strongly agree).

- Goal: The initiator of the dialog achieved his/her goal.
- Quality: The overall quality of the translations was adequate.
- Clarification: The clarification(s) helped during the dialog.

Note that the clarification scale is applicable only when the S2S system was configured to perform error recovery. In this paper, we will report average computed over the multiple judgements.

3.2. Results

Clarification→	ON	OFF
#Trials ↑	113	125
Avg. Trial Duration ↓	396.2	378.2
Turns per Trial (En) ↓	9.2	8.0
Turns per Trial (IA) ↓	9.7	8.5
#Clarifications (En)	191	-
#Clarifications (IA)	229	-
ASR Performance		
WER (En) ↓	5.7	5.6
WER (IA) ↓	6.8	7.0
OOV Rate (En) ↓	0.9	0.8
OOV Rate (IA) ↓	1.8	2.2
ASR Error Detector Performance		
Recall (En) ↑	30.4	-
Recall (IA) ↑	24.0	-
SMT Performance		
TER (E2I) ↓	58.9	58.5
TER (I2E) ↓	49.8	51.7
UNK rate (En) ↓	0.57	0.44
UNK rate (IA) ↓	0.42	0.44

Table 3. Statistics of trials collected at BOLT evaluation of BBN's S2S system & Component performance metrics

We logged 238 conversational trials over 18 sessions of use of our S2S system. S2S system configured for ASR error detection and recovery was used in 9 of these sessions during which

113 trials were logged. Table 3 provides summary statistics about these trials and reports the performance of key system components (ASR, AED, SMT) in these trials both with and without error recovery (Clarification ON or OFF).

ASR performance, reported as WER and OOV rate, is consistent across the two types of sessions. The WER is significantly lower than on the development sets, especially for Iraqi Arabic (IA). This is likely because the development set was collected in a human-mediated cross-lingual communication setting unlike the computer-mediated setting in which the S2S system is used. AED performance is reported as recall at the operating points mentioned in section 2.2.3. Despite a more conservative choice of operating point, AED is able to accurately detect a larger fraction of erroneous words for English. SMT performance is measured both as TER on 1-best ASR hypothesis and as the rate of untranslatable input words (UNKs). TER is comparable to development sets. UNK rate for English is lower than Iraqi Arabic.

While error recovery appears to not harm ASR and MT performance, we note that it achieves lower conversational throughput. Speakers have 10% fewer conversations when error recovery is enabled, and the average duration of a conversational trial increased by 5%. User effort, quantified as the number of turns spoken per trial, increases by 15% with error recovery dialogs. Also, we found that 9% of the English error clarification sub-dialogs required more than one turn (7% for Iraqi Arabic). This is an improvement over previously reported measures of error recovery cost [16], where it was found that erroneous inputs consumed 1.4 clarification turns.

Correlations	**English**		**Iraqi Arabic**	
Clarification→	**ON**	**OFF**	**ON**	**OFF**
Words per turn	-0.32	-0.35	-0.30	-0.45
OOV Rate↓	0.73	0.05	-0.28	-0.14
WER↓	0.31	-0.59	0.33	0.12
AED Recall ↑	-0.06		-0.19	
UNK Rate↓	0.31	-0.06	0.42	0.05
TER↓	-0.02	0.05	0.17	-0.20

Table 4. Correlations between metrics and user effort across different speaker pairs

We found significant variation in all of the metrics reported in Table 3 across different speaker pairs and across different domains covered in the scenarios. The standard deviation for user effort across different speaker pairs (s.d.=2.89) was higher than across different domains (s.d.=1.26). Table 4 characterizes the effect of different system components on user effort by reporting Pearson correlations of various metrics with user effort. The correlations are computed over the variations among the speakers.

Without error recovery, higher WER corresponds to fewer turns taken by the English speaker. This is likely to be indicative of fewer concepts discussed by the speaker. More English OOVs lead to increased effort when error recovery is enabled. However, use of OOVs by the Iraqi speaker does not increase the number of turns. We found that the correlation between OOV rate and AED Recall for English is positive (r=0.18), but its negative for Iraqi Arabic (r=-0.23). This is due to the difference in AED performance between the two languages. For both the speakers, more untranslatable tokens lead to an increase in effort when error recovery is enabled.

Finally, we examine the human judgements of the conversational trails. Averages for the three scales are shown in Table 5. The use of error recovery does not improve goal achievement or quality of translation. However, a closer examination of the scenarios that

were trialed with both error recovery enabled and disabled (N=112) shows that for difficult scenarios, error recovery infact improves conversational goal achievement. Difficult scenarios are defined as subset of scenarios that had below average goal achievement score in the baseline system (i.e. without error recovery).

Clarification→	ON	OFF
Goal ↑	2.27	2.40
Quality ↑	1.79	1.89
Clarification ↑	1.70	-
#Difficult Trials ↑	38	74
Goal (Difficult Trials) ↑	2.10	2.00
Goal (Easy Trials) ↑	2.47	2.73

Table 5. Judgement scores for BBN's S2S systems

Intuitively, this can be interpreted as indication that error recovery is only helpful when the users are having a conversation that the S2S system has difficulty with. For easy scenarios, error recovery does not help. One of the immediate design implications of this for S2S systems is to empower the users with the ability to turn error recovery on only when needed. However, user enabled error recovery assumes that the users are able to judge when the S2S system is having difficulty with a conversation.

4. Conclusion

Application of interactive error recovery have been investigated for multiple spoken language technologies including spoken dialog systems [18][19] and S2S systems [16][20]. Our prior work attempted to address various types of errors encountered due to imperfections of S2S system components. In contrast to that, the work presented in this paper builds on recent improvements in component performance and focuses only on errors introduced by misrecognition of input speech. This leads to simplification of interaction design as well as reduction in the cost of error recovery, quantified here in terms of user effort and time. Extending the need to minimize the cost of error recovery, we have principally chosen conservative false-alarm rate for AED based on language specific WERs. While in the work presented here we focus on an eyes-free use case, another configuration of BBN's S2S system employs a touch screen interface along with visual cues to resolve errors without requiring spoken commands.

References

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. S. Zhang, H. Yamamoto, E. Sumita and S. Yamamoto, "The ATR Multilingual Speech-to-Speech Translation System", IEEE Trans. on Audio, Speech, and Language Processing, 14(2), pp. 365- 376, 2006.

[2] D. Stallard, R. Prasad, P. Natarajan, F. Choi, S. Saleem, R. Meermeier, K. Krstovski, S. Ananthakrishnan and J. Devlin, "The BBN TransTalk Speech-to-Speech Translation System", Speech and Language Technologies, InTech, (Ed. Ivo Ipsic), pp. 31-52, 2011.

[3] F. Choi, S. Tsakalidis, S. Saleem, C. L. Kao, R. Meermeier, K. Krstovski, C. Moran, K. Subramanian, D. Stallard, R. Prasad, P. Natarajan, "Recent Improvements in BBN's English/Iraqi Speech-

to-Speech Translation System", IEEE Spoken Language Technology Workshop, pp. 245–248, 2008.

[4] L. Nguyen and R. Schwartz, "Efficient 2-pass n-best Decoder", DARPA Speech Recognition Workshop, pp. 167–170, 1997.

[5] S. Tsakalidis, R. Hsiao, D. Karakos, T. Ng, S. Ranjan, G. Saikumar, L. Zhang, L. Nguyen, R. Schwartz and J. Makhoul, "The 2013 BBN Vietnamese Telephone Speech Keyword Spotting System", IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 7829-7833, 2014.

[6] M. Karafiat, F. Grezl, M. Hannemann, K. Vesely, and J. H. Cernocky, "BUT Babel system for spontaneous Cantonese", Interspeech, pp. 2589-2593, 2013.

[7] T. Ng, B. Zhang, S. Matsoukas, and L. Nguyen, "Region Dependent Transform on MLP Features for Speech Recognition", Interspeech, pp. 221–224, 2011.

[8] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz and J. Makhoul. "Fast and Robust Neural Network Joint Models for Statistical Machine Translation", Proc. of the Association for Computational Linguistics, pp. 1370-1380, 2014.

[9] F. J. Och and H. Ney. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, 29(1). pp. 19-51, 2003.

[10] D. Chiang. "Hierarchical Phrase-based Translation", Computational Linguistics, 33(2), pp. 201-228, 2007

[11] A. Rosti, B. Zhang, S. Matsoukas, and R. Schwartz. "BBN System Description for WMT10 System Combination Task", Proc. of WMT/MetricsMATR, pp. 321-326, 2010

[12] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. "Bleu: a Method for Automatic Evaluation of Machine Translation", Proc. of the Association for Computational Linguistics, pp. 311-318, 2002

[13] M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation", Proc. of the Association for Machine Translation in the Americas, pp. 223-231, 2006

[15] Y.-C. Tam, Y. Lei, J. Zheng, and W. Wang, "ASR Error Detection Using Recurrent Neural Network Language Model and Complementary ASR," Proc. of ICASSP, pp. 2331-2335, 2014

[16] R. Kumar, M. Roy, S. Ananthakrishnan, S. Hewavitharana and F. Choi. "Interactive Error Resolution Strategies for Speech-to-Speech Translation Systems", Proc. of the SIGDIAL, pp. 142-144, 2013.

[17] S. Ananthakrishnan, D. Mehay, S. Hewavitharana, R. Kumar, M. Roy and E. Kan. "Lightly-Supervised Word Sense Translation Error Detection and Resolution in an Interactive Conversational Spoken Language Translation System," Machine Translation, Kluwer Academic Publishers, vol. 29, no. 1, pp. 25-47, 2015.

[18] M. Turunen, and J. Hakulinen, “Agent-based Error Handling in Spoken Dialogue Systems”, Proc. of Eurospeech, pp. 2189–2192, 2001

[19] D. Bohus, and A. I. Rudnicky, "Sorry, i didn’t catch that!- an investigation of non-understanding errors and recovery strategies", Proc. of SIGDIAL, pp. 128-143, 2005

[20] E. Pincus, S. Stoyanchev and J. Hirschberg, “Exploring Features For Localized Detection of Speech Recognition Errors,” Proc. of SIGDIAL, pp. 132—136, 2013

[21] W. Chen, S. Ananthakrishnan, R. Kumar, R. Prasad, and P. Natarajan, “Automated ASR error detection in a conversational spoken language translation system”. Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing, pp. 7418-7422, 2013.

[22] Zhang, R., Kikui, G., Yamamoto, H., Watanabe, T., Soong, F., and Lo, W. K. “A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation”, *Proc. of 20th COLING, Stroudsburg, PA, USA, 2004*

[23] He, X. and Deng, L. “Optimization in Speech-Centric Information Processing: Criteria and techniques”, *Proc. of ICASSP, 2012, p. 5241-5244*

[24] Matsoukas, S., Bulyko, I., Xiang, B., Nguyen, K., Schwartz, R. and Makhoul, J. "Integrating Speech Recognition and Machine Translation," *Proc. of ICASSP, 2007, p. 1281- 1284*

[25] Suhm, B., Myers, B. and Waibel, A. "Interactive recovery from speech recognition errors in speech user interfaces," *Proc. of 4th ICSLP, 1996. p.865-868*

[26] N. F. Ayan, A. Mandal, M. Frandsen, J. Zheng, A. Kathol, F. Bechet, B. Favre, A. Marin, T. Kwiatkowski, M. Ostendorf, L. Zettlemoyer, P. Salletmayr, J. Hirschberg, and S. Stoyanchev, “Can you give me another word for hyperbaric?: Improving speech translation using targeted clarification questions,” Proc. of ICASSP, 2013