Source-Error Aware Phrase-Based Decoding for Robust Conversational Spoken Language Translation

Sankaranarayanan Ananthakrishnan, Wei Chen, Rohit Kumar, and Dennis Mehay

Speech, Language, and Multimedia Business Unit
Raytheon BBN Technologies
Cambridge, MA 02138, U.S.A.
{sanantha,wchen,rkumar,dmehay}@bbn.com

Abstract

Spoken language translation (SLT) systems typically follow a pipeline architecture, in which the best automatic speech recognition (ASR) hypothesis of an input utterance is fed into a statistical machine translation (SMT) system. Conversational speech often generates unrecoverable ASR errors owing to its rich vocabulary (e.g. out-of-vocabulary (OOV) named entities). In this paper, we study the possibility of alleviating the impact of unrecoverable ASR errors on translation performance by minimizing the contextual effects of incorrect source words in target hypotheses. Our approach is driven by locally-derived penalties applied to bilingual phrase pairs as well as target language model (LM) likelihoods in the vicinity of source errors. With oracle word error labels on an OOV word-rich English-to-Iraqi Arabic translation task, we show statistically significant relative improvements of 3.2% BLEU and 2.0% METEOR over an error-agnostic baseline SMT system. We then investigate the impact of imperfect source error labels on error-aware translation performance. Simulation experiments reveal that modest translation improvements are to be gained with this approach even when the source error labels are noisy.

1. Introduction

Conversational speech translation enables monolingual speakers of different languages to communicate with one another. The pipeline consists of ASR transcription of the input source language utterance, followed by text-to-text translation by SMT, and optional text-to-speech synthesis (TTS) in the target language. ASR performance is often a crucial bottleneck in the performance of speech translation systems, because it has a significant downstream impact on the SMT component. This is an important issue especially for spontaneous conversational speech, which exhibits a rich vocabulary even in domain-constrained applications, often resulting in a high OOV word rate. In the force protection and medical assistance domains, targeted under the DARPA TransTac and BOLT programs, a significant fraction of OOV entities refer to names of people, places, organizations, and objects. These OOV entities cause acoustically similar in-vocabulary words that best fit the linguistic context to be substituted in the 1-best ASR transcription, as illustrated in Figure 1. Furthermore, ASR errors caused by OOV entities are *unrecoverable*, i.e. there is no path in the ASR lattice that corresponds to the correct transcription.

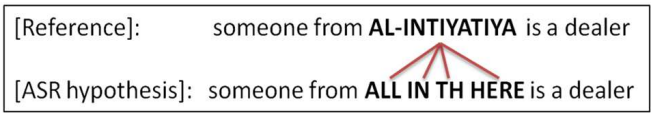


Figure 1: Unrecoverable ASR misrecognition caused by an OOV named-entity.

Translation errors caused directly by unrecoverable ASR errors, e.g. due to translation of source words substituted or inserted in place of an OOV entity, are unavoidable. However, these unrecoverable source language errors also affect translations of surrounding regions of error-free source words due to contextual effects. The goal of error-aware translation is to minimize the contextual impact of source errors and obtain the best possible translation for the correctly recognized portions of the utterance. We study this possibility by modifying a phrase-based SMT decoder to include penalties for bilingual phrase pairs spanning erroneous and error-free regions of input, and target language model (LM) likelihoods in the vicinity of source

errors. The proposed features are naturally integrated within a standard log-linear phrase-based translation model, resulting in a straightforward development and tuning process.

The remainder of this paper is organized as follows. Section 2 presents an overview of related work in this area. Section 3 describes the baseline speech translation pipeline, including details on the ASR and SMT systems. A detailed description of the proposed error-aware SMT decoding approach is given in Section 4. Experimental results are presented in Section 5. Finally, Section 6 concludes this paper with a brief discussion of our contribution and presents directions for future research in this area.

2. Relation to prior work

Integration of ASR and MT has gained popularity in the SLT community as a way of improving translation performance with potentially noisy input. This ranges from simple ASR post-processing to obtain segment boundaries or to insert punctuation [1, 2] to more sophisticated techniques such as joint decoding [3] and/or augmenting the SMT search space with ASR n-best lists, lattices, or word graphs (confusion networks) [4, 5]. The latter approach relies on the fact that the n-best list or lattice might contain a better hypothesis that could generate a more accurate translation. However, it is of limited utility in improving translation performance for utterances that generate unrecoverable ASR errors. Furthermore, the joint search space can be very large, making it difficult to implement some of these approaches for low memory, small form-factor devices that are preferred for SLT applications.

Our proposed approach is inspired by the idea of *attention-shift decoding* for ASR [6], where an input utterance is comprised of reliable *islands* and unreliable *gaps*. In this framework, initial hypotheses are constructed for the islands, and used to fill in the intermediate gaps in conjunction with additional information sources. In the case of SLT, islands refer to correctly recognized segments of the input utterance, while gaps consist of unrecoverable ASR errors. Our goal is to maximize translation performance on the correct islands, while minimizing interference from the incorrect gaps. In the SLT task domain, gaps will always generate translation errors and can only be filled in through additional external input (e.g. clarification dialog with the user). We refer the reader to our previous work [7] for more details on some of these interactive methods. In this paper, we focus solely on improving translation performance on the islands.

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

3. Baseline systems

The baseline ASR and SMT systems for our SLT application are built on data from the DARPA TransTac English-Iraqi Arabic parallel two-way spoken dialogue collection. These data span a variety of domains including force protection (e.g. checkpoint, reconnaissance, patrol), medical diagnosis and aid, maintenance and infrastructure, etc., and are conversational in genre. We focused on the English-to-Iraqi Arabic direction because this was a primary requirement of the ongoing DARPA BOLT program, under which a significant part of this research was conducted.

The baseline English ASR was 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 [8]. The acoustic model was trained on approximately 200 hours of transcribed English speech from the TransTac corpus. The LM was trained on 5.8M English sentences (60M words), drawn from both in-domain and out-of-domain sources. LM and decoding parameters were tuned on a held-out development set of 3,534 utterances (45k words). With a dictionary of 38k words, we obtained 12.8% WER on a separate held-out test set of 3,138 utterances.

Our English-to-Iraqi Arabic SMT system was trained on a parallel corpus derived from the TransTac collection (773k sentence pairs, 7.3M words). Phrase pairs were extracted from bidirectional IBM Model 4 word alignment [9, 10] based on the heuristic approach of [11]. The target LM was trained on Iraqi Arabic transcriptions from the parallel corpus. Our phrase-based decoder (similar to Moses [12]) performs beam search stack decoding based on a standard log-linear model, whose parameters were tuned with MERT [13] on a held-out development set (3,534 sentence pairs, 45k words). The BLEU and METEOR scores of this system on a noise-free held-out test set (3,138 sentence pairs, 38k words) were 16.1 and 42.5, respectively.

4. Error-aware SMT decoding

Phrase-based SMT decoders rely on context in order to construct a reasonably fluent translation of an input source sentence. Local source context is captured by multi-word phrase pairs, while local target context is modeled both by phrase pairs as well as a n-gram target LM. By definition, error regions in source input (gaps) produce incorrect translations. This affects translation of surrounding regions of error-free input (islands) due to two primary contextual effects:

1. Selection of phrase pairs whose source phrases span islands and gaps, leading to mixing of correct and incorrect words in the source context.

2. Erroneous target LM history causing propagation of bad hypotheses at the boundaries between translations of source gaps and islands.

Our proposed approach to error-aware phrase-based SMT decoding involves minimizing the contextual impact gaps can have on the translation of islands. We encourage this separation between translation of islands and gaps in two different ways: (a) by discouraging the decoder from choosing phrase translation pairs whose source phrases span island-gap boundaries; and (b) by preventing the propagation of bad target hypotheses generated by source gaps through the application of dynamic target language model penalties.

Throughout this paper, we assume that each ASR-hypothesized source word s_i is tagged with a corresponding probability of error e_i, ranging from 0.0 (correct) to 1.0 (error). These error probabilities might be based on oracle error labels (e.g. Levenshtein alignment of ASR transcription with the reference), or automatically estimated through some machine learning inference process. In interactive spoken language translation systems, source error information may also be gleaned directly from the user through clarification techniques such as ASR confirmation [7]. In the latter approach, the user hears a synthesized version of the ASR 1-best hypothesis, and can inform the system of incorrect regions (gaps) in the hypothesis.

We introduce two new features that leverage source error probabilities to minimize gap interference in translation of islands. These features are evaluated at run-time and integrate directly within the log-linear translation model framework. Tunable parameter weights associated with these features can be optimized with MERT on an appropriate development set. The proposed approach is highly efficient because it preserves the original search space and adds virtually no complexity to the SMT decoder.

4.1. Phrase pair error span penalty

We introduce a penalty term that applies to phrase pairs whose source phrases span the boundary between an island and a gap, thereby discouraging selection of erroneous source contexts for translation of correctly recognized words. This also encourages separation of incorrect target words generated by gaps from correct hypotheses due to islands, permitting replacement with other information that can render the translation comprehensible. For instance, the interactive SLT system described in [7] automatically identifies source gaps generated by OOV named entities, and replaces incorrect target words due to them with an audio segment corresponding to the spoken name.

The error span penalty is evaluated at run-time for each candidate phrase pair in the search graph based on the source words it spans. It is computed as the maximal difference between error probabilities of successive constituent words in the source phrase, and applies equally to all translation options generated by that source phrase.

$$F_{X \rightarrow Y}(s_i, s_j) = - \max_{i \leq k < j} \mid e_{k+1} - e_k \mid \qquad (1)$$

Equation 1 illustrates the evaluation of this feature for a sample phrase pair $X \rightarrow Y$ which spans contiguous source words $(s_i, \ldots, s_j)$ with error probabilities $(e_i, \ldots, e_j)$. The rationale behind this feature is that source phrases spanning island-gap boundaries are likely to exhibit large internal differences in source error probability. The error span penalty discourages the decoder from choosing translations whose source phrases potentially span island-gap boundaries. However, it does not penalize phrase pairs that exclusively span either correct source words (islands) or incorrect source words (gaps).

4.2. Target language model penalty

Bad phrase translations generated by source gaps can negatively influence the target context through the n-gram target LM. To prevent the propagation of errors in this manner, we introduce a dynamic target LM penalty that is applied to each translation hypothesis in the

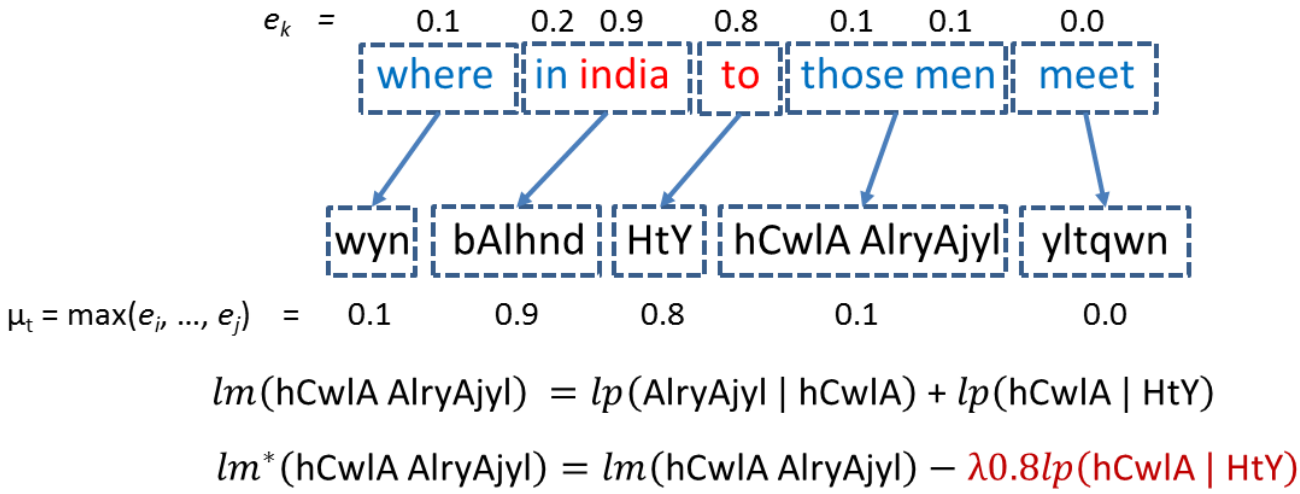


Figure 2: LM penalty highlighted for erroneous bigram context. Incorrect source words are highlighted in red.

beam search stack decoding process. In our decoder, each hypothesis, regardless of which stack it is placed in, records the most recent phrase pair (source/target span) that was used to arrive at that hypothesis. The total LM log-likelihood of the current hypothesis is evaluated as the sum of LM log-likelihoods of each constituent target word given its n-gram context. Depending on the n-gram order, the context may extend to target words from the previous hypothesis on a lower-order stack. If the most recent phrase pair used in obtaining the previous hypothesis corresponds to a source gap, we dynamically adjust LM log-likelihoods for target words in the current hypothesis whose context includes target words from the previous hypothesis.

In the example of Figure 2, the total LM score $lm(\texttt{hCwlA AlryAjyl})$ of the target phrase `hCwlA AlryAjyl` is the sum of the LM log-likelihoods $lp(\texttt{AlryAjyl} \mid \texttt{hCwlA})$ and $lp(\texttt{hCwlA} \mid \texttt{HtY})$ of the constituent words given the local context (without loss of generality, we illustrate using a bigram LM context). However, the bigram likelihood of the first word $lp(\texttt{hCwlA} \mid \texttt{HtY})$ is invalid due to the erroneous context `HtY`, which in turn was generated by the incorrectly recognized English source word (gap) `to`. Therefore, we apply a penalty factor to this term, weighted by the projection of the corresponding source error probability to the target context ($\mu_3 = 0.8$), in addition to the globally tunable feature weight λ. Thus, the penalty term attenuates the effect of the incorrect target hypothesis `HtY` to obtain the discounted LM score $lm^*(\texttt{hCwlA AlryAjyl})$, thereby alleviating the impact of erroneous LM context at run-time.

There is a subtle difference between discounting the total target phrase LM score via subtraction as described above, versus modifying the LM score directly via a multiplicative penalty factor. Our discounting approach is more flexible because it allows us to tune a feature weight specifically for the penalty discount, without affecting the main LM feature weight. In other words, the total penalty can be separated from the total LM score. An alternate solution would have been to back-off to the unigram likelihood of `hCwlA` instead. However, back-off can only be applied with categorical error labels, precluding the use of soft weighting and tunable parameters.

5. Experimental results

To evaluate the proposed approach, we designed and created high-error development/test (HED/HET) sets consisting of spoken utterances rich in OOV entities. Table 1 summarizes these datasets, which exhibit very high OOV/ASR error rates compared to the baseline development/test sets. Consequently, the baseline translation scores of 1-best ASR hypotheses of the HET set were significantly lower (Table 2). For reference, the noise-free test set baseline BLEU and METEOR scores were 16.1 and 42.5, respectively.

We offer a proof-of-concept evaluation of error-aware SMT decoding using oracle source error labels for the HED/HET sets, i.e. with error probability of all correct and incorrect source words set to 0.0 and 1.0, respectively. The oracle error labels were obtained by

Dataset	#Utts	#Words	OOV%	WER
HED	627	6.4k	2.9%	31.8%
HET	507	5.3k	8.9%	46.8%

Table 1: High-error dev/test data statistics.

automatic alignment of ASR hypotheses to the reference transcriptions. Because of the relatively small size of the HED set compared to the baseline development set, we only optimized the weights of the two proposed features on the HED set, carrying over all other tunable parameters from the baseline system. This also allowed a fair comparison, summarized in Table 2, between the baseline and error-aware systems. In combination, the proposed features produced relative gains of 3.2% BLEU and 2.0% METEOR over the baseline system on error-labeled ASR transcriptions of the HET set. Because it is impossible to translate gaps correctly, these improvements are attributable solely to better translations of the islands.

To verify the statistical signficance of this improvement, we performed the non-parametric Wilcoxon signed-rank test based on pair-wise bootstrap resampling [14] of the baseline and error-aware SMT hypotheses. With 100 randomized samples, the p-value returned by this test was 5.14×10^{-17}, thus confirming statistical significance of the improvement at $\alpha = 0.01$.

System	BLEU	METEOR
Baseline	5.67	24.62
EAD (oracle)	**5.85**	**25.12**
EAD (estimated)	5.61	**24.86**

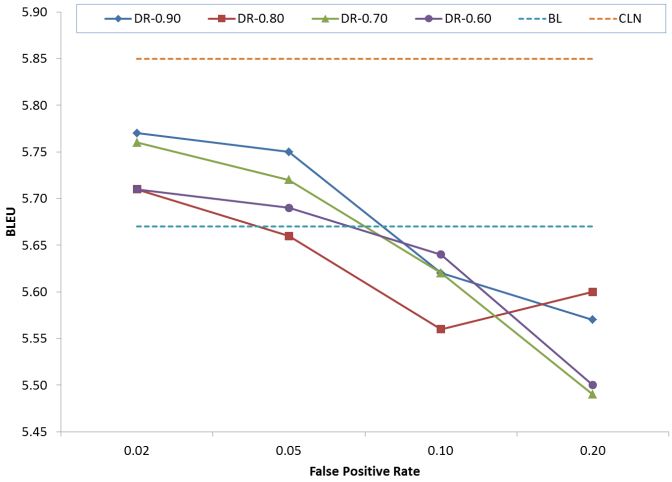
Table 2: HET set translation scores for error-aware decoding (EAD) with oracle/estimated error probabilities.

Achieving perfect ASR error detection is nearly impossible with current technology. We investigated the impact of noisy source error labels on translation performance in order to determine the noise level at which error-aware SMT decoding no longer achieves its goal. We simulated false alarms and missed detections by deliberately injecting noise into the oracle error labels, i.e. randomly changing 0.0 error probabilities to 1.0, and vice-versa in the desired proportion. Figure 3 illustrates the trajectories of BLEU/METEOR scores of error-aware decoding on the HET set across a range of false alarm rates (x-axis). Each curve corresponds to a specific detection rate; for instance, "DR-0.90" refers to 90% error detection rate. Each data point on every curve is the average of 10 independent noise simulations, giving a smooth trajectory of the performance trend. The simulation results are consistent with our intuition that there must be a gradual degradation in translation performance (BLEU/METEOR scores) as the noise level in the source word error labels (false alarm rate) increases. We note that error-aware decoding provides modest BLEU score improvements over the baseline SMT system as long as the false alarm rate is low (2-5%) and detection rate is high (70-80%). METEOR improvements persist at noisier operating points.

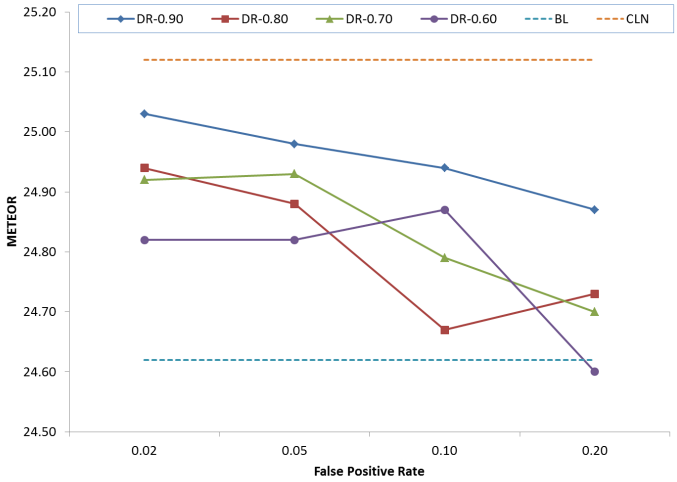
In our final experiment, we attempted to determine whether automatic detection of ASR errors could be used in conjunction with error-aware SMT decoding to improve translation performance in the absence of oracle ASR error labels. To this end, we coupled error-aware SMT decoding with a CRF-based automated ASR error predictor trained on a variety of features, including ASR and SMT confidence scores, subword ASR hypothesis mismatch, word boundary verification, named-entity detection, etc. The predictor infers a real-valued probability of error in [0.0, 1.0] for each source word in the HED/HET sets. Our recent work [15] provides more details on this system. ROC analysis showed that the automated predictor achieved 68% ASR error detection rate at 10% false alarm rate on the HET set. Error probabilities inferred by this system were used to evaluate the proposed penalties for SMT decoding. While the corresponding BLEU score does not improve (final row of Table 2), the METEOR score is slightly better than the baseline system. Given the current performance level of the automated error predictor, these results are in complete agreement with our simulation experiments.

6. Conclusion and future directions

ASR performance is a crucial bottleneck for downstream SMT quality in conversational speech translation systems. Unrecoverable ASR errors due to OOV words can also impact subsequent translation of surrounding, correctly recognized words due to contextual effects. Thus, errors in the source input can cause imperfect or incorrect translation of error-free neighboring words. Besides being less effective on utterances that generate unrecoverable ASR errors, traditional methods of integrating ASR and SMT (for instance, via lattice or



(a) BLEU Trajectories



(b) METEOR Trajectories

Figure 3: Trajectory of BLEU and METEOR scores for error-aware decoding at various false alarm and detection rates for error labels. Dashed horizontal lines represent the baseline (lower) and error-aware decoding with perfect error detection (upper). Figures show a gradual degradation in SMT performance as the noise level in the error labels increases.

n-best based search space augmentation) can be computationally expensive as well as memory intensive.

We presented an exploratory study in which we made targeted modifications to a phrase-based SMT decoder that reduce interference of incorrect gaps on translation of correct islands by introducing dynamic penalties applied to bilingual phrase pairs and the target LM. The new features were directly integrated within the log-linear model, resulting in straightforward development and tuning of the modified SMT system.

In the proof-of-concept experiment where we assumed perfect knowledge of source errors, the proposed modifications gave statistically significant relative improvements of 3.2% BLEU and 2.0% METEOR over the baseline system. Comprehensive simulation experiments revealed that modest translation improvements persist even in the presence of false alarms and missed detections of source errors, subject to certain thresholds. Coupling automated ASR error detection with error-aware SMT decoding yielded small gains in METEOR. We expect translation performance to improve as error prediction accuracy increases.

Based on these observations, one of our primary goals for the future is to improve automated ASR error detection capability for coupling with error-aware decoding. On the other hand, interactive, clarification-enabled SLT systems (e.g. two-way speech-to-speech translation systems) permit us to leverage user feedback to obtain source error labels. For example, based on cues from the automated ASR error detector, the system may request the speaker to confirm whether a sequence of ASR-hypothesized words is incorrect. In this way, user feedback can be used to construct oracle source error labels as input to the error-aware SMT decoder.

7. Acknowledgements

This paper is based upon work supported by the DARPA BOLT program. The views expressed here are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

8. References

[1] S. Matsoukas, I. Bulyko, B. Xiang, K. Nguyen, R. Schwartz, and J. Makhoul, "Integrating speech recognition and machine translation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, HI, April 2007, pp. 1281–1284.

[2] Y. Al-Onaizan and L. Mangu, "Arabic ASR and MT integration for GALE," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, HI, April 2007, pp. 1285–1288.

[3] E. Matusov, S. Kanthak, and H. Ney, "Integrating speech recognition and machine translation: Where do we stand?" in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006, pp. 1217–1220.

[4] R. Zhang and G. Kikui, "Integration of speech recognition and machine translation: Speech recognition word lattice translation," *Speech Communication*, vol. 48, no. 3-4, pp. 321–334, 2006.

[5] L. Mathias, "Statistical machine translation and automatic speech recognition under uncertainty," Ph.D. dissertation, Baltimore, MD, USA, 2008.

[6] R. Kumaran, J. Bilmes, and K. Kirchhoff, "Attention shift decoding for conversational speech recognition," in *INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 1493–1496.

[7] R. Prasad, R. Kumar, S. Ananthakrishnan, W. Chen, S. Hewavitharana, M. Roy, F. Choi, A. Challenner, E. Kan, A. Neelakantan, and P. Natarajan, "Active error detection and resolution for speech-to-speech translation," in *International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, December 2012, pp. 150–157.

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

[9] P. E. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, pp. 263–311.

[10] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, March 2003. [Online]. Available: http://dx.doi.org/10.1162/089120103321337421

[11] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 48–54.

[12] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557769.1557821

[13] F. J. Och, "Minimum error rate training in statistical machine translation," in *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 160–167.

[14] P. Koehn, "Statistical significance tests for machine translation evaluation," in *EMNLP*, Barcelona, Spain, July 2004, pp. 388–395.

[15] W. Chen, S. Ananthakrishnan, R. Kumar, R. Prasad, and P. Natarajan, "ASR error detection in a conversational spoken language translation system," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 7418–7422.