THE IMPACT OF ASR ON SPEECH–TO–SPEECH TRANSLATION PERFORMANCE

Ruhi Sarikaya, Bowen Zhou, Daniel Povey, Mohamed Afify and Yuqing Gao
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{sarikaya,bzhou,dpovey,maafify,yuqing}@us.ibm.com

ABSTRACT

This paper reports on experiments to quantify the impact of Automatic Speech Recognition (ASR) in general and discriminatively trained ASR in particular on the Machine Translation (MT) performance. The Minimum Phone Error (MPE) training method is employed for building the discriminative ASR acoustic models and a Weighted Finite State Transducer (WFST) based method is used for MT. The experiments are performed on a two–way English/Dialectal–Arabic speech–to-speech (S2S) translation task in the military/medical domain. We demonstrate the relationship between ASR and MT performance measured by BLEU and human judgment for both directions of the translation. Moreover, we question the use of BLEU metric for assessing the MT quality, present our observations and draw some conclusions.

Index Terms: Speech Recognition, ASR, Machine Translation, MT, Performance Metric.

1. INTRODUCTION

Despite recent effort to improve integration of ASR and MT components through a word graph interface [1] so as to reduce the impact of ASR errors on the S2S translation performance, current state-of-the-art S2S translation systems are built by cascading ASR and MT components. In the cascade integration paradigm the ASR and MT operate independently without tight coupling. Typically, the MT component is presented with the single-best recognition hypothesis. As such, poor ASR performance should have a great impact on the S2S performance. This impact should be particularly pronounced if the system is operating in a mismatched acoustic/environmental condition where the Word Error Rate (WER) is high.

It is generally assumed that improving the ASR performance improves the MT performance. However, so far this relationship has not been thoroughly studied, since researchers have been primarily focusing on improving automatic performance metric scores that are in some way correlated with actual performance improvement perceived by the users. However, taking the S2S systems out of the laboratory to deploy in real world requires not only the examination of the "true" impact of the ASR on the S2S performance but also questioning the automatic performance metrics used for assessing the MT quality. Even though automatic metrics such as BLEU [2] have previously been shown to correlate well with human judgment, there is a new study questioning this correlation [3]. Therefore, it is essential to quantify the impact of ASR improvements on the S2S translation performance using not only BLEU [2] but also human judgment. Here, the particular questions we want to answer are: 1) How does WER effect human judgment of translation quality as compared to an automatic metric like BLEU, 2) How much MT improvement one can get from discriminative training of ASR acoustic models, 3) What are the issues with BLEU in assessing MT quality.

The rest of the paper is organized as follows. Section 2 describes the discriminative ASR training method. A brief description of WFST based MT is provided in Section 3. The experimental setup introducing the data as well as the ASR architecture is presented in Section 4. Results and discussion are provided in Section 5. Finally, Section 6 summarizes the findings.

2. DISCRIMINATIVE ACOUSTIC MODEL TRAINING

Until the last few years discriminative training techniques were thought to be ineffective in reducing WER for large vocabulary ASR tasks using HMM systems. The key issues were a viable computational framework which allows incorrect word hypotheses to be efficiently processed and good generalization to test data. The computation issue is circumvented by using a lattice-based framework along with the Extended Baum–Welch (EBW) algorithm [4] for Maximum Mutual Information Estimation (MMIE) training of the parameters. Generalization is improved by using acoustic scaling to increase the effective amount of confusable data [5] and a weak unigram language model during training.

As an alternative to MMIE training the Minimum Word Error (MWE) objective function was previously proposed [6]. MWE maximizes the expected word accuracy and can easily be computed in a lattice framework. As a natural extension to MWE the Minimum Phone Error (MPE) criterion which uses the same approach at the phone level was also proposed [6]. MPE is the summed "Raw Phone Accuracy" (RPA) times the posterior sentence probability. For R training observation sequences $\{\mathcal{O}_1, ..., \mathcal{O}_r, ..., \mathcal{O}_R\}$ with corresponding transcriptions s_r, the MPE objective function for HMM parameter set λ, including the effect of scaling the acoustic and LM probabilities can be written:

$$\begin{aligned} \mathcal{F}_{MPE}(\lambda) &= \sum_{r=1}^{R} \frac{\sum_s p_\lambda(\mathcal{O}_r|s)^{\mathcal{K}} P(s)^{\mathcal{K}} \mathrm{RPA}(s,s_r)}{\sum_s p_\lambda(\mathcal{O}_r|s)^{\mathcal{K}} P(s)^{\mathcal{K}}} \\ &= \sum_{r=1}^{R} \sum_s P^{\mathcal{K}}(s_r|\mathcal{O}_r,\lambda)\mathrm{RPA}(s,s_r), \quad (1) \end{aligned}$$

This function measures the expected phone accuracy of a sentence drawn randomly from the possible transcriptions. The summation in the denominator is taken over all possible word sequences allowed in the task. Hence MPE training maximizes the posterior probability of the correct phone sequence. The denominator can be approximated by a word lattice of alternative sentence hypotheses. MPE was shown to outperform both MMIE and MWE on a number of large vocabulary ASR tasks [6].

3. WFST BASED MACHINE TRANSLATION

The statistical MT problem has been formulated as that of finding the most likely word sequence, $\hat{e}$, in some target language E, given the word sequence, f, in the source language F [7]:

$$\hat{e} = \arg\max_{e} P(f|e)P(e), \tag{2}$$

where $P(e)$ is the language model of E, $P(e|f)$ is the translation model and the argmax operation denotes the search problem. Hence, a statistical MT system consists of a training phase to construct the translation and language models and a search phase to decode the most likely word sequence in a target language. For this study, we use a memory efficient and fast phrase-based statistical machine translation system introduced in [8]. In this approach, we statistically construct a single optimized WFST, which is titled Statistical Integrated Phrase Lattice (SIPL). A beam Viterbi decoder employing a multilayer search algorithm [8] is developed to combine the translation model and language model FSTs with the input lattice efficiently. The translation problem can be framed as finding the best path in the full search lattice given an input sentence/automaton I. To address the problem of efficiently computing $I \ o \ M \ o \ L$, we have developed a multilayer search algorithm. Here, o is the FST composition operator, L ($P(e)$ in Eq. 2) is the target language model, and M ($P(f|e)$ in Eq. 2) is the SIPL that encodes the translation model, which is computed as follows:

$$M = Min(Min(Det(P)oT)oW), \tag{3}$$

where Det and Min denote the determinization and minimization operations respectively and P, T, and W refer to the transducers of source language segmentation, the phrase translation, and the target language phrase-to-word transducers, respectively. The P in Eq. 3 becomes determinizable due to an auxiliary symbol, EOP that marks the end of each distinct source phrase [8]. In spite of the fact that T and W in Eq. 3 are not deterministic and that minimization is formally defined on deterministic machines, in practice we often find that minimization can help reduce the number of states of non-deterministic machines. It should also be noted that due to the determinizability of P, M can be computed offline using a moderate amount of memory. The multilayer search algorithm has not only significant memory efficiency and being faster than general composition implementations found in FSM toolkits, but it can also incorporate information sources that cannot be easily represented using a available WFST toolkits. For example, the decoder allows us to apply the translation length and phrase penalties to score the partial translation candidates during search.

4. SPEECH RECOGNITION SYSTEM ARCHITECTURE

The Iraqi–Arabic acoustic training data consists of about 200 hours of speech collected in the context of a S2S translation project [9, 10], which covers the military and medical domains. The acoustic features and model training algorithms are common to both Iraqi–Arabic and English. The speech data is sampled at 16kHz and the feature vectors are computed every 10ms. The 24 dimensional MFCC features are then mean normalized, and 9 vectors are stacked leading to a 216–dimensional parameter space. The feature space is finally reduced to 40 dimensions using a combination of linear discriminant analysis (LDA), and maximum likelihood linear transformation (MLLT). There are 33 graphemes representing the speech and silence. Each phone is modeled with a 3-state left-to-right HMM. Building the decision tree for the Iraqi–Arabic data results in about 2K leaves and 75K Gaussians. On the other hand the English acoustic models are trained on about 400 hours of acoustic data that is largely collected for non–S2S applications. The English system uses an alphabet of 52 phones. This system has approximately 3.5K context–dependent states modeled using 42K Gaussian distributions.

A statistical trigram language model is built for both English and Iraqi Arabic side of the S2S system. The English language model uses a corpus of 6.4M words with 30K unique vocabulary items. The Iraqi–Arabic language model uses about 400K utterances with 98K vocabulary items for language modeling. All language models are built using modified Knesser-Ney smoothing technique [12].

5. RESULTS and DISCUSSION

5.1 ASR Results

The ASR test data consists of 1440 parallel utterances on the English (EN) and Iraqi–Arabic (IA) sides. This data is obtained from real dialogs spoken by native speakers of each language. Table 1 shows the WERs for MLE and MPE trained ASR acoustic models. MPE gives the largest reduction in WER for the English (EN) side (42% relative reduction). The improvement on the Iraqi–Arabic (IA) side is not as large (19% relative reduction). It should be noted that MPE iterations beyond two did not provide further improvement.

5.2 MT Results

The MT component uses a parallel corpus of about 337K utterances to build the translation models and also uses the same test data as that of the ASR. In order to simulate the ASR output at different WERs, speech is recognized using both the MLE acoustic model and

	EN	IA	EN → IA	IA → EN
Models	WER (%)		BLEU	
Perfect Text	0	0	0.2379	0.3987
MLE	12.7	33.4	0.2043	0.3150
MPE.1	8.2	28.8	0.2219	0.3765
MPE.2	7.3	27.2	0.2211	0.3805

Table 1: Word Error Rates (WER) for various English (EN) and Iraqi–Arabic (IA) ASR acoustic models and the evaluation of the translation performance using the BLEU Score.

	$EN \rightarrow IA$		$IA \rightarrow EN$	
	Trans. Reference	Perf. Text Trans.	Trans. Reference	Perf. Text Trans.
# Words	8385	7980	10425	9655
Perplexity	386	281	77	56
4gr Hit Rate	46%	49%	58%	68%

Table 2: Perplexity and 4gr Hit Rates for the Translation Reference and Perfect Text Translation.

different iterations of MPE trained models. The ASR output is then fed into the MT unit. The MT results are evaluated using the BLEU metric and Human Evaluators (HE). The BLEU metric is defined as follows:

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^{N} w_n \log p_n), \qquad (4)$$

where N (typically 4) is the maximum n-gram length, w_n and p_n are the corresponding weight and precision, respectively, and BP is the brevity penalty.

The BLEU scores for the translations of the ASR outputs are also shown in Table 1 along with the translations of the perfect text (i.e, bypassing ASR and using human transcription of the speech). The scores are higher for $IA \rightarrow EN$ compared to $EN \rightarrow IA$. The BLEU scores of the perfect text translations suggest that the translation is more accurate for the $IA \rightarrow EN$ direction. We believe this is due to two factors contributing to large -what we call- the *Intrinsic Language Perplexity* (ILP) for IA: i) low n-gram hit rate that the BLEU is based on and ii) highly inflected nature of the Arabic language. For example, despite using the parallel corpora the vocabulary size for IA (80K) is more than three times as large as that of EN (24K). In addition, in Table 2 the perplexity and the 4gr hit rates are presented for the "Translation Reference" and "Perfect Text Translation" for both directions of the translation. This table essentially compares the respective ILPs for EN and IA by building language models using the parallel corpus and testing on the translation of the same test data. The perplexity figures for IA are about five times as large as those for EN. Likewise, the 4gr hit rates for EN are 12% to 19% higher than those for IA. All of these results suggest that ILP for IA is higher than that for EN.

Discriminatively trained ASR output results in a significant 6.5 points improvement in BLEU scores for the $IA \rightarrow EN$, whereas the improvement is about 1.7 point for the reverse direction. It is worth noting that for both translation directions there is a mere 1.7 (or 1.8) point difference between perfect text and MPE.2 ASR output. Another interesting observation is the fact that improving the WER from 27% to 0% (perfect text) results in only a 1.8 point improvement in BLEU score for IA.

Very Good [4]	Perfect Translation
Good [3]	Fluent translation with all information conveyed, there may be extra words or some unimportant words are missing without affecting the meaning of the sentence.
OK [2]	All important information translated correctly but some details missing or translation is awkward
Bad [1]	Some important information is missing that can lead to wrong understanding
Very Bad [0]	Unacceptable translation, almost all of the important information is missing

Table 3: Translation Quality Grades.

5.3 BLEU vs. Human Judgment

Despite widespread use of automatic metrics, human judgement is still valuable in evaluating the true impact of major changes to conventional translation systems [11]. For the human evaluation of translation quality three raters are instructed to assign one of the five translation quality/accuracy grades listed in Table 3. The $EN \rightarrow IA$ BLEU scores (scaled by 100) and the human judgment scores (scaled by 25) as a function of WER are plotted in Fig. 1 for the entire test set (BLEU-1) with 1440 sentences and a subset of the test data (BLEU-2), which is scored by Human Evaluators (HE). The subset is obtained by taking the union of those utterances that either MLE or MPE or both ASR outputs have an error. There are 721 and 1064 such utterances (out of 1440) for the $EN \rightarrow IA$ and $IA \rightarrow EN$ directions, respectively.

BLEU–1 starts at 23.8 for perfect text and linearly decreases to 10.4 when the WER is at 41.4%. The first four points in the graph correspond to perfect text, MLE, MPE.1 and MPE.2 ASR outputs, respectively. The rest of the WER points are generated by deliberately degrading the ASR through increasing the acoustic scale beyond the optimal value. For each percentage point degradation in WER BLEU–1 decreases by 0.32 point. BLEU–2 also linearly decreases with increasing WER at a rate of 0.28, whereas the human evaluation scores decrease by about 1.0 point for each percentage increase in the WER after fitting a linear line to all three human evaluation scores. Based on these results it appears that the human judgment is more sensitive to ASR errors than BLEU.

The inter–rater agreement between pairs of HE was computed using Cohen's Kappa statistic. Kappa scores were 0.19 (HE1,HE2), 0.48 (HE2,HE3), and 0.32 (HE1,HE3), for an average score of 0.33. However, the raters may agree on the ordering of the translations with respect to their quality, but not on the overall quality of each translation, which is not reflected in the Kappa scores. Therefore, a more appropriate statistic for this task is Pearson's r which were, 0.75, 0.79 and 0.75, respectively. All of these scores are significant ($p < 0.05$) showing strong correlation between the raters. The first human evaluator (HE1) is the most forgiving for translation errors where he assigned scores between 74 to 85 that are "Good" or better according to Table 3.

The $IA \rightarrow EN$ translation scores are provided in Fig. 2 where two HE are employed. Both Kappa score (0.48) and Pearson's r (0.79) exhibit strong agreement between the raters. There are several interesting results one can extract from the figure. First, unlike $EN \rightarrow IA$ direction the relationship between the WER and trans-

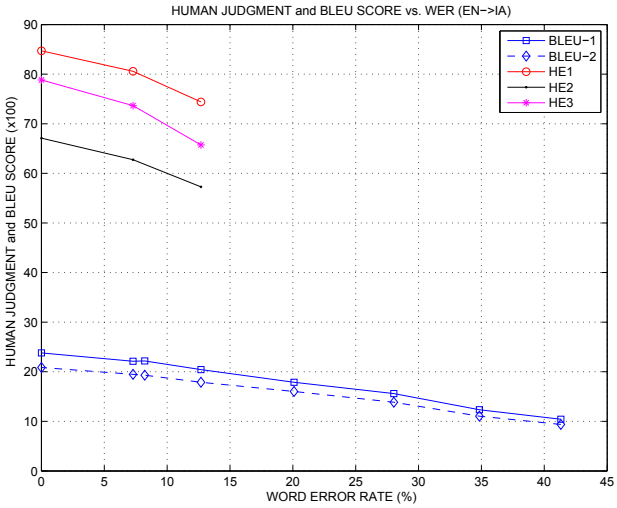

Figure 1: BLEU and Human Judgment vs. WER for EN → IA.

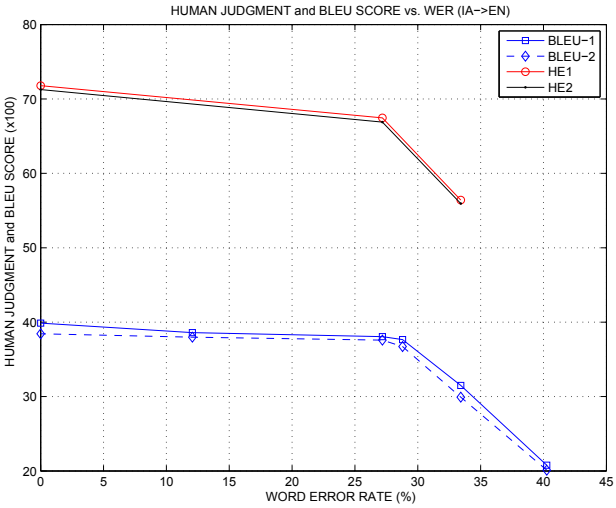

Figure 2: BLEU and Human Judgment vs. WER for IA → EN.

lation scores is a piecewise linear function. There is a critical WER level (∼27% for this case) that has to be attained to achieve large improvements both in BLEU and HE scores. However, after the critical level the improvement in ASR results in marginal improvements in BLEU metric. Additionally, human judgment is again more sensitive to ASR errors than the BLUE score; where BLUE scores decrease at a rate of 0.07 (BLEU-1) and 0.03 (BLEU-2) for each percentage increase in WER, whereas human judgment scores degrade at a rate of 0.17 when the WER increases from 0 to 27% WER. Further increases in WER (from 27% to 40%) results in a faster rate of decrease for BLEU-1 (1.32) and BLEU-2 (1.34). It is worth noting how small the reductions in BLEU scores when the WER degrades from 0 to 27%. The human judgment scores degrade with a rate of about 1.0 between 27–33% WER.

Learning word and phrase relationships has been proved to be difficult for statistical MT when the extent of morphological expression differs significantly across the source and target languages [13]. Hence, in the future it is worth investigating whether the findings of this work generalize to translations between English and other inflected languages.

6. CONCLUSIONS

We quantified the impact of ASR on the S2S performance on a two–way English/Dialectal–Arabic speech–to-speech (S2S) translation task in the military/medical domain. The results demonstrate that the relationship between WER and BLEU–based MT performance metric is not necessarily linear. In fact it can also be piecewise linear function depending on the direction of the translation and the language pair. Human evaluations of the MT results show that even though human judgment is correlated with the BLEU metric in differentiating alternative translations, humans are more sensitive to variations in the WER. Discriminative training of ASR has a significant impact on the translation performance. More importantly, even though the (average) human judgments are better for $EN \rightarrow IA$ direction the BLEU scores are about half as large as that of the $IA \rightarrow EN$ direction, which indicates that the absolute magnitude of the BLEU score is not a good indicator of the actual translation quality.

References

[1] E. Matusov, S. Kanthak and H. Ney, "On the Integration of Speech Recognition and Statistical Machine Translation", *Interspeech-05*, Lizbon, Portugal, 2005.

[2] K. Papineni, S. Roukos, T. Ward and W. Zhu, "Bleu: A Method for Automatic Evaluation of Machine Translation", *Proc. ACL*, 2002, Philadelphia, PA.

[3] C. Callison-Burch, M. Osborne and P. Koehn, "Re-evaluating the Role of BLEU in Machine Translation Research", *Proc. EACL*, Trento Italy, 2006.

[4] P.S. Gopalakrishnan, *et.al.*, "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," *IEEE Trans. on Information Theory*, v. 37, pp 107-113, 1991.

[5] P.C. Woodland and D. Povey, "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition," *Computer Speech Language*, v. 16, no 1, pp. 25-48, Jan 2002.

[6] D. Povey and P.C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," *ICASSP–02*, pp. 105–108, Orlando. FL 2002.

[7] P. F. Brown, *et al.*, "The Mathematics of Statistical Machine Translation: Parameter Estimation", *Comp. Linguistics*, 19(2):263–311, 1993.

[8] B. Zhou, S.F. Chen and Y. Gao, "Folsom: A Fast and Memory-Efficient Phrase-based Approach to Statistical Machine Translation," *IEEE SLT Workshop*, Aruba, 2006.

[9] Y. Gao, *et al.*, "IBM MASTOR: Multilingual Automatic Speech-to-Speech Translator," *ICASSP–2006*, Toulouse France, 2006.

[10] D. Stallard, *et al.*, "Design and Evaluation of the 2006 BBN English/Iraqi two-way speech translation system", *IEEE SLT–2006*, Aruba, 2006.

[11] D. Marcu, *et al.*, "SPMT: Statistical Machine Translation with Syntactified Target Language Phrases", *EMNLP–2006*, Sydney, 2006.

[12] S. Chen, J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling", *ACL-96*, Santa Cruz, CA, 1996.

[13] A. Zollmann, A. Venugopal and S. Vogel, "Bridging the Inflection Morphology Gap for Arabic Statistical Machine Translation", *HLT-2006*, New York, NY 2006.