

# The USFD SLT system for IWSLT 2014

*Raymond W. M. Ng, Mortaza Doulaty, Rama Doddipatla, Wilker Aziz, Kashif Shah,  
Oscar Saz, Madina Hasan, Ghada AlHarbi, Lucia Specia, Thomas Hain*

Department of Computer Science, University of Sheffield, United Kingdom

{wm.ng,mortaza.doulaty,r.doddipatla,w.aziz,kashif.shah,  
o.saztorralba,m.hasan,GAlHarbi,l.specia,t.hain}@sheffield.ac.uk

## Abstract

The University of Sheffield (USFD) participated in the International Workshop for Spoken Language Translation (IWSLT) in 2014. In this paper, we will introduce the USFD SLT system for IWSLT. Automatic speech recognition (ASR) is achieved by two multi-pass deep neural network systems with adaptation and rescoring techniques. Machine translation (MT) is achieved by a phrase-based system. The USFD primary system incorporates state-of-the-art ASR and MT techniques and gives a BLEU score of 23.45 and 14.75 on the English-to-French and English-to-German speech-to-text translation task with the IWSLT 2014 data. The USFD contrastive systems explore the integration of ASR and MT by using a quality estimation system to rescore the ASR outputs, optimising towards better translation. This gives a further 0.54 and 0.26 BLEU improvement respectively on the IWSLT 2012 and 2014 evaluation data.

## 1. Introduction

In this paper, the University of Sheffield (USFD) system for the International Workshop on Spoken Language Translation (IWSLT) 2014 is introduced. USFD participated in English-to-French and English-to-German SLT tasks. The ASR and MT systems made use of state-of-the-art technologies. On the ASR side, two deep neural network systems built on partially different data and different tandem configurations were used. On the MT side, phrase-based translation models were built. ASR and MT system integration attempts were made by using a translation quality estimation system. It considered the system scores from both ASR and MT, as well as features extracted from the ASR outputs in source language. The ASR hypotheses were then rescored based on the predicted translation quality. This gives performance improvements in terms of BLEU score increase.

In the following, the data used for system training is introduced in §2. §3 and §4 give the details of the ASR and MT systems. The decoding algorithm and system results are given in §5. Besides the primary submission, USFD also submitted contrastive systems which implement system integration. These systems used a quality estimation module and performed ASR  $N$ -best list rescoring based on predicted translation quality. This would be described in §6.

## 2. Data processing and selection

The ASR and MT systems were primarily trained on TED lecture data [1]. For ASR, TED and the additional data form two data subsets, on which two systems were trained. For MT, out-of-domain data after data selection were incorporated in the training of translation models and target language models.

### 2.1. ASR acoustic modelling

Two data sets were used for ASR system training. For the ease of discussion they are hereinafter referred to as ASR<sub>1</sub> and ASR<sub>2</sub>. The composition of the two data sets is shown in Table 1.

Table 1: Data for acoustic model training

ASR <sub>1</sub>		ASR <sub>2</sub>	
Data	Hours	Data	Hours
TED	132	TED	112
LLC	106	AMI+AMIDA+ICSI	165
ECRN	60	ECRN	60

TED serves as a common data set in both ASR<sub>1</sub> and ASR<sub>2</sub>. Their segmentations in ASR<sub>1</sub> and ASR<sub>2</sub> differ slightly and this is explained later. The two data sets are augmented by e-corner lecture data (ECRN) with a duration of 60 hours [2]. ASR<sub>1</sub> also contains 106 hours of LLC lecture data. In ASR<sub>2</sub>, 165 hours of meeting data from the AMI, AMIDA and ICSI corpora are added so the trained model will reflect also generic domains other than lectures [3, 4].

The TED portions in both ASR<sub>1</sub> and ASR<sub>2</sub> originate from 734 TED talks published before 31 Dec 2010. Each talk has a duration of around 15 minutes. Human annotations in the form of subtitles are also available, giving rough segmentation with segment duration from 3 to 5 seconds and time accuracy to the nearest second.

Exact segmentations and transcriptions of TED were derived in different ways in ASR<sub>1</sub> and ASR<sub>2</sub>. In ASR<sub>1</sub>, all segments from the same talk were merged and the speech was forced aligned, resegmented before another forced alignment run determined the final training set. This gave a total of 132 hours of speech for AM training. In ASR<sub>2</sub>, forced alignment

Table 2: Amount of text data used in different training tasks in En→Fr translation (#Full data set was used for building target LM)

Data	Number of words/million			
	Target LM#	Source LM	Punct TM	TM
TED	3.17	3.17	3.17	3.17
News Commentary	4.0	0.9	0.2	0.7
Common crawl	70.7	36.1	3.6	10.8
Gigaword	575.7	271.2	26.3	14.9
Europarl	50.3	10.8	4.3	1.9

was performed on the rough segmentation, after which contiguous segments were merged when there was tight silence at the segment boundaries. A further run of forced alignment determined the final training set. This gave a total of 112 hours of speech.

To evaluate the performance of different segmentations, PLP-based state-tied triphone models with cepstral mean and variance normalisation were trained on these data and decoding was performed on the IWSLT 2010 evaluation data set. The WERs for the ASR<sub>1</sub> and ASR<sub>2</sub> settings are 25.7% and 26.2% respectively. When the models are trained directly on the roughly segmented data (no adjustment of segmentations), the total duration of training data is 109 hours and the corresponding WER is 28.1%.

## 2.2. Language models and MT

Textual data for the training of language models and translation models were obtained from the affiliated websites of the IWSLT and WMT evaluations [5, 6]. TED was considered as the in-domain training data and the full data set was used. Four out-of-domain (OOD) data sets from News commentary v9, Common Crawl, Gigaword and Europarl v7 were also used, after a data selection process.

The OOD corpora were selected with the cross entropy difference criterion [7]. Given a sentence  $x_1^I = [x_1 \dots x_I]$  with  $I$  words, cross entropy values  $H(x_1^I, ID)$  and  $H(x_1^I, OOD)$  were computed using  $\mathcal{G}_{ID}$ , the ID language model (in this case, TED) and  $\mathcal{G}_{OOD}$ , the OOD language model (built on the corpus from which the sentence was taken). The cross entropy difference (CED) was given by,

$$CED(x_1^I) = H(x_1^I, \mathcal{G}_{ID}) - H(x_1^I, \mathcal{G}_{OOD}) \quad (1)$$

Sentences were ranked by the CED values and 25% of the sentences with the lowest CED values were selected from each corpus. Furthermore, CED values were calculated on sentence batches with increasing sizes. A line search was done to find the optimal batch giving the minimum CED value. All data selection was done on the English text. For data selection to translation model training, the corresponding sentences in the target languages were extracted after selection was done on English sentences.

Table 2 shows the amount of the full text data set, and the

selected text data in different systems in the English→French translation task. The full data set contains 703.9M words. They were used for training the target language model in MT, which was a 5-gram interpolated LM with punctuation and out-of-vocabulary word modelling, modified Kneser-Ney smoothing and was in standard ARPA format. The source language model for ASR was built on the full TED data set and 25% or 50% of the OOD data, making up to 322.2M words. A monolingual translation model was trained for punctuation insertion and case conversion. The training took the full TED data and 5-10% of the OOD data, resulting in a total of 37.6M words. The translation model was trained on the full TED data set and other optimally selected OOD data sets, where only around 5% of the sentences were selected. The total number of words is 31.7M.

## 3. Automatic speech recognition

There are two DNN systems with tandem configurations in ASR [8]. Bottleneck (BN) features were derived from deep neural network (DNN)s [4], and GMM-HMM systems were trained on these bottleneck features. The two tandem systems were trained on ASR<sub>1</sub> and ASR<sub>2</sub> data respectively (Table 1). Different portions of data were used in different stages of training. Let DNN<sub>1</sub> and DNN<sub>2</sub> denote the two DNN systems for ASR<sub>1</sub> and ASR<sub>2</sub>. DNN<sub>1</sub> was trained on TED data only. DNN<sub>2</sub> was trained on TED and AMI+AMIDA+ICSI data only. The remaining data listed in Table 1 were added to the training pool in the GMM-HMM training stage.

DNN<sub>1</sub> has 4 hidden layers, each having 1,745 hidden units. The BN layer is placed just before the output layer and has 26 units. The output layer has 4,320 units. DNN<sub>2</sub> has 5 hidden layers, with the first 3 layers having 1,745 units and the fourth hidden layer having 65 units. A BN layer is placed just before the output layer and has 39 units. The output layer has 5,691 units.

Both the DNNs were trained using log filter-bank outputs and concatenating 31 adjacent frames, which were decorrelated using DCT to form a 368-dimensional feature vector. The filter-bank outputs were mean and variance normalised at the speaker level. Global mean and variance normalisation was performed on each dimension before feeding the input for training the DNN. The GMM-HMM systems trained using the BN features were different. The model for ASR<sub>1</sub> was trained on the concatenated features with the 26-dimension BN features from DNN<sub>1</sub> and the 39-dimension PLP features. The model for ASR<sub>2</sub> was trained on the 39-dimension BN features from DNN<sub>2</sub>. Both the GMM-HMM models were trained as tied-state triphone systems with the final models having 16 mixture Gaussians per state.

All systems are vocal tract length normalised (VTLN). In the training stage, a PLP system was used to obtain the warp factors for each speaker. Then the filter-bank and PLP features were VTLN-warped, which were in turn used for DNN and GMM-HMM training in the tandem configuration. In the decoding stage, a non-VTLN DNN and GMM-HMM tandem

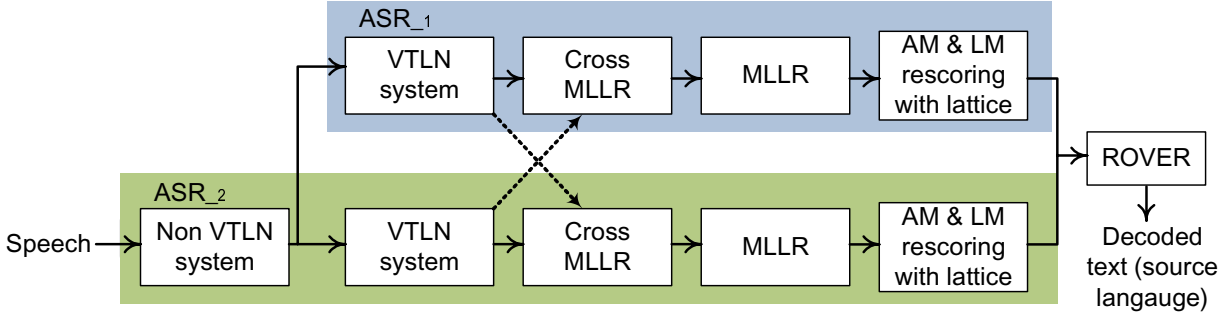


Figure 1: System diagram for multi-pass ASR decoding.

system trained on  $ASR_2$  data replaced the PLP system for the derivation of warp factors.

To improve the performance of the acoustic model, minimum phone error (MPE) training was performed using the lattices which were generated using a uni-gram language model [9].

Language models for ASR are all interpolated LMs built on the English text data described in Table 2 and tuned on IWSLT 2010 dev and eval data. 2-gram and 4-gram ARPA language models were trained for lattice generation and expansion. The 4-gram LM was pruned with a threshold  $10^{-10}$  and a weighted-finite-state transducer (WFST) was constructed for fast decoding in the pre-final passes in the ASR systems.

All ASR LMs were based on a word-list with a 60k word vocabulary extracted based on our standard English ASR inventory and the English part of the TED MT training data for IWSLT 2014 [3, 5]. Pilot ASR experiments on the IWSLT 2011 and 2012 eval data show the drop of perplexity with the addition of Common crawl and Gigaword data. For these two corpora, the rate of data selected for LM building was set to 50%, while the rate for other OOD corpora was kept 25%. This made the total number of words 322.2M as shown in Table 2.

Pronunciation probabilities were incorporated in final stage decoding [10]. These probabilities were extracted based on the Viterbi alignment of the phoneme level transcription of the  $ASR_1$  training data. When a word allowed multiple pronunciations, the frequency of each pronunciation was calculated and stored. These frequencies were then applied to the words in the decoding dictionary for words that appeared in both training and decoding stages. Words with multiple pronunciations appearing only in the decoding stage were given equal probability.

#### 4. Machine translation

A phrase-based model using MOSES [11] in a standard setting was employed. For phrase extraction all of the TED data (3.17 million words) was used. Following previous findings [12], data selection via a cross-entropy difference criterion (detailed in §2.2) was used to select the optimal batch of

the OOD data, which amounts to about 5% of the total data or 30.58M words. The phrase length was limited to 5 and word-alignment was obtained with FASTALIGN [13]. Lexicalised reordering models were trained using the same data. For language modelling, we used the complete sets of OOD data (i.e. no data selection). 5-gram LMs were trained using LMPLZ [14]. 100-best MIRA tuning was employed [15]. For the English-to-French system, tuning was done on the IWSLT 2010 development and evaluation data with a total of 2,551 sentences. For the English-to-German system, tuning was done on the IWSLT 2010 development data with 887 sentences.

In SLT, the input to the MT system was ASR output, which typically lacks casing and punctuation. Following previous work [16, 17], a monolingual translation system was trained to recover casing and punctuation from the ASR output, thus producing source sentences which are more adequate for translation. The training data for this monolingual MT system was obtained by pre-processing an actual corpus of the source language to form *pseudo ASR* outputs, which contained no case and punctuation information. Numbers, symbols and acronyms were also converted to their verbal forms with lookup tables. We then used this synthesised corpus of pseudo ASR as the source, and the original corpus as the target of our monolingual MT. The monolingual translation system was trained on 37.6M words (Table 2). It performed monotonic translation with phrases of as long as 7 words.

#### 5. Decoding

The evaluation systems for ASR and MT are multi-pass systems with resource optimisation and environment management capabilities [11, 18]. The ASR is a two-stream multi-pass system. It is illustrated in Figure 1. The two streams  $ASR_1$  and  $ASR_2$  differ by the acoustic model training data (detailed in Table 1) and also the tandem configurations (detailed in §3). Both streams follow the same routine along the multi-pass decoding system. In pass 1, a unified decoding result was generated using a non-VTLN DNN and GMM-HMM tandem system with cepstral mean and variance (CMVN) normalisation trained on  $ASR_2$  data. These

Table 3: Tree-search and WFST decoder

Decoder	Tst11		Tst12	
	WER	RT	WER	RT
Tree-search	23.7%	18.4	27.0%	19.8
WFST	23.7%	3.0	27.0%	3.3

hypothesis transcripts were used for inferring the warp factors. The filterbank (for both ASR<sub>1</sub> and ASR<sub>2</sub>) and PLP (for ASR<sub>1</sub> only) features were then warped and CMVN normalised, and the system branched off into two streams with two VTLN decoders trained on ASR<sub>1</sub> and ASR<sub>2</sub> data respectively.

After pass 2 decoding, speaker-based MLLR cross adaptations were carried out. The transcripts from ASR<sub>1</sub> was used for the model transformation in ASR<sub>2</sub> system and vice versa. The number of regression classes was set to 16. When pass 3 decoding was done, MLLR self adaptations were performed. The number of regression classes was also set to 16.

All pre-final stage decoding made use of weighted finite state transducers (WFSTs) for fast implementation. In a pilot experiment, PLP systems with heteroscedastic linear discriminant analysis (HLDA) were trained on the ASR<sub>2</sub> data [19]. WFST decoding with a pruned 4-gram grammar network was compared with the standard tree search with an unpruned 3-gram LM. The WER and real-time factor (RT) on IWSLT 2011 evaluation and IWSLT 2012 evaluation data are shown in Table 3. WFST was shown to achieve the same performance as tree-search decoding, with much faster decoding speed.

In the final stage, acoustic and language model rescoring were performed. Base lattices were generated with 2-gram LM pruned with a threshold  $10^{-10}$ . Lattice expansion was done with 4-gram unpruned language models. Three settings were tried and the results were compared,

- (i) Language model rescoring with the 4-gram LM
- (ii) Considering pronunciation probability (Pron. prob.) on top of (i)
- (iii) Acoustic and language model rescoring with the setting of (ii)

ASR performance in terms of WER are shown in Table 4. The initial non-VTLN system gave WER of 16.9% and 17.7% on IWSLT 2011 and 2012 data respectively. Moving towards the VTLN systems, when ASR<sub>1</sub> and ASR<sub>2</sub> branched off, it is observed that the ASR<sub>1</sub> model gave 1.0% to 1.4% lower WER than the ASR<sub>2</sub> model. This is because the data in ASR<sub>1</sub> had a better match in terms of domain. Incremental performance gains can be observed in individual steps, particularly MPE, cross-adaptation and language model rescoring. The WER difference between ASR<sub>1</sub> and ASR<sub>2</sub> diminished to 0.4-0.5% after all optimisation steps. After system combination, the final WER is 21-25% relatively lower compared with the initial system.

MT Decoding was performed with cube pruning [20] both in tuning and testing. Decoding was done with the min-

Table 4: WER of the multi-pass ASR systems

ASR system	Tst11		Tst12	
	ASR <sub>1</sub>	ASR <sub>2</sub>	ASR <sub>1</sub>	ASR <sub>2</sub>
Non-VTLN	–	16.9%	–	17.7%
+VTLN	15.4%	16.4%	16.4%	16.8%
+MPE	14.7%	15.7%	16.0%	16.1%
+Cross-adapt	14.0%	14.9%	14.2%	14.8%
+Self-adapt	14.0%	15.0%	14.2%	14.7%
+LM rescoring	13.4%	14.5%	13.5%	14.2%
+Pron. prob.	13.3%	14.2%	13.4%	14.0%
+AM rescoring	13.3%	13.8%	13.4%	13.7%
ROVER	—13.3%—		—13.2%—	

Table 5: MT system performance on eval data

Language pair	BLEU(c)	
	Dev10	Tst12
(MT with true transcript)		
En→Fr		40.9
En→De	21.5	
(Monolingual translation)		
En(pseudo ASR)→En		88.0
En(ASR)→En		69.0
(SLT)		
En(ASR)→En→Fr		31.7
En(ASR)→En→De	16.8	

imum Bayes risk criterion and reordering over punctuations was forbidden. To restore the correct case of the output the truecasing heuristic was employed. The same set of standard techniques was applied on En→Fr and En→De translation.

The MT system was tested on IWSLT 2010 development data and 2012 evaluation data, and the results are shown in Table 5. Performance are shown in terms of cased and punctuated BLEU scores. When given the reference transcript, the MT system gave 40.9 and 21.5 BLEU score for MT tasks in En→Fr and En→De respectively. The monolingual translation system (§4) restored case and punctuation information. It was tested on pseudo ASR and real ASR output and yielded 88.0 and 69.0 BLEU score. Finally in the SLT setting, the decoded ASR result was fed to the monolingual translation system and the output were subsequently translated. The BLEU score is 31.7 and 16.8 for SLT tasks in En→Fr and En→De respectively.

In Table 6, the official IWSLT 2014 evaluation performance in terms of BLEU and TER (cased, punctuated and non-case, non-punctuated) for the USFD primary system is shown.

Table 6: Primary SLT system performance (Tst14)

Language pair	BLEU(c)	TER(c)	BLEU	TER
En→Fr	23.45	59.94	24.14	58.97
En→De	14.75	70.15	15.24	69.15

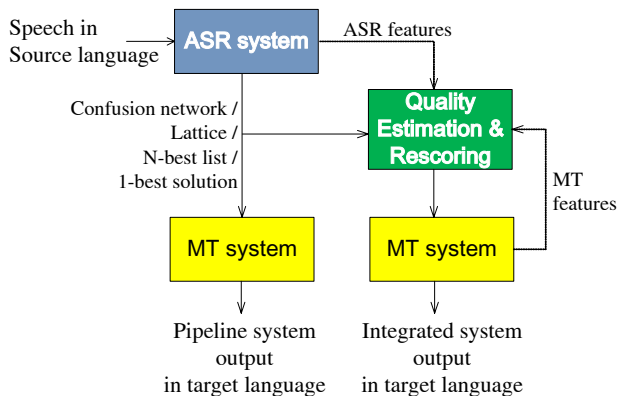


Figure 2: System integration with ASR and MT

## 6. System integration

The USFD primary system is a pipeline SLT system in which 1-best ASR result was directly fed to the MT system. System integration experiments were tried in the En→Fr SLT task and the results were submitted as contrastive systems. Figure 2 depicts the integrated system and its comparison with the pipeline system. In the integrated system, ASR system hypotheses are expanded in the form of lattices, confusion networks or  $N$ -best lists. A quality estimation (QE) module evaluated and rescored the ASR outputs before they were fed to the MT system.

In our implementation, 10-best outputs from the ASR system on the IWSLT 2011 evaluation data were used for QE training. The QE module derived 117 QuEst [21, 22] features from each sentence to describe its linguistic, statistical properties as well as the statistics from the ASR and MT models. Out of the 117 features, top 58 features were selected using the Gaussian Process (GP) with RBF kernel as described in [23]. Further, GP was used to learn the relationship between the selected features and the translation performance of the sentence (in this case, sentence-based METEOR score) [24]. During testing, the estimated translation performance was used to rescore the 10-best ASR output. Details of the integrated system were described in [25].

Table 7: Contrastive SLT system performance (En→Fr)

Setting	Tst12	Tst14
Contrastive 1 (baseline)	31.33	23.18
Contrastive 2 (+ 10-best list rescoring)	31.51	23.27
Contrastive 3 (+ ASR confidence-informed rescoring)	31.87	23.44

The ROVER combination of  $ASR_1$  and  $ASR_2$  systems only provided 1-best output. In the integration experiment, the 10-best output from  $ASR_1$  was used instead.

Performance of the contrastive systems in terms of cased and punctuated BLEU score is shown in Table 7. Contrastive

1 result is from the baseline system with pipeline setting. Contrastive 2 and 3 show the results of two different system integration settings. The baseline system gave BLEU scores 31.33 and 23.18 on IWSLT 2012 and IWSLT 2014 data. The baseline numbers are inferior to the primary system number (IWSLT 2012: 31.7; IWSLT 2014: 23.45) as shown in Table 5 and 6. This is because the baseline here did not benefit from ASR system combination.

Rescoring gives 0.18 and 0.09 BLEU improvements to IWSLT 2012 and IWSLT 2014 data respectively. By inspecting the results, it was found that rescoring generally had higher effectiveness for the sentences with low ASR confidence. Therefore, a confidence threshold was set, and rescoring was only performed when the ASR confidence dropped below this threshold. For IWSLT 2012 data, optimality was reached when 55% of the sentences were selected by this confidence criteria to rescore, resulting a further 0.36 BLEU score gain. This threshold was applied on IWSLT 2014 data, a 0.17 BLEU score gain was observed.

## 7. Summary

In this paper, the USFD SLT system for IWSLT 2014 was described. Automatic speech recognition (ASR) is achieved by two multi-pass deep neural network systems with slightly different tandem configurations and different training data. Machine translation (MT) is achieved by a monolingual phrase-based monotonic translation system which recovers case and inserts punctuation, followed by a bilingual phrase-based translation system. The USFD contrastive systems explore the integration of ASR and MT by using a quality estimation system to rescore the ASR outputs, optimising towards better translation. This gives noticeable BLEU improvement on the IWSLT 2012 and 2014 evaluation data.

## 8. References

- [1] TED, “Technology entertainment design,” <http://www.ted.com>, 2006.
- [2] M. Hasan, R. Doddipatla, and T. Hain, “Multi-pass sentence-end detection of lecture speech,” in *Proc. Interspeech*, 2014.
- [3] T. Hain, L. Burget, J. Dines, P. N. Garner, A. E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, “The AMIDA 2009 meeting transcription system,” in *Proc. Interspeech 2010*, 2010, pp. 358–361.
- [4] R. Doddipatla, M. Hasan, and T. Hain, “Speaker dependent bottleneck layer training for speaker adaptation in automatic speech recognition,” 2014.
- [5] M. Cettolo, C. Girardi, and M. Federico, “WIT3: Web inventory of transcribed and translated talks,” in *Proceedings of Conference of European Association for Machine Translation Trento (Italy)*, 2012, pp. 261–268.

- [6] “ACL 2014 ninth workshop on statistical machine translation,” <http://www.statmt.org/wmt14/translation-task.html>, 2014.
- [7] R. C. Moore and W. Lewis, “Intelligent selection of language model training data,” in *Proceedings of the ACL 2010 Conference Short Papers*, ser. ACLShort ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 220–224.
- [8] H. Hermansky, D. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems,” in *Proc. ICASSP*, 2000.
- [9] D. Povey and P. C. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” in *Proc. ICASSP*, 2002.
- [10] T. Hain, “Implicit modelling of pronunciation variation in automatic speech recognition,” *Speech Communication*, vol. 46, pp. 171–188, 2005.
- [11] 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.
- [12] A. Birch, N. Durrani, and P. Koehn, “Edinburgh SLT and MT system description for the IWSLT 2013 evaluation,” in *Proceedings of International Workshop on Spoken Language Translation*, 2013.
- [13] C. Dyer, V. Chahuneau, and N. A. Smith, “A simple, fast, and effective reparameterization of IBM model 2,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 644–648.
- [14] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, “Scalable modified Kneser-Ney language model estimation,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 690–696.
- [15] C. Cherry and G. Foster, “Batch tuning strategies for statistical machine translation,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. NAACL HLT ’12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 427–436.
- [16] S. Peitz, M. Freitag, A. Mauser, and H. Ney, “Modelling punctuation prediction as machine translation,” in *Proc. IWSLT*, 2011.
- [17] E. Cho, J. Niehues, and A. Waibel, “Segmentation and punctuation prediction in speech language translation using a monolingual translation system,” in *Proc. IWSLT*, 2012.
- [18] T. Hain, A. E. Hannani, S. N. Wrigley, and V. Wan, “Automatic speech recognition for scientific purposes - webASR,” in *Proc. Interspeech*, 2008, pp. 504–507.
- [19] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition,” *Speech Communication*, vol. 26, pp. 283–297, 1998.
- [20] L. Huang and D. Chiang, “Forest rescoring: Faster decoding with integrated language models,” in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 144–151.
- [21] L. Specia, K. Shah, J. G. C. d. Souza, and T. Cohn, “QuEst - A translation quality estimation framework,” in *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics: Demo Session*, Sofia, Bulgaria, 2013, p. 794.
- [22] K. Shah, E. Avramidis, E. Biçici, and L. Specia, “QuEst - design, implementation and extensions of a framework for machine translation quality estimation,” *Prague Bull. Math. Linguistics*, vol. 100, pp. 19–30, 2013.
- [23] K. Shah, T. Cohn, and L. Specia, “An Investigation on the Effectiveness of Features for Translation Quality Estimation,” in *Machine Translation Summit XIV*, Nice, France, 2013, pp. 167–174.
- [24] M. Denkowski and A. Lavie, “Meteor universal: Language specific translation evaluation for any target language,” in *Proceedings of WMT14*, 2014.
- [25] R. W. M. Ng, K. Shah, W. Aziz, L. Specia, and T. Hain, “Quality estimation for ASR K-best list rescoring in spoken language translation,” Submitted to *Proc. ICASSP*, 2015.