ADAPTATION OF LECTURE SPEECH RECOGNITION SYSTEM WITH MACHINE TRANSLATION OUTPUT

Raymond W. M. Ng, Thomas Hain, Trevor Cohn

Department of Computer Science, The University of Sheffield, United Kingdom

{w.ng,th}@dcs.shef.ac.uk,T.Cohn@sheffield.ac.uk

ABSTRACT

In spoken language translation, integration of the ASR and MT components is critical for good performance. In this paper, we consider the recognition setting where a text translation of each utterance is also available. We present experiments with different ASR system adaptation techniques to exploit MT system outputs. In particular, N-best MT outputs are represented as an utterance-specific language model, which are then used to rescore ASR lattices. We show that this method improves significantly over ASR alone, resulting in an absolute WER reduction of more than 6% for both in-domain and out-of-domain acoustic models.

Index Terms— TED talks, speech translation, language model adaptation

1. INTRODUCTION

Spoken language translation (SLT) is a challenging problem, combining the difficulties of automatic speech recognition (ASR) with those of machine translation (MT), and possibly speech synthesis. Most prior work on SLT has adopted a pipeline in which each step is done independently of the other steps [1]. Development in both ASR and MT technologies however has shown considerable task dependence, i.e., tuning to a specific domain is vital for competitive performance. Secondly, most SLT technologies exhibit better performance under integrated optimisation. For example, SLT integration was shown to be beneficial in works by Zhou [2] or Zhang [3]. Such integration is often applied as an extended interface between the two systems.

Better ASR output tends to lead to better translation results [4, 5] although the standard metrics used, word error rate (WER) and BLEU scores, were shown to be poorly correlated [6] under some circumstances. From the above it seems only natural that not only the constraint from the spoken language, but also that of the foreign target language can be used to obtain better performance. The transfer of knowledge can work both ways: Information about speech can be an input to the MT system, and prior knowledge about the target language can be input to improve recognition of the source. In this paper we focus on the latter case, an approach to exploit machine translation (MT) output for ASR system adaptation.

The current high interest in SLT technologies through internet and general globalisation trends has a positive impact on the availability of data in the last decade. More and more multi-lingual data for SLT system training can be found, but such data is often inaccurate or incomplete. For example the minutes of the European Parliament Plenary Sessions contain transcriptions that are only an approximate representation of the spoken words, but are translated into many languages. In contrast many television programmes or movies are broadcast in some countries (e.g., the Netherlands, Belgium or Hong Kong) in a foreign-language with subtitles, but no transcript of the source language [7]. A closely related conventional application scene is computer-assisted translation (CAT) [8][9]. Nevertheless, in CAT the ASR system deals with mostly clean speech from a designated human translator. In our case, environment and domain mismatch pose a greater challenge to the ASR. To benefit from the MT system, we apply backward translation to the text transcription in foreign language, retrieve semantically relevant text in the source language and perform adaptation of an ASR system.

After a brief review of key works in this area, the paper includes an outline of language model adaptation by interpolation as typically used for ASR system adaptation (Section 2). This is followed by a description of the experimental setup (Section 3) for MT and ASR as well as the TED talk corpus used in the experiments. Experimental results are described in section 4, followed by a detailed discussion and analysis of the outcome.

1.1. Prior work

SLT is of major interest to many groups, especially those participating in evaluation competitions (e.g., SLT task at IWSLT since 2010 [1, 10]). In part driven by large scale projects in SLT, several attempts at integrating ASR and MT can be found in the literature. [3, 5, 11] make use of ASR lattices as inputs to a phrased-base translation system, while [2] perform integration at the weighted finite state transducer level. [12] performs integration of knowledge from the source side through language model interpolation on the target output.

Adaptation of language and acoustic models using reverse translated text has been investigated in a series of publications [13, 9, 14, 15] developed on a smaller corpus (BTEC, a travel domain corpus). Paulik's work [13] is most similar to ours in which he developed a method for integrating ASR and MT using N-best lists. His experiments used very small $N \leq 150$, and showed diminishing performance with larger N. In contrast, we apply our experiment on a more challenging broader-domain data set, TED talks, and focus on the out-of-domain setting for both the translation and acoustic models. Our technique also uses N-best lists, however we use utterance rather than corpus-level LM to encode the MT output strings. Our results demonstrate consistent gains over a competitive baseline recogniser, and continuing improvement with larger N-best lists.

2. LANGUAGE MODEL INTERPOLATION

Language model adaptation through linear or log-linear interpolation can be effective for adaptation to a new domain [16]. For best performance it is common to obtain a language model trained on text from the target domain, the so-called *foreground* (FG) language model. However, especially for spoken text there is usually little precious data available for such purpose. In our setting, we con-

struct FG language model from the machine translation output of a foreign sentence, which has a very sparse vocabulary and is overall highly noisy, further exacerbating the problem. Hence a *background* (BG) language model is trained on significant amount of (quality) data in the generic domain. A simple linear combination has been shown to work well in a variety of different tasks (e.g., [16]). The interpolated language model has the form,

$$P_{INT}(w_1^N) = \lambda P_{BG}(w_1^N) + (1-\lambda) P_{FG}(w_1^N) \ , \qquad (1)$$

where $P(w_1^N)$ is the word n-gram probabilities and λ is the combination weight to tune.

Two questions to be addressed are training strategies for the FG language model and selection of the optimal interpolation weight λ. A single foreground language model trained on the whole corpus in the target domain can yield a very robust solution [17]. Alternatively class-based foreground models can better deal with small amounts of data [18]. Decreasing the amount of training text is only desirable when highly biased solutions are required. The extreme case here is a foreground language model for every utterance to be recognised. Such utterance-specific modelling can also be interpreted as a significantly weighted constraint on the search space without requiring strict ordering as is present in a lattice. The interpolation with the BG model still allows for substantial variation.

In tuning, the optimal interpolation weight λ is typically obtained on a held-out development set using maximum likelihood optimisation. In practice one can be faced with a situation where there simply is no reference data at all, and thus no means for optimising λ. Where reference in-domain data is used for tuning λ we have *supervised adaptation*. In contrast, *unsupervised adaptation* involves using errorful outputs. Such output can be either obtained from a first pass ASR system or from reverse translation into the source language.

In situations where significant amounts of errors are present in the hypothesised text richer representations, i.e., lattices or N-best lists, are more desirable. Statistical MT systems are typically capable of producing N-best lists with large values of N and the training of language models on these N-best lists is straight-forward. Such a model then includes on the one hand the semantic and syntactic constraints for a given utterance. Secondly, simply by repeated occurrence in the N-best list, certain n-grams are weighted higher, encoding the confidence of the translation system in the output. Training a language model in such a way, as well as selecting the interpolation weight, requires no prior knowledge and depends only on a first-pass system output from both ASR and MT systems.

3. EXPERIMENTAL FRAMEWORK

3.1. Data

All experiment reported in this paper are performed on TED talk data [19]. TED is an organisation that organises and records short lectures by important figures of the public, in English. They are then made available on the TED web-site. All lectures have English subtitles created by professional human transcribers. Translations to different languages are provided by a community of volunteers with less rigorous quality control. The data is an excellent multi-lingual resource for realistic speech and translation data. TED data have been used in one of the sub-tasks in IWSLT Evaluation Campaign from 2010-2012. The choice of our training and testing data were adapted from the IWSLT2012 Evaluation guidelines [20].

In-domain (ID) data comprises 174 hours of 762 TED talks published before 31 Dec 2010 chosen to exclude the talks in the IWSLT2012 development and test sets. We also make use of out-of-domain (OOD) data. For the ASR model, this comprises 170 hours of conversational speech from meetings [21] for acoustic training purposes only. Similarly, the MT model is trained exclusively using OOD data, namely the Europarl v7 English-French parallel corpus of EU parliamentary debates [22] which contains 1.96M parallel sentence pairs, and tuned on the news-test2008 dataset, 2051 sentence pairs drawn from the news domain.

For tuning and testing our integrated model, we use the IWSLT2012 development set.[1] The IWSLT2012 development set comprises the IWSLT2010 development (DEV10, 934 utterances) and test (TST10, 1664 utterances) sets, with a total speech durations of 1.52 hours and 2.48 hours respectively.[2] They are used as the development and test sets in these experiments.

3.2. Acoustic modelling

Two sets of acoustic models are used: one trained on out-of-domain (OOD) data and one on in-domain (ID) data. Both are decision tree phonetically state-tied triphone models with comparable numbers of Gaussian components. They are trained using perceptual linear prediction (PLP) coefficients [23], with cepstral mean and variance normalisation per speaker. The OOD models were developed for the AMIDA 2007 meeting recognition system [21], include Heteroscedastic Linear Discrimination Analysis (HLDA) [24] and are trained using the Minimum Phone Error Criterion [25]. The ID models are trained using a standard mixup procedure and the maximum likelihood criterion.

As mentioned earlier, resources not specifically targeted for recognition exhibit accuracy issues. For ID training this results in time alignment issues of the data. Although there are 174 hours of data as indicated by the closed caption, 30 hours are silence. Moreover, the closed caption comes with timing information are inaccurate. Only a certain percentage of data remains after a forced alignment process, depending on the pruning settings used. An experiment two different pruning settings yielded substantially different amounts of data, with 132 hours of non-silence speech data for wide pruning and 109 hours with tight pruning. Models derived from the wide-pruning set gave higher error rates and hence models trained on 109 hours only are used.

3.3. Machine translation

The machine translation (MT) system is an OOD system without any ASR-specific knowledge incorporated. We use the Moses phrase-based translation system [26], which we trained on the Europarl French-English data. The translation model was trained using Moses' default settings, which has been demonstrated to work well for many European language pairs [27]. In brief, the training process first performs unsupervised word alignment using GIZA++ [28], followed by heuristic extraction of a weighted collection of phrase-pairs to form the phrase-table [29], and various other features including lexicalised distortion [26]. The English side of the training set was used to build a 3-gram language model, using modified Kneser-Ney smoothing. Finally for system tuning, we performed maximum BLEU training [30] to learn feature weights on the OOD MT development set.

3.4. Integration

The integration technique uses linear interpolation between two language models – background (OOD) and foreground (ID) LMs. For

[1] The reference to the IWSLT2012 test set is not yet publicly available

[2] Note that utterances here correspond to full sentences rather than subtitle fragments. These were manually annotated for the IWSLT competition.

Table 1. *WER and BLEU between different data sets in IWSLT DEV10 and TST10*

Hypothesis	Reference: Gold		Reference: ASR output	Reference: MT output
	WER	BLEU	WER	WER
ASR (ID AM)	30.1%	52.1%	n/a	72.7%
MT (1-best)	56.9%	20.5%	69.1%	n/a

the background language model we use the rt09 3-gram language model. It has a vocabulary of 50k and comprises 10-component language models including those for conversational speech, broadcast news speech and considerable amounts of web data [31]. All experiments use maximum entropy pruning with a threshold of 10^{-8}.[3]

A number of alternatives for the foreground language models are considered, incorporating different knowledge about the target. The two main configurations are *offline* and *online integration*. For *offline integration*, foreground LM is a corpus-level LM constructed on TED training corpus. The interpolation weight λ is tuned on the development data using line-search to minimise the WER. The following development texts are considered: i) reference transcriptions (labelled *ref*); ii) outputs from the ID ASR system (*ASR hyp*); and iii) outputs from the MT system (*MT hyp*). Note that the ASR and MT settings use 1-best hypotheses. Each of these configurations have different levels of supervision in the target domain.

The second scenario is *online integration*, where the foreground LM is specific to each utterance. Here an ASR lattice and MT N-best are obtained in a first-pass decoding. Next the utterance-level LM is fit to the MT output, which is then interpolated with the background LM. Finally, the ASR lattice is rescored with the foreground LM to produce the overall best hypothesis. According to the terminology in [12] and [13], this is "dynamic loose coupling" of the ASR and MT systems.

The ASR base lattice is obtained with the background trigram language model. For the MT side, we consider various sizes of N-best lists, $N \in \{100, 1000, 2000\}$. These output lists are normalised, and then stripped of punctuation and out-of-vocabulary words. Each N-best list is used to train a 5-gram language model, which is smoothed using Kneser-Ney smoothing, where possible, or Witten-Bell otherwise. Note that there is no development text to govern the language model interpolation, so instead we experimentally evaluate a range of λ values. For the final lattice rescoring step, for computational reasons we consider 4-grams in the interpolated language model.

4. RESULTS

4.1. Comparison between ASR and MT output

Table 1 shows the results for the ASR system using ID acoustic model and the MT system as compared to the ground truth and each other. It is clear that the MT system outputs are drastically poorer than the ASR outputs in terms of both WER and BLEU (for BLEU bigger numbers are better). This raises the question of whether or not combining these outputs can improve overall performance. To gain insight into this question, we compare the MT and ASR outputs, and observe even higher WERs than compared with the gold transcript. This indicates that the two systems make very different errors, which augurs well for their combination. A related question is whether the poor MT performance might be improved by considering a wider range of outputs, that is, using N-best lists. Table 2 shows that the oracle WER does fall with larger N, motivating our use of large values of N in our subsequent experiments. Nonetheless, just acoustic rescoring of the N-best lists is not a solution.

[3]This excludes 97% of bigrams and trigrams, leaving 2.6M bigrams and 1.4M trigrams. Pruning has a small detrimental effect on the WER, amounting to about 1.5% absolute for both ID and OOD acoustic models.

Table 2. *WER of MT output with different size of N-best list*

$N=1$	2	5	10	20	50	100
56.9%	55.1%	52.7%	51.0%	49.4%	47.5%	46.2%

Table 3. *Offline integration with language model interpolation*

Condition	OOD-AM DEV10 WER	OOD-AM TST10 WER	ID-AM DEV10 WER	ID-AM TST10 WER	Perplexity on TST10
Unadapted	30.9%	31.3%	30.8%	29.6%	173.39
Supervised (ref)	29.6%	30.4%	29.7%	28.9%	157.57
Unsupervised (ASR hyp)	29.7%	30.7%	29.8%	28.9%	160.65
Unsupervised (MT hyp)	29.6%	30.5%	29.9%	28.8%	158.14

4.2. Offline integration

The results for offline integration is shown in Table 3. Experiments make use of out-of-domain (OOD) and in-domain (ID) acoustic model respectively. There is almost no benefit from using an ID AM over the OOD AM on DEV10 data, while on TST10 the in-domain model gives a WER reduction between 1.5% and 1.7%. This may in part be explained by better training technologies for building the OOD model (HLDA, MPE), however, normally the benefit of in-domain data usually far outweighs the benefit of such technologies.

The effect of LM integration can be observed by comparing different rows in Table 3. Supervised adaptation with interpolation weights tuned on reference text, gives the lowest perplexity on test data. For unsupervised adaptation, interpolated LMs tuned on MT hypothesis text give slightly lower perplexity than that tuned on ASR hypothesis text. The corresponding WER also reflects this trend. LM adaptation with OOD AM gives 0.6-0.9% WER reduction on test set TST10, with ID AM the WER reduction on TST10 is 0.7-0.8%.

Both supervised and unsupervised methods do not give large wins over the unadapted setting. This may be partly explained by the large amounts of differing texts used for training of the background LM. However it is surprising to see that integration with MT and with ASR yields very similar performance.

4.3. Online integration

For online integration a new language model is created for every utterance, derived from reverse translation. Evidently these language models are strongly biased and thus a high LM scale factor is to be expected. Figure 1 illustrates the relationship of WER, λ and scale factors s on the DEV10 set, for both the OOD and ID models. Foreground LM is trained on 100-best MT output. λ denotes the weight on the background LM. Tried values of λ include 0.05, 0.5 and 0.95. $\lambda = 0.95$ implies only low contribution of the MT hypothesis.

As can be observed typical scale factors used for this task, e.g., $s = 13$, are suitable, regardless of the acoustic model used. For the interpolation factor no strong bias to either component seems to yield the best result, equal weighting appears to be sufficient. A small weight for the MT hypothesis language model reduces the WER significantly. At $s = 13$, the WER is 30.8% when $\lambda = 1$

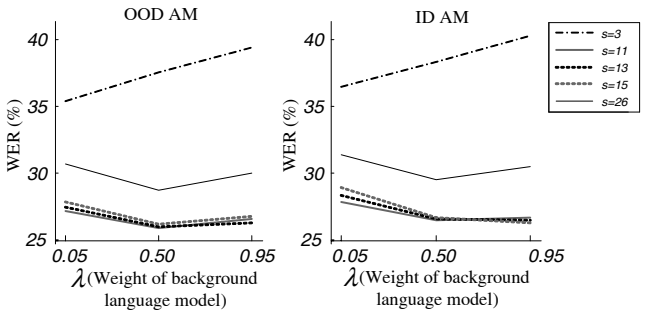

Fig. 1. *WER against interpolation weight with different scale factors on DEV10 (Foreground LM trained on* 100*-best MT system output*

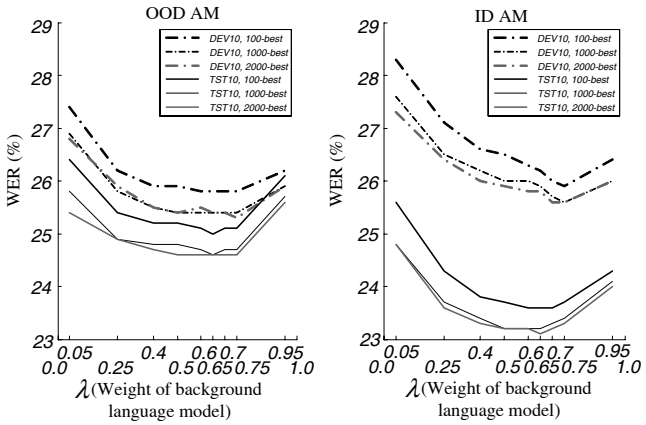

Fig. 2. *WER against interpolation weight with different N-best sizes on DEV10 and TST10*

(Table 3, ID-AM Unadapted), while $\lambda = 0.95$ already yields a reduced absolute WER of 26.4%, a reduction by 14.3% relative.

Figure 2 shows the dependence of interpolation factors on the test sets with different N-best list sizes. To increase the resolution, more values of λ are tested. The associated WER and perplexity numbers for the 2000-best case are included in Table 4. The effect of λ on the WER is quite consistent for different acoustic models, data sets, and N-best list sizes. The minimum WER is found for values of λ in the range between 0.65 and 0.75. On TST10, the overall best performance is obtained by use of the 2000-best utterance-specific foreground language model, with WERs 24.6% and 23.1% for OOD and ID AMs respectively. The associated perplexities (Table 2) are significantly reduced in contrast to offline integration.

The effect of N-best list size increase is significant. From Figure 2 one can observe a significant drop in error rate with an increase of N from 100 to 1000. Thereafter only modest changes are visible. Inspecting the TST10 results, for the OOD case, the WER changes from the baseline 31.3% (unadapted, Table 3) to 24.6%, for ID the reduction is from 29.6% to 23.1%. These amount to relative performance improvements by 21.4% and 22.0% respectively. Somewhat surprisingly these results are obtained with a weight of 65% on the background LM.

5. DISCUSSION

In this paper we have demonstrated how ASR performance can be improved with supervised information in a foreign language. The mismatch between ASR and MT systems can be as high as 72.7% WER, which implies that the output from the two systems are quite different. For the MT output, deletion and substitution rates are high, with 17.8% and 30.1% respectively, much higher than those observed for the ASR output (8.4% and 17.7% respectively). One reason is word-ordering and phrasing. Without the acoustic constraints, many paraphrases are observed in MT output. For instance, "adaptive imperatives" is substituted as "imperatives of adaptation" in one example utterance; "adaptive" counts as error twice because of mis-positioning.

Table 4. *Online adaptation with language model from 2000-best MT hypothesis outputs under different interpolation weight*

λ	OOD AM WER		ID AM WER		TST10
	DEV10	TST10	DEV10	TST10	Perplexity
0.05	26.8	25.4	27.3	24.8	85.47
0.25	25.9	24.9	26.4	23.6	51.92
0.40	25.5	24.7	26.0	23.3	48.12
0.50	25.4	24.6	25.9	23.2	48.15
0.60	25.5	24.6	25.8	23.2	49.85
0.65	25.4	24.6	25.8	23.1	51.43
0.70	25.4	24.6	25.6	23.2	53.63
0.75	25.3	24.6	25.6	23.3	56.67
0.95	25.9	25.6	26.0	24.0	96.19

Table 5. *Confusion pairs in ASR and MT system outputs*

	ASR	MT
"THAT"	it→that (25),	that→this (103),
	the→that (24),	that→which (66),
	that→the (19),	that→it (59),
	→that (96 del),	→that (203 del),
	→that (150 sub),	→that (467 sub),
"IS"	is→as (21),	is→are (16),
	is→was (13),	
Other	could→can (30),	so→therefore(32),
	our→are (11),	maybe→perhaps (13)
	one→a (11),	gamers→players (10)

Table 5 shows examples of the confusion word pairs between reference and hypothesis for ASR and MT systems. Among the top words in the confusion sets, the word "that" is much less preferred in MT output. Substitutions of 'is" with "was" in the ASR output may be triggered by phonetic or semantic reasons, while the reciprocal for MT output ("is" → "are") is clearly not phonetically similar. Many of the confusion pairs in the MT output are phonetically different synonyms, e.g., "so → therefore", "maybe → perhaps".

6. CONCLUSION

In this paper, we show integration of ASR and MT systems can lead to significantly improved ASR output. Online integration with utterance-level MT language models has given the best results, with relative reductions in WER by more than 20%. We observe that such improvements are possible because of significantly different error patterns between ASR and MT. The integration behaviour and consistency gives confidence that tighter coupling regimes may yield further improvements, even for completely unsupervised scenarios.

7. ACKNOWLEDEGMENT

This research was supported by Google and the EPSRC (Ref: EP/I034750/1).

8. REFERENCES

[1] Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker, "Overview of the IWSLT 2011 evaluation campaign," in *Proc. 8th International Workshop on Spoken Language Translation*, 2011, pp. 11–27.

[2] B. Zhou, L. Besacier, and Y. Gao, "On efficient coupling of ASR and SMT for speech translation," in *Proc. ICASSP*, 2007, vol. IV, pp. 101–104.

[3] R. Zhang and G. Kikui, "Integration of speech recognition and machine translation: Speech recognition word lattice translation," *Speech Communication*, vol. 48, pp. 321–334, 2006.

[4] P. R. Dixon, A. Finch, C. Hori, and H. Kashioka, "Investigation on the effects of ASR tuning on speech translation performance," in *Proc. 8th International Workshop on Spoken Language Translation*, 2011, pp. 167–174.

[5] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Proc. Interspeech*, 2005.

[6] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?," in *Proc. ICASSP*, 2011, pp. 5632–5635.

[7] C. M. Koolstra and J. W. J. Beentjes, "Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home," *Educational Technology Research and Development*, vol. 47, no. 1, pp. 51–60, 1999.

[8] Shahram Khadivi and Hermann Ney, "Integration of speech recognition and machine translation in computer-assisted translation," *IEEE Trans. Audio, Speech, Lang. Prcs*, vol. 16, no. 8, Nov 2008.

[9] M. Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, and A. Waibel, "Speech translation enhanced automatic speech recognition," in *Proc. ASRU*, 2005.

[10] Michael Paul, Marcello Federico, and Sebastian Stüker, "Overview of the IWSLT 2010 evaluation campaign," in *Proc. 7th International Workshop on Spoken Language Translation*, 2010, pp. 3–27.

[11] G. Saon and M. Picheny, "Lattice-based Viterbi decoding techniques for speech translation," in *Proc. ASRU*, 2007, pp. 386–389.

[12] A. Reddy, R. Rose, and A. Désilets, "Integration of ASR and machine translation models in a document translation task," in *Proc. Interspeech*, 2007.

[13] M. Paulik, "Machine translation enhanced automatic speech recognition," Universität Fridericiana zu Karlsruhe Master Thesis, 2005.

[14] M. Paulik, C. Fügen, S. Stüker, T. Schultz, T. Schaaf, and A. Waibel, "Document driven machine translation enhanced ASR," in *Interspeech*, 2005, pp. 2261–2264.

[15] S. Stüker, M. Paulik, M. Kolss, C. Fügen, and A. Waibel, "Speech translation enhanced ASR for european parliament speeches - on the influence of ASR performance on speech translation," in *Proc. ICASSP*, 2007.

[16] G. Lecorvé, J. Dines, T. Hain, and P. Motlicek, "Supervised and unsupervised web-based language model domain adaptation," in *Proc. Interspeech*, 2012.

[17] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication*, vol. 42, pp. 93–108, 2004.

[18] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, pp. 467–479, Dec 1992.

[19] "Technology entertainment design," `http://www.ted.com`.

[20] "Iwslt 2012 ted task specification," `http://hltc.cs.ust.hk/iwslt/index.php/evaluation-campaign/ted-task`.

[21] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, D. van Leeuwen, and V. Wan, "The 2007 AMI(DA) system for meeting transcription," in *Proc. MLMI*. 2007, Springer.

[22] Philipp Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. MT Summit 2005*, 2005, pp. 79–86.

[23] Hynek Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Soc.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[24] 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.

[25] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training.," in *Proceedings ICASSP'92*, 2002, pp. 105–108, MPE first paper.

[26] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst, "Moses: open source toolkit for statistical machine translation," in *Proc. ACL 2007*, 2007, pp. 177–180.

[27] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, "Findings of the 2012 workshop on statistical machine translation," in *Proc. WMT 2012*, June 2012, pp. 10–51.

[28] Franz Josef Och and Hermann Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, mar 2003.

[29] Philipp Koehn, Franz Josef Och, and Daniel Marcu, "Statistical phrase-based translation," in *Proc. NAACL '03*, 2003, pp. 48–54.

[30] Franz Josef Och, "Minimum error rate training in statistical machine translation," in *Proc. ACL 2003*, 2003, vol. 1, pp. 160–167.

[31] T. Hain, L. Burget, J. Dines, P. N. Garner, A. El Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "The AMIDA 2009 meeting transcription system," in *Proc. Interspeech 2010*, 2010, pp. 358–361.