The UEDIN English ASR System for the IWSLT 2013 Evaluation

Peter Bell, Fergus McInnes, Siva Reddy Gangireddy,
Mark Sinclair, Alexandra Birch, Steve Renals

School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK
{peter.bell,fergus.mcinnes,a.birch,s.renals}@ed.ac.uk,
{s.gangireddy,m.sinclair-7}@sms.ed.ac.uk

Abstract

This paper describes the University of Edinburgh (UEDIN) English ASR system for the IWSLT 2013 Evaluation. Notable features of the system include deep neural network acoustic models in both tandem and hybrid configuration, cross-domain adaptation with multi-level adaptive networks, and the use of a recurrent neural network language model. Improvements to our system since the 2012 evaluation – which include the use of a significantly improved n-gram language model – result in a 19% relative WER reduction on the `tst2012` set.

1. Introduction

We report on experiments carried out for the development of automatic speech recognition (ASR) systems on the English datasets of the International Workshop on Spoken Language Translation (IWSLT) 2013. We report our work on the new TED German task in an accompanying paper [1] since the development of the two systems was largely independent. Work on our machine translation system may be found in [2]. Significant changes to the English ASR system since 2012 include improvements to our baseline language models, described in Section 2.1, and the use of recurrent neural network language models, described in Section 2.2. The acoustic models are described in Section 3 – the main addition is that we now use deep neural networks in a hybrid configuration, and apply automatic voice activity detection to the `tst2013` test set.

2. Language modelling

The ASR system used Kneser-Ney smoothed N-gram language models for decoding and lattice rescoring, and a recurrent neural network (RNN) language model for a final rescoring stage based on N-best lists. These models are described in the subsections below.

This work was supported by the European Union under the FP7 projects inEvent (grant agreement 287872) and EU-Bridge (grant agreement 287658), and by EPSRC Programme Grant grant EP/I031022/1, *Natural Speech Technology*.

2.1. N-gram models

The N-gram language models were obtained by interpolating individual modified Kneser-Ney discounted LMs trained on the small in-domain corpus of TED transcripts and the larger out-of-domain (OOD) sources. The OOD sources were Europarl (v7), News Commentary (v7), News Crawl (2007 to 2011) and Gigaword (Fifth Edition).

The News Crawl and Gigaword sources in particular contained a wide variety of phenomena such as money amounts and other numerical expressions, abbreviations, and listed and tabulated information, which required normalisation to create data resembling spoken word sequences. Considerable effort was put into developing appropriate text normalisation scripts. Starting from the scripts used in LM training for the IWSLT 2012 evaluation, over 1000 lines of Perl code and 1400 abbreviation entries were added (expanding the original files by more than 50%). The processing applied to the data can be summarised as follows.

1. Remove documents that are not of type *story*, strip out markup and split text into sentences (required for Gigaword only).
2. Eliminate duplicate lines (common in some newswire sources, where multiple copies or variants of the same story may occur).
3. Convert Unicode characters and encodings for fractions, symbols etc into standard ASCII forms such as "1/4" (for subsequent conversion to words).
4. Filter out newswire datelines, e.g. "LONDON, Nov 2", and other extraneous material.
5. Normalise punctuation, abbreviations, units of measurement etc.
6. Convert numerical expressions to words.
7. Remove punctuation and convert to lower-case without diacritics.
8. Convert British to American English spellings and correct some common spelling errors.

The vocabulary for the ASR system was defined so as to include all words occurring in the in-domain training corpus (other than words which occurred only once and were not in a standard dictionary) and all words exceeding specified occurrence count thresholds in the OOD corpora, while remaining below the maximum of 64K words imposed by the version of HDecode in use here. The vocabulary size was 62,522.

Initialisms included in the vocabulary were treated as single words for LM purposes, e.g. "u.s." (with the dots retained to distinguish them from words such as "us"). Once the vocabulary had been defined, out-of-vocabulary initialisms were broken into single letters, e.g. "m. f. n.", so as to be modelled as sequences of in-vocabulary words (letter names) rather than treated as OOV.

In view of the mismatch in content and style between the target domain (TED talks) and the OOD data, a data selection process [3, 4] was applied to the OOD corpora to obtain an appropriate subset of data for LM training. The set of out-of-domain data D_S was chosen by computing a cross-entropy difference (CED) score for each sentence s:

$$D_S = \{s | H_I(s) - H_O(s) < \tau\} \quad (1)$$

where $H_I(s)$ is a cross-entropy of a sentence with a LM trained on in-domain data, $H_O(s)$ is a cross-entropy of a sentence with a LM trained on a random subset of the OOD data of similar size to the TED corpus, and τ is a threshold to control the size of D_S

Language models were trained on the in-domain and OOD data using the SRILM toolkit [5], and were interpolated with weights optimised on the TED development set (dev2010 and tst2010: total 44,456 words).

Perplexities on the development set with 3-gram and 4-gram models trained on the TED corpus and selected OOD data are shown in Table 1. Selecting 25% of the OOD sentences yielded an OOD training set of 751M words; setting the CED threshold to 0 gave a smaller but more targeted set of 312M words, which gave a lower perplexity on the TED data than the 751M word set when used alone to train the LM, but a slightly higher perplexity after interpolation with the TED LM. The perplexities obtained here are substantially lower than the values of 160 (3-gram) and 159 (4-gram) with the LMs used in our IWSLT 2012 system [6], which were trained using a much smaller set of OOD data with no CED filtering.

The LMs finally used in the ASR system were the TED+312MW trigram model (for decoding) and the TED+312MW 4-gram model (for lattice rescoring). The amounts of data from the respective sources used in these LMs are shown in the "Selected" column of Table 2. Comparison with the total sizes of the source corpora (after text normalisation) given in the preceding column shows that the proportion of data selected by the CED criterion ranged from 8% for the Gigaword corpus to 15% for News Commentary.

Language model	Perplexity
TED 3-gram	183.2
OOD (312MW / 751MW) 3-gram	133.5 / 138.3
TED+OOD (312MW / 751MW) 3-gram	125.1 / 124.9
TED 4-gram	179.9
OOD (312MW / 751MW) 4-gram	123.9 / 126.4
TED+OOD (312MW / 751MW) 4-gram	114.9 / 113.4

Table 1: Perplexities of N-gram language models on TED development set.

Corpus	Total	Selected
TED	2.4M	2.4M
Europarl	53.1M	6.3M
News Commentary	4.4M	0.7M
News Crawl	693.5M	72.9M
Gigaword	2915.6M	232.9M
OOD total	3666.6M	312.8M

Table 2: Numbers of words in LM training sets.

2.2. RNN models

Neural network language models have shown to consistently improve the word error rates (WER) of LVSCR tasks [7, 8, 9]. For this year's evaluation, we investigated the effectiveness of RNN LMs for TED lecture transcription. To study the effectiveness of RNNs we rescored the n-best hypothesis using RNNs trained on in-domain and different subsets of out-of-domain (OOD) data, shown in Table 3, selecting according to the CED score as in Section 2.1 In-domain data consists of 2.4M tokens. Since it is very difficult to train the RNNs on large amounts of OOD data, we restrict the maximum size of OOD data to 30M.

The number of hidden neurons ranged from 300 to 500 and number of classes in the output layer was 300. Models are trained using RNN training tool of [10]. Table 4 shows the perplexity (PPL) and WER on on development data provided by IWSLT evaluation campaign. We can observe that rescoring the n-best hypothesis with the RNNs reduce the WER by 0.8%. We choose the best model from this experiments to rescore the n-best hypothesis from `tst2011`, `tst2012` and the `tst2013` test sets. The interpolation weight between n-gram and RNNLM is optimised on devel-

Table 3: Subsets of OOD data

#Words	#Sentences	Threshold(τ)
5M	664.2K	-1.14
10M	1156.7K	-0.963
15M	1596.7K	-0.862
20M	2011.3K	-0.79
25M	2412.6K	-0.733
30M	2792.4K	-0.687

Table 4: Perplexity and WER on development data

Tokens	Vocabulary	PPL	WER(%)
n-gram	-	-	15.6
7.4M	47.7K	171.56	15.2
12.4M	54.8K	161.66	15.2
17.4M	61.7K	147.17	15.0
22.4M	68K	142.22	14.9
27.4M	74.3K	133.5	14.8
32.4M	80K	**126.0**	**14.8**

opment data, to minimise WER.

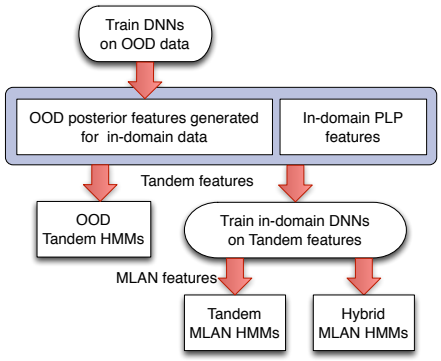

Figure 1: Tandem and hybrid MLAN training

3. Acoustic modelling

For the acoustic modelling components of the system, we used a setup identical to that described in [11], where more details may be found. Briefly, we used a combination of tandem and hybrid deep neural network (DNN) systems trained on a corpus of in-domain TED talks, incorporating out-of-domain data of multi-party meetings from the AMI corpus using the multi-level adaptive networks (MLAN) technique [12]. Compared to our 2012 system, the main addition is the use of DNNs with MLAN features in the hybrid framework. We describe this further below. Additionally, unlike earlier test sets from the IWSLT evaluation, the 2013 test set was not provided with a manually derived segmentation; we therefore employed an automatic segmentation system, described in Section 3.3.

3.1. Training data

For in-domain training data, we used 813 TED talks recorded prior to the end of 2010. The talks were segmented and aligned to the crowd-sourced transcriptions available online using a lightly-supervised technique described in [13]. This produced 143 hours of labelled speech segments for use in acoustic model training. Additionally, we used 127 hours of out-of-domain data from the AMI Corpus of multi-party meetings[1] using a setup based on [14]. This data is not in general well-matched to the TED-domain. The OOD data was not used directly in acoustic model training, but used to generate out-of-domain neural network features for the in-domain data.

3.2. Deep neural network systems

For our 2012 system, we used neural networks within the tandem framework [15, 16], using DNNs to generate log probabilities over monophones. The monophone probabilities are decorrelated and projected to 30 dimensions, then augmented with the original acoustic features to give a total feature vector of 69 dimensions. These vectors are used for standard HMM-GMM training. Additionally in this year's system, we used DNNs in a hybrid configuration, generating posterior probabilities over tied-state triphones, as proposed in [17]. These are converted to pseudo-likelihoods for use in the decoder.

Both tandem and hybrid nets used PLP input features with 9 frames of temporal context. For the tandem systems, the final nets used had four hidden layers with 1024 hidden units per layer; the hybrid systems used six hidden layers with 2048 hidden units per layer. The tandem nets had an output layer of size 46; the size of the output layer of the hybrid nets varies according to the number of tied states, which resulting from clustering with a GMM; it was typically around 6,000. The nets were trained with a tool based on the Theano library [18] on NVIDIA GeForce GTX 690 GPUs. For the tandem systems, we applied speaker adaptive training of the GMMs using CMLLR [19] regression class trees with 32 classes. For the hybrid systems, we performed adaptation of the input feature space at training and test time using a global CMLLR transform for each speaker. Tandem systems were discriminatively trained with MPE.

As in the 2012 system, we incorporated out-of-domain data using the MLAN technique. Neural networks were trained on the AMI corpus and the resulting nets used to generate posterior features for each utterance in the TED corpus. Th2ese neural net features are known to provide a degree of domain-independence [20]. In the MLAN scheme, the OOD features are augmented with the original acoustic features and a further DNN is trained on these features, allowing further adaptation to the target domain. This second adaptive network may be used to generate tandem features, or used in a hybrid system. The possible configurations are illustrated in Figure 1.

3.3. Voice activity detection

The voice activity detection component of the system comprises a GMM-HMM based model which is used to perform a Viterbi decoding of the audio. The HMM has 2 classes: speech and non-speech. These are modelled with diagonal-covariance GMMs with 12 and 5 mixtures respectively. We allow more mixture components for speech to

[1] http://www.amiproject.org/

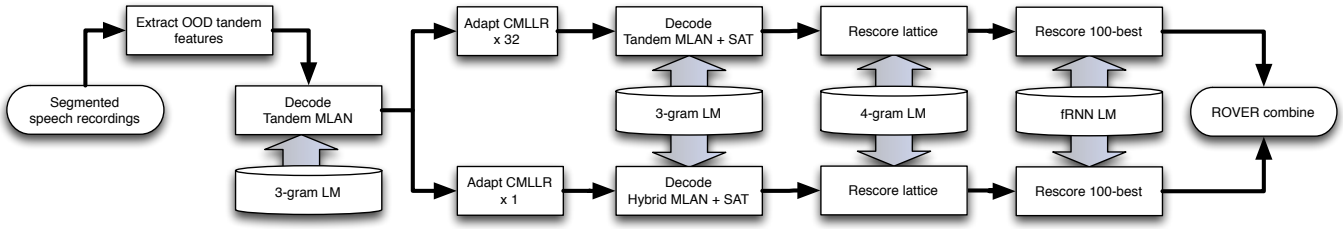

Figure 2: The full decoder architecture

cover its greater variability. Features are calculated every 10ms from a 30ms analysis window and have a dimensionality of 14 (13 PLPs and energy). Models were trained on 70 hours of scenario meetings data from the AMI corpus using the provided manual segmentations as a reference. To avoid over segmentation a minimum duration constraint of 50ms is enforced by inserting a series of 50 states per class that each have a transition weight of 1.0 to the next, the final state has a self transition weight of 0.9.

4. Decoder architecture

Figure 2 shows the complete decoding architecture. After an initial pass, used to generate transcripts to estimate speaker transforms, we operate two parallel decoding sequences for the tandem and hybrid acoustic models. For each model, the complete process consists of a decoding with the trigram LM using HTK's HDecode[2]. Lattices output from the this pass were rescored using the 4-gram LM, generating 100-best lists, which were rescored with the final interpolated RNN LM. Finally, the one-best outputs from tandem and hybrid systems are combined at the hypothesis level using ROVER.

5. Results

In this section we first present development results from individual components of the complete system pipeline. Table 5 shows results using the manual segmentations provided for earlier evaluations. The results may differ slightly from official results due to variations in scoring procedure. It may be observed that there is no clear winner out of the tandem and hybrid systems; however, they are clearly complementary as system combination consistently yields improved performance.

The trends are similar when the automatic segmentation is used, shown in Table 6. When the automatic segmentation is used there is a deterioration in performance of up to 3% WER. Some of this may be attributed to an increase in insertion and deletion errors of the result of segmentation errors; however, an additional source of error, particularly affecting the RNN LM, is that the automatic segmenter typically results in shorter segments, not divided along semantic lines as the manual version is, resulting in reduced language modelling power, since we do not propagate LM probabilities across segment boundaries. Note that the results with the RNN model are available only for a subset of experiments as this component of the system was not fully automatic at the time of system development.

System	dev2010	tst2010	tst2011
Tandem MLAN	15.9	14.1	11.2
+ 4gram	15.6	13.6	10.8
+ RNN	-	-	10.4
Hybrid MLAN	15.6	13.9	11.5
+ 4gram	15.2	13.5	11.3
+ RNN	-	-	10.5
ROVER combination			
4gram	**14.7**	**12.6**	10.3
+ RNN	-	-	**9.9**

Table 5: *Development system results with manual segmentation (WER%)*

System	dev2010	tst2010	tst2011
Tandem MLAN	18.8	17.6	14.9
+ 4gram	18.4	17.2	14.5
+ RNN	17.6	16.6	-
Hybrid MLAN	18.6	17.4	14.6
+ 4gram	18.4	17.2	14.3
+ RNN	17.6	16.7	-
ROVER combination			
4gram	17.6	16.2	**13.2**
+ RNN	**17.0**	**16.1**	-

Table 6: *Development system results with automatic segmentation (WER%)*

Finally, we provide the official results from the 2013 evaluation in Table 7. Automatic segmentation is used only for `tst2013` set. It is notable that the WER is substantially higher on this set than on the other development and evaluation sets. A preliminary analysis suggests that this is probably not due to problems with the segmentation, as insertion and deletion errors do not make up a noticeably higher proportion of the total errors than for the other test sets. Over the talks, the WER ranges from 9% to 48%, suggesting that

[2] `http://htk.eng.cam.ac.uk`

	tst2011	tst2012	tst2013
Primary system	10.2	11.6	22.1

Table 7: *Official system results from the 2013 evaluation (WER%)*

perhaps this year's test set contains a more diverse range of acoustic conditions.

6. Machine translation

We applied machine translation to the ASR output. Details may be found in the accompanying paper [2]. Table 8 compares MT performance for various inputs from the ASR system. Note that performing translation from a confusion network containing multiple ASR hypotheses resulted in worse results that using the one-best output. We are investigating the reasons for this – one theory is that, due to the generally low WER of the systems, the alternative hypotheses are rarely correct, often simply indicating OOV errors when they have high acoustic scores. Table 9 presents, for reference, the official 2013 BLEU results comparing, as inputs, the use of our best system, and the transcription by the IWSLT organisers.

ASR input	en-fr
1-best	22.9
1-best punctuated	24.1
Confusion net	18.4

Table 8: Cased BLEU results for models when tuned and tested on ASR output in different formats.

	en-fr
Edinburgh ASR system	22.45
IWSLT ASR system	23.00

Table 9: Official test 2013 cased BLEU results for 1Best SLT input. The Edinburgh ASR system input was our primary system.

7. Conclusions

We have described our ASR system for the English 2013 IWSLT evaluation Improvements to our system since the 2012 evaluation result in relative WER reductions of 17% 19% on the `tst2011` and `tst2012` sets respectively. The use of RNN LMs does not give improved performance on the `tst2013` set, a result that is probably due to the shorter utterances derived from the automatic segmentation.

Improvements planned for future systems include the use of neural network based voice activity detection, and the pooling of German and English audio data in multi-condition DNN training, whereby both systems are trained simultaneously, sharing lower layers of the network. We also plan to apply talk-level language model adaptation.

8. References

[1] J. Driesen, P. Bell, and S. Renals, "Description of the UEDIN system for german ASR," in *Proc. IWSLT*, 2013.

[2] A. Birch, N. Durrani, and P. Koehn, "Edinburgh SLT and MT system description for the IWSLT 2013 evaluation," in *Proc. IWSLT*, Heidelberg, Germany, 2013.

[3] R. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proc. ACL Conference Short Papers*, Uppsala, 2010, pp. 220–224.

[4] H. Yamamoto, Y. Wu, C. Huang, X. Lu, P. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The NICT ASR system for IWSLT 2012," in *Proc. International Workshop on Spoken Language Translation*, Hong Kong, Dec. 2012.

[5] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proc. ICSLP*, vol. 2, 2002, pp. 901–904.

[6] E. Hasler, P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski, "The UEDIN systems for the IWSLT 2012 evaluation," in *Proc. International Workshop on Spoken Language Translation*, Hong Kong, Dec. 2012.

[7] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, pp. 1137–1155, 2003.

[8] H. Schwenk, "Continuous space language models," *Computer Speech & Language*, vol. 21, no. 3, pp. 492–518, 2007.

[9] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH*. ISCA, 2010, pp. 1045–1048.

[10] T. Mikolov, M. Karafiát, L. Burget, J. Černokcý, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010.

[11] P. Bell, H. Yamamoto, P. Swietojanski, Y. Wu, F. McInnes, C. Hori, and S. Renals, "A lecture transcription system combining neural network acoustic and language models," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[12] P. Bell, M. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. Woodland, "Transcription of multi-genre media archives using out-of-domain

data," in *Proc. IEEE Workshop on Spoken Language Technology*, Miama, Florida, USA, Dec. 2012.

[13] A. Stan, P. Bell, and S. King, "A grapheme-based method for automatic alignment of speech and text data," in *Proc. IEEE Workshop on Spoken Language Technology*, Miama, Florida, USA, Dec. 2012.

[14] T. Hain, L. Burget, J. Dines, P. Garner, F. Grezl, A. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[15] H. Hermanksy, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1630.

[16] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features in LVCSR," in *Proc. Interspeech*, 2004.

[17] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[18] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. SciPy*, June 2010.

[19] M. Gales, "Maximum likelihood linear transforms for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 75-98, 1998.

[20] A. Stolcke, F. Grézl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006.