Combining Feature and Model-Based Adaptation of RNNLMs for Multi-Genre Broadcast Speech Recognition

Salil Deena, Madina Hasan, Mortaza Doulay, Oscar Saz and Thomas Hain

Department of Computer Science

University of Sheffield, UK

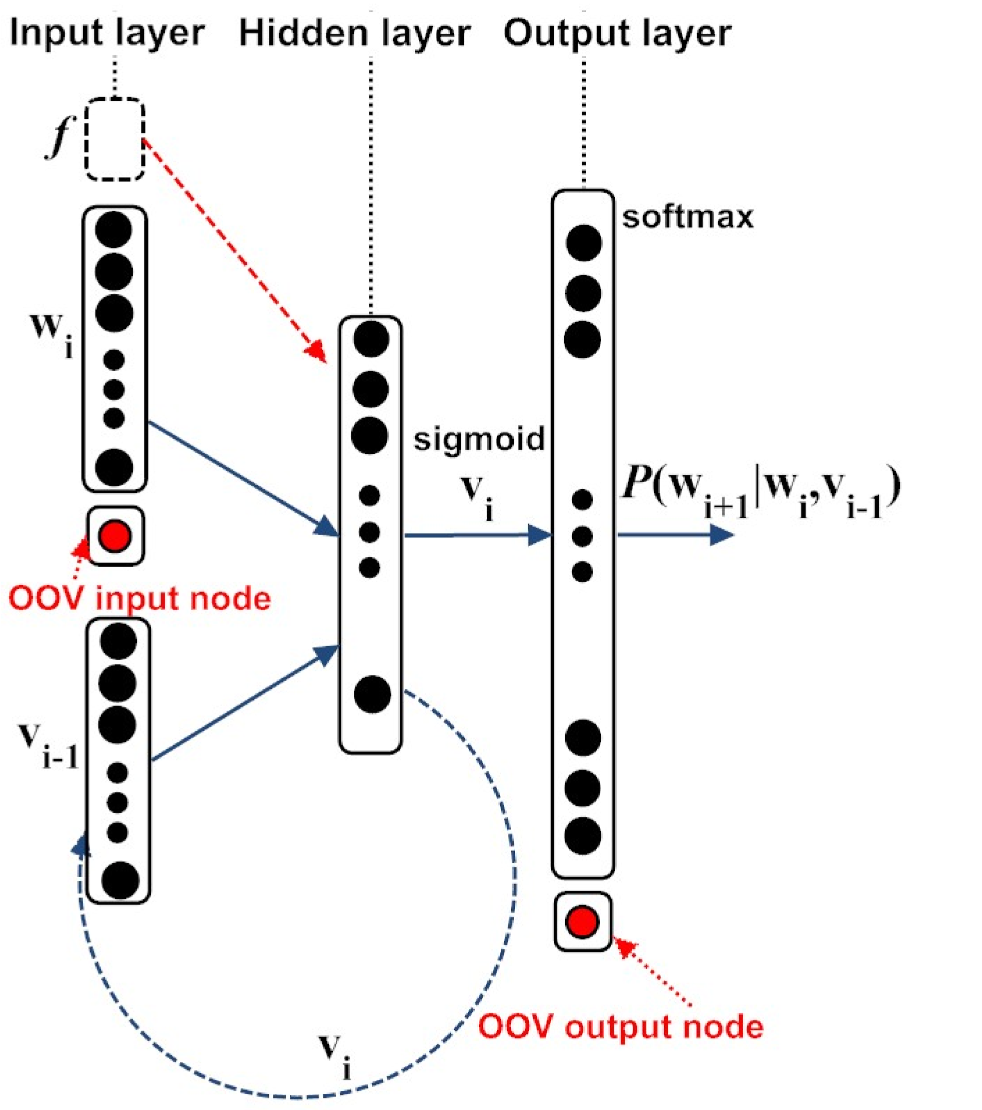
{s.deena,m.hasan,m.doulaty,o.saz,t.hain}@sheffield.ac.uk

The University Of Sheffield.

Natural Speech Technology

Recurrent Neural Network LMs

- RNNLMs [1] outperform n-grams in many ASR tasks due to the following:
  - RNNLMs allow robust parameter estimation through a continuous-space representation
  - RNNLMs can model longer context dependencies than n-grams
- Recurrent layer can represent full history $< w_{i-1}, \ldots, w_1 >$ for word w_i using concatenation of word w_{i-1} and remaining context vector v_{i-2}



Multi-Genre Broadcast Data

- The data were BBC broadcasts and subtitles **officially distributed for the MGB challenge** [2]
  - Acoustic training data: 2,193 shows with 1,580 hours of audio and lightly supervised transcripts
  - Language training data: 648M words from historical subtitles ($LM1$) and 10M words from 2,193 training shows ($LM2$)

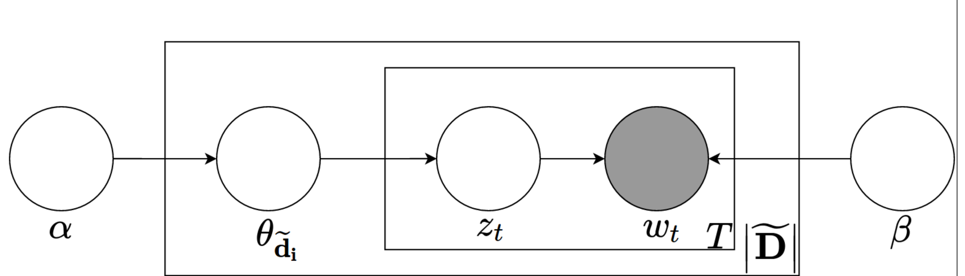
Subtitles	#sentences	#words	#unique words
$LM1$ (1979-2008)	72.9M	648.0M	752,875
$LM2$ (Apr/May '08)	633,634	10.6M	32,304

  - Development data: 47 shows with 28 hours of audio
  - 8 genres: **Advice, children's, comedy, competition, documentary, drama, events and news**

	Train		Development	
Genre	Shows	Time	Shows	Time
Advice	264	193.1h.	4	3.0h.
Children's	415	168.6h.	8	3.0h.
Comedy	148	74.0h.	6	3.2h.
Competition	270	186.3h.	6	3.3h.
Documentary	285	214.2h.	9	6.8h.
Drama	145	107.9h.	4	2.7h.
Events	179	282.0h.	5	4.3h.
News	487	354.4h.	5	2.0h.
Total	2,193	1580.5h.	47	28.3h.

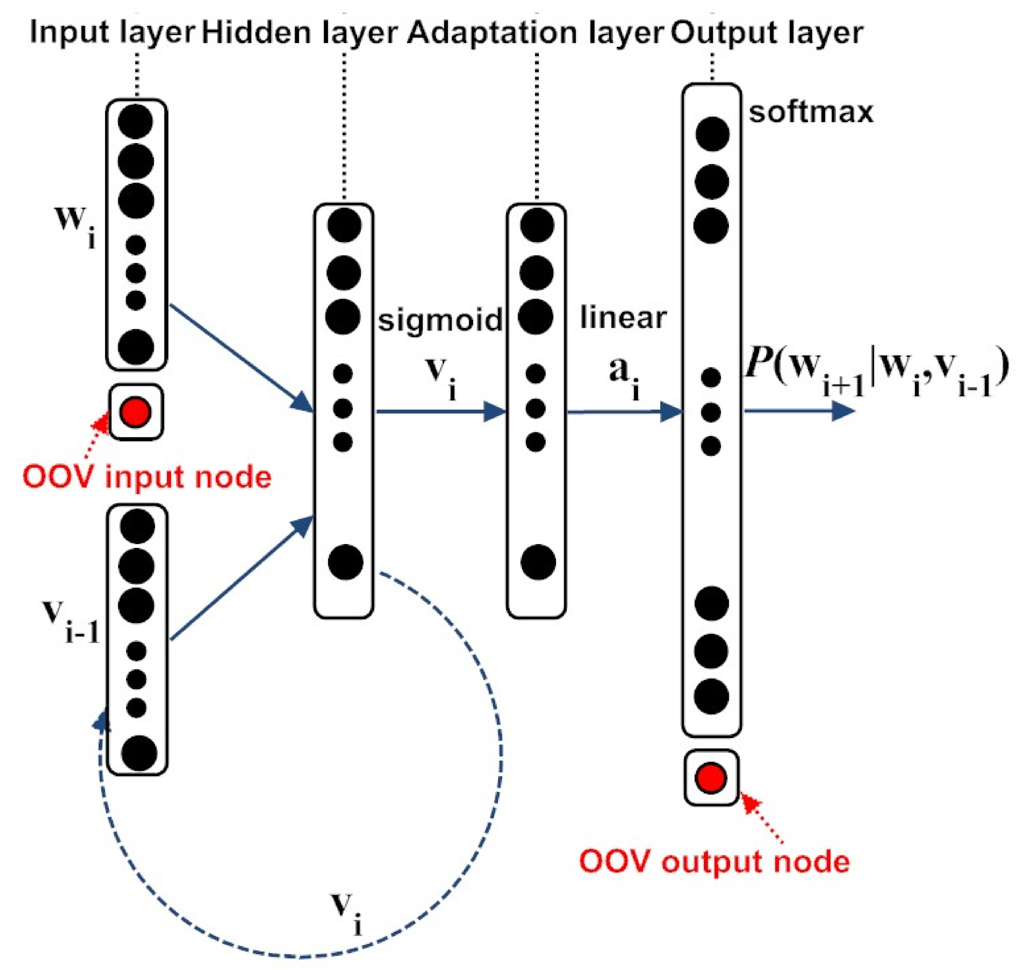
Feature-Based RNNLM Adaptation

- Append a feature vector f to the input of the RNNLM
- Two features used in this work:
  - Genre 1-hot auxiliary codes, which represent genre as a 1-of-K vector
  - Latent Dirichlet Allocation (LDA) auxiliary features, obtained by computing Dirichlet posteriors over latent topics after training models by first computing term frequency-inverse document frequency (TF-IDF) vectors on text data



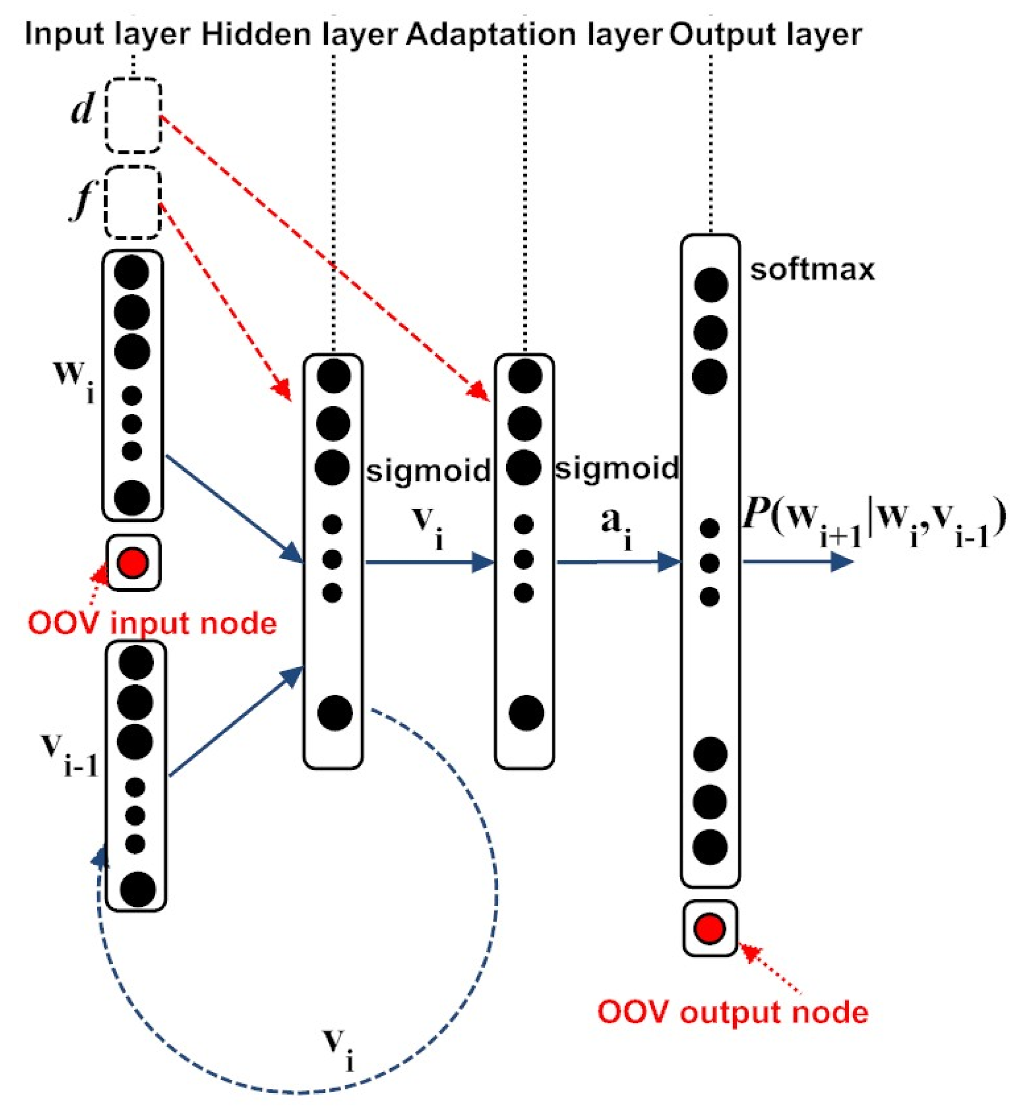
Model-Based RNNLM Adaptation

- The following two model adaptation techniques were compared:
  - Genre fine-tuning, which involves further training a RNNLM model using genre-specific text data
  - Linear hidden network (LHN) adaptation layer, which introduces a linear multiplicative transform to the hidden layer to adapt to genre-specific text data
    - The adaptation layer is cascaded between the hidden and output layers respectively
    - The weights connecting the adaptation and the output layers are initialised using the identity matrix
    - At time of adaptation, only those weights are updated whilst keeping the rest of the network unchanged
    - We are the first to apply LHN adaptation layer to RNNLMs



Hybrid RNNLM Adaptation

- The following two hybrid adaptation techniques were proposed:
  - Fine-tuning feature-based RNNLM, which involves further training LDA adapted RNNLMs on genre-specific text, thus combining topic and genre domain representations
  - Feature-Based RNNLM with adaptation layer, which involves having an adaptation layer with genre 1-hot features input, together with LDA feature input at the hidden layer
    - Feature-based adaptation layer provides an additive transform through bias adaptation whilst LHN adaptation layer provides a multiplicative transform of the weights at the hidden layer
    - Additive transform was shown to be less prone to overfitting in acoustic domain
    - Overfitting can happen when amount of domain-specific data is small, which is the case for genres such as comedy and drama



Semi-supervised RNNLM Adaptation

- Genre labels are available for $LM2$ text but not for larger $LM1$ text
- In order to make the best of the hybrid adaptation techniques, need to generate genre labels for $LM1$ text
  - LDA features with 1024 topics were extracted from $LM2$ text and a SVM classifier was used to predict genre from the LDA features
  - Classification accuracy obtained on held-out development data was 94.79%
  - Same LDA+SVM model was used to predict genre labels for $LM1$ text, which was then used for hybrid RNNLM adaptation

Experiments and Results

- DNN-GMM-HMM Bottleneck acoustic models [3]
- 200k vocubulary used to build baseline 4-gram LM on $LM1 + LM2$ text
- Trained both $LM1$ and $LM1 + LM2$ RNNLMs
- Used a modified version of RNNLM toolkit [4]
- Gained improvements with hybrid RNNLM adaptation compared to previous work [5]
- LDA topic features and genre 1-hot features were found to be complimentary
- LHN Adaptation Layer gives improvements over fine-tuning
- Adaptation layer with additive transform gives better results than with multiplicative transform

		Global	
System	Adaptation	PPL	WER
$LM1\&LM2$ 4-gram and RNNLM baselines			
4-gram	None	100.1	30.1
4-gram+RNNLM interp (lattice rescoring)	None	88.6	29.4
4-gram+RNNLM interp (n-best rescoring)	None	88.6	29.2
$LM1\&LM2$ 4-gram + $LM2$ RNNLM (0.3 interp) with RNNLM adaptation			
RNNLM Baseline	None	93.7	29.8
Genre feat. at hidden layer	Feature	91.9	29.7
Genre fine-tuning	Model	90.6	29.6
Genre LHN adaptation layer fine-tuning	Model	90.4	29.6
Genre feat. at adaptation layer	Hybrid	90.7	29.6
LDA feat. at hidden layer	Feature	88.3	29.5
LDA feat. at hidden layer and genre fine-tuning	Hybrid	86.7	29.4
LDA feat. at hidden and genre feat. at adaptation layer	Hybrid	86.9	29.2
$LM1\&LM2$ 4-gram + $LM1\&LM2$ RNNLM (0.5 interp) with RNNLM adaptation			
RNNLM Baseline	None	88.6	29.2
Genre feat. at hidden layer	Feature	85.4	29.0
Genre fine-tuning	Model	82.2	29.0
Genre LHN adaptation layer fine-tuning	Model	81.9	28.9
Genre feat. at adaptation layer	Hybrid	83.4	28.7
LDA feat. at hidden layer	Feature	81.6	28.7
LDA feat. at hidden layer and genre fine-tuning	Hybrid	80.4	28.7
LDA feat. at hidden and genre feat. at adaptation layer	Hybrid	79.4	28.6

References

[1] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." *INTERSPEECH'10: Proc. of the 11th Annual Conference of the International Speech Communication Association*, vol. 2, p. 3, 2010.

[2] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Webster, and P. Woodland, "The MGB challenge: Evaluating multi–genre broadcast media transcription," in *ASRU'15: Proc. of IEEE workshop on Automatic Speech Recognition and Understanding*, Scottsdale, AZ, 2015.

[3] O. Saz, M. Doulaty, S. Deena, R. Milner, R. W. M. Ng, M. Hasan, Y. Liu, and T. Hain, "The 2015 Sheffield system for transcription of multi–genre broadcast media," in *ASRU'15: Proc. of the IEEE Automatic Speech Recognition and Understanding workshop*, 2015.

[4] X. Chen, X. Liu, Y. Qian, M. Gales, and P. Woodland, "CUED-RNNLM – An Open-Source Toolkit for Efficient Training and Evaluation of Recurrent Neural Network Language Models," in *ICASSP'16: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[5] X. Chen, T. Tan, X. Liu, P. Lanchantin, M. Wan, M. J. F. Gales, and P. C. Woodland, "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition," in *INTERSPEECH'15: Proc. of the 16th Annual Conference of the International Speech Communication Association*, 2015, pp. 3511–3515.