Babelon

2016 summary report

Prepared by:

Brno University of Technology

Martin Karafiat, Mirko Hannemann, Igor Szoke, Frantisek Grezl,
Karel Vesely, Honza Cernocky,

1 Highlights/Significant Technical Achievements This Period

- Adding Y4 languages into multilingual NN feature extractor training data gives 0.3-0.5% absolute improvement.

- Adding noise variants into multilingual NN feature extractor fine-tuning gives 0.5-0.7% even if final classifier was not trained on noises.
- Silence reduction in GMM training gives a small gain in basic ML system. Slightly bigger, 0.2-0.3% absolute improvement, was generated by silence reduced GMM in CMLLR estimation for DNN features.

- Adding Sequence Summarizing Neural Networks (SSNN) output as auxiliary information to standard DNN training gives 0.5% absolute gain.

2 Other Actions Worked on This Period

- Fine-tune 28L Multilingual Stacked Bottle-Neck NN to all Y4 languages and delivering CMLLR adapted output from 1StageNN to the team.

3 Concise Description of Each of the Experiments Being Done on the Language and its Outcome

- Self-stabilized NN - inspired by Jasa Droppo ICASSP 2016 work:

  - Adding learnable scalar parameter "β" into each layer:

$$h^i = f\left(W^i * h^{i-1} + b^i \right) \rightarrow h^i = f\left(\exp(\beta) * W^i * h^{i-1} + b^i\right)$$

  - They presented fast training without special care about LearningRate.

  - Tested on Javanese GMM non-adapted system. The Self-stabilized NN approach was applied on feature extraction level with SBN NN. The output tested functions were Sigmoid and Re-Lu.

Output fce	Normal train - XE	Self-Stabilized - XE
Sigmoid	61.3 (30 epochs)	62.0 (13 epochs)
Re-Lu	61.6 (27 epochs)	61.4 (14 epochs)

  - Faster convergence but degradation with sigmoid output function. The more investigation is needed.

- Multilingual NN based feature extraction.

  - Multilingual NN had Convolutive Stacked Bottle-Neck architecture. We were experimenting with:
    - Adding more languages into Multilingual NN training.
    - Effect of fine-tuning into target language.
    - Effect of extension of fine-tuned data for noise and speed variants. The additional data were generated by BBN. The 2 copies were added for each conversational waveform.
    - Effect of multiple GPU training.
  - Tested on Javanese GMM non-adapted system.

NN train data	**Vanilla Mult**	**Fine-tuned**	**Noises**
17L	61.4	59.5	
17L+4nonBabel (4GPU)	61.6	59.5	58.8
24L+4nonBabel (3GPU)	61.1	59.2	58.6
24L+4nonBabel (1GPU)	60.6	-	-

  - No gain by adding non-Babel data.
  - 0.3-0.5% absolute improvement by adding Y4 languages.
  - About 1% absolute improvement by fine-tuning into target language.
  - 0.5-0.7% absolute improvement by adding noise variants.
  - Next, we observed performance degradation by multiple GPU training of BN neural networks.

- Multilingual NN based feature extraction - DNN systems:
  - The features generated from 1stageNN from previous set of experiments (28Languages NN, fine-tuned into noised Javanese) were used for DNN system training. The DNN was trained only on clean data – no noise variants.
    - The 1stageNN features were processed by CMLLR based speaker adaptation. The single CMLLR is used, therefore we investigated an effect of reduction of silence. The training segmentation for GMM was processed by silence removal on the edges – 30frames of silence was allowed.

1stage BN Features	**DNN-XE**	**DNN-sMBR**

Original segmentation for GMM-CMLLR	52.4	50.6
Silence limited segmentation for GMM-CMLLR	52.2	50.3

  - 0.2-0.3% absolute gain from estimation of CMLLR with silence reduced GMMs.

- Building DNN systems for all languages.

  - Comparing old version of features based on 17 lang. Multilingual NN and latest version based on 28 languages, fine-tuning to noised data and silence reduced CMLLR.

Lang	17L based NN features WER[%]		28L based NN features WER[%]	
	XE	sMBR	XE	sMBR
Javanese	52.5	?	52.2	50.3
Pashto	50.0	48.5	47.8	46.2
Amharic	44.5	42.7	42.1	40.2
Mongolian	52.9	50.6	48.9	46.3
Igbo	59.0	57.1	56.9	54.2
Dholuo	41.1	38.8	38.1	36.1
Guarani	47.7	?	45.2	?

  - 2-4% gains from new features.

- Testing of VAD

  - Following experiment is comparing effect of silence reduction in GMM training similarly to previous experiment with CMLLR adaptation in DNN system.

  - Next the systems degradations caused by automatic VAD based segmentation were investigated for all languages. The VAD was trained on year 1+2 (11) languages.

  - The GMMs were trained on MultRDTv1 features without speaker adaptation.

Lang	Original training segmentation	Silence limited

	WER[%]		segmentation WER[%]	
	dev	**dev_VAD**	**dev**	**dev_VAD**
Javanese	59.7	60.4	59.4	60.0
Pashto	53.0	54.3	52.8	54.2
Amharic	49.1	59.9	48.5	49.4
Mongolian	58.3	68.9	58.1	58.6
Igbo	63.3	63.7	63.2	63.6
Dholuo	44.7	45.3	44.6	45.1
Guarani	52.0	52.9	51.8	52.6

  - Small improvement by removing silence in GMM training
  - Automatic segmentation is causing less than 1% degradation.

- DNN system with sentence summarizing information
  - The idea is in training classical DNN together with sentence summarizing NN (SSNN). Output of SSNN is averaged across the sentence and added to some layer of main DNN. This work was investigated during JSALT workshop 2015.
  - This technique was tested on Javanese. The DNNs were trained on top multilingual features based on 21L NN.
  - Two approaches were investigated:
    - Joint training of SSNN and main DNN.
    - After joint training the SSNN was cut-out from architecture and used as a sentence based feature extractor. The SSNN output can be added as auxiliary information to input of new DNN training in similar way as i-vector.

Experiment	**DNN-XE**	**DNN-sMBR**
Baseline	52.4	50.5
Joint training with SSNN	52.6	-
SSNN as i-vectors (slightly different architecture than above)	51.9	50.0

  - 0.5% absolute gain by adding of SSNN output as auxiliary information.

- Residual NN

  - Residual networks consists of residual units which can be expressed as:

$$y = h(x) + \mathrm{F}(\mathrm{x}, \mathrm{W}),$$

where $h(x)$ is an identity mapping function.

  - Two main advantages of this network are:

    1. Deeper Network can be trained as the gradient does not vanish even when the weights are small as the errors through residual connections are summed after each layer.

    2. This model does not require pre-training and shows comparable performance to DBN models.

  - The experiments run on Javanese. The DNNs were trained on top of multilingual features from 17L NN.

Type	N Layers	hidden dim	bn dim	Activation function	WER[%]
Baseline DNN	6	2048	0	Sigmoid	53.6
Resnet	4	512	256	ReLU	53.2
ResNet with Batch Normalization	3	512	256	ReLU	53.0

  - Over 0.5% absolute gain was observed but further investigation is needed.

4 Progress on Languages (report waypoints when possible)

5 Issues/Problems and Proposed Solutions (potential work plan risks and course corrections)

6 Babel-related Scholarly Activities including papers published and presentations given during this period (provide full reference for the material, and upload the item to SharePoint at least 2 weeks prior to publication or presentation; don’t forget the acknowledgment)

Presentation of 2papers on SLTU 2016:

    - Frantisek Grezl, :“*Bottle-neck feature extraction structures for multilingual training and porting*”
    - Frantisek Grezl, Ekaterina Egorova, Martin Karafiat: “*Study of large data resources for multilingual training and system porting*”

7 Significant Anticipated Activities and Opportunities Next Period (next month)

- Further experiments with CNTK toolkit – multilingual LSTM training and porting to target language.
- Further experiments with residual and SSNN.
- Experimenting with speech enhancement - WPE dereverberation.