Aalto system for the 2017 Arabic multi-genre broadcast challenge

Aalto University
School of Electrical Engineering

Peter Smit, Siva Reddy Gangireddy, Seppo Enarvi, Sami Virpioja, Mikko Kurimo
Department of Signal Processing and Acoustics, Aalto University, Finland

MGB CHALLENGE

The challenge

- Arabic MGB-3 transcription challenge: build a speech-to-text system for transcribing Egyptian Arabic.
- Arabic MGB-2 (re-run): transcription system for Modern Standard Arabic (MSA).
- Audio data: 1200 hrs MSA + 6 hrs Egyptian Arabic. LM data: 121M tokens.

Results

	MGB-3		MGB-2	
	Dev.	Eval.	Dev.	Eval.
Primary	**28.2**	**29.3**	**14.8**	**13.2**
Contr. 1 (char)	31.8	31.3	16.3	14.4
Contr. 2 (sub-17k)	31.3	30.5	15.9	14.0
Contr. 3 (word)	31.3	31.2	16.3	14.3

TOP-3

	MGB-3		MGB-2
	MR-WER	AV-WER	WER
Aalto	**29.3**	**37.5**	**13.2**
NDSC-THUEE	32.5	40.7	14.5
JHU	32.8	40.7	16.0

Why did we succeed

- Effective acoustic model adaptation.
- Use of multiple different lexical units (words, subwords, characters).
- Wide variety of tools (Kaldi, Morfessor, TheanoLM, VariKN).

Acoustic modeling

Organizer's baseline -

- The provided baseline was a model trained on the MGB-2 data. The acoustic model was a Time-Delay Neural Network (**TDNN**) trained on all MSA data, without any filtering.
- The language model was a **3-gram** trained on the provided background text corpus.
- For acoustic modeling and decoding **Kaldi** was used, for language modeling **SRILM**.

Aalto baseline relative improvement +11%

Our baseline was the same as the organizer's, except for

- We used only data with high confidence for GMM training.
- All provided training data was **automatically cleaned and segmented** using standard Kaldi scripts. This resulted in **1022 hours** of cleaned data.
- The language model was a small **varigram** model trained with the **VariKN** toolkit.

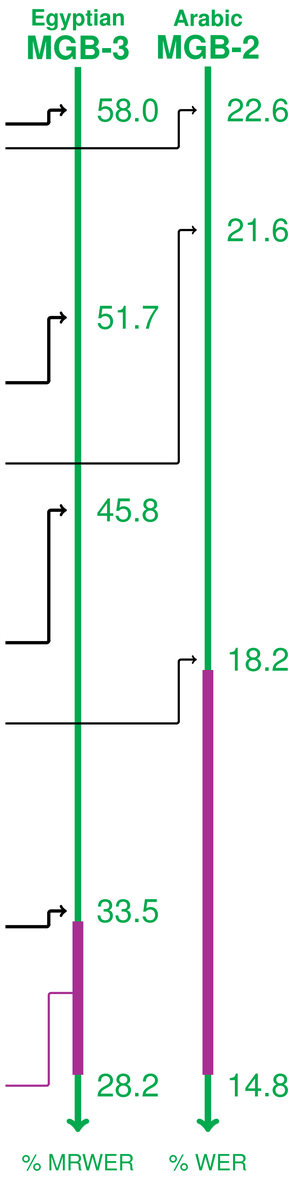
RNN acoustic model +11.4%

- We experimented with different recurrent and non-recurrent architectures.
- Bidirectional long short-term models combined with regular TDNN layers (**TDNN-BLSTM**) gave the best results. This model had 3 TDNN layers (1024 units) and 3 BLSTM layers in both directions (256 units).
- We used WER on the MGB-2 development set as the stopping criterion.

Dialect adapted Acoustic model +26.7%

- We used our best TDNN-BLSTM model and **adapted** with a simple procedure.
- The Egyptian data was used to continue training the model for a small number of epochs. Same parameters were used as for the last regular training iteration.
- We used MRWER on the MGB-3 development set as the stopping criterion.

Scale of language modeling improvements in the next column.



Language / lexical modeling

Interpolation and Rescoring +2.8%

- Besides a small language model on the background text data, we also trained bigger models on the background data, as well as separate models on the transcriptions of the MGB-2 and MGB-3 training sets.
- We **interpolated** these n-gram models, with the interpolation weight optimized for the respective development sets.
- We **rescored** the first-pass results with the larger interpolated language model.

Subwords -0.1%

- Instead of words as units, we also tried **subwords** (trained with **Morfessor**) and **characters** as language modeling units.
- Subwords reduce the vocabulary size while increasing the coverage of the language, as unseen words can be constructed.
- For MGB-3 this did not give an immediate improvement, but in combination with later steps, it did improve over word models.

Recurrent Neural Network Language Model +4.5%

- We build Recurrent Neural Network Language Models (**RNNLM**) with the **TheanoLM** toolkit.
- Our models were **LSTM** models with a projection layer (300/200 units), a hidden layer (1500/1000 units) and a highway layer (1500/1000 units) with a *tanh* activation function (units for subwords/words). The models were trained with Adagrad.
- We rescored the lattices and interpolated them with the n-gram lattices.

System combination +9.9%

- Different systems were built, including TDNN-LSTM and TDNN-BLSTM acoustic models, and models with various lexical units and boundary marking methods.
- Systems were combined and decoded using Minimum Bayesian Risk (**MBR-decoding**).
- In total 40 systems were used.