Unsupervised Language Model Adaptation for Lecture Speech Transcription

Thomas Niesler[†] and Daniel Willett[‡]

[†]Department of Electronic Engineering, University of Stellenbosch, South Africa

[‡]Speech Open Lab, NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

1. Introduction

- Aim: specialise well−trained baseline model to particular character of current recognition task
- Supervised adaptation performs well but requires reference sample from target domain
- Unsupervised adaptation uses transcript generated by recogniser for adaptation
  - No reference material required
  - Errors in transcript: deteriorated performance
- Issues: data sparseness and error reinforcement

2. The task

- Japanese speech & language conferenceslectures
  - ⇨ Domain is highly constrained

Data set	#Lectures	Length	#Words
Dev	7	2.0h	23k
Eval	7	3.2h	35k
Train	144	38h	413k

- Only source of language model training data
- Baseline trigram language model
- Tree−based state−clustered SI cross−word triphones
- Closed 13K vocabulary

3. Adaptation process

Text selection

- Use 1st−pass transcription to identify subset of training corpus most relevant to current lecture
- Measure similarity using **term frequency inverse document frequency (tf−idf) :**

$$\mathcal{T}(d_i, w_j) = tf(d_i, w_j) \cdot \log(idf(w_j))$$

where

$$tf(d_i, w_j) = \#\text{occurrances of } w_j \text{ in } d_i$$

and

$$idf(w_j) = \frac{\text{total \#documents}}{\#\text{documents containing } w_j}$$

- Similarity between two documents:

$$\mathcal{S}(d_i, d_k) = \frac{\sum_{j=1}^{V} (\mathcal{T}(d_i, w_j) \cdot \mathcal{T}(d_k, w_j))}{\sqrt{\left(\sum_{j=1}^{V} \mathcal{T}(d_i, w_j)^2\right) \cdot \left(\sum_{j=1}^{V} \mathcal{T}(d_k, w_j)^2\right)}}$$

- The tf−idf is large when w_j occurs often in d_i but not in many other documents
- Linear interpolation with baseline LM

Minimum Discriminant Information

- Estimate unigram distribution $P_a(w)$ from adaptation data
- MDE finds adapted model $P_a(w|h)$ that is as close as possible (KL distance) to baseline model $P_b(w|h)$ while keeping marginal $P_a(w)$

$$\sum_h P_a(w|h) \cdot P_a(h) = P_a(w) \quad \forall w$$

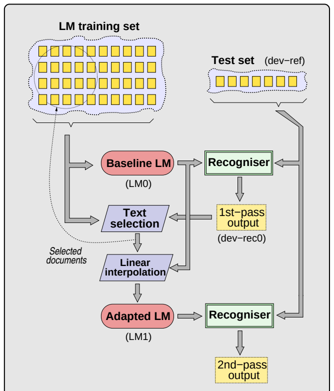
- Approximate solution (Kneser et al, Eurospeech 97)

$$P_a(w|h) = \frac{\alpha(w) \cdot P_b(w|h)}{\sum_w \alpha(w) \cdot P_b(w|h)}$$

$$\alpha(w) = \sqrt{\frac{P_a(w)}{P_b(w)}}$$

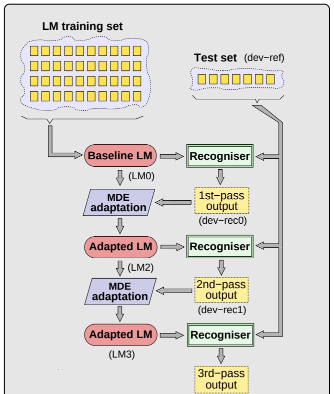
4. Experiments

Text−selection only



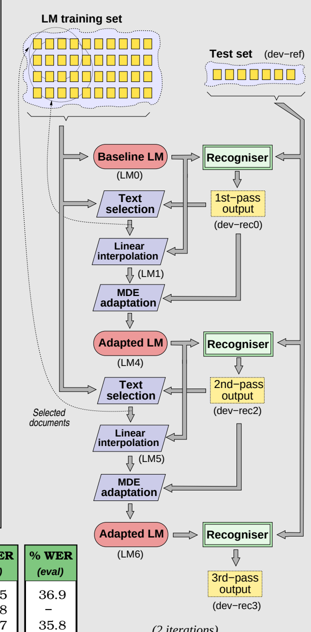
LM	Perplexity		% WER
	dev−ref	dev−rec0	
LM0	130.1	107.2	33.5
LM1	121.8	100.7	32.8

MDE only



LM	Perplexity			% WER
	dev−ref	dev−rec0	dev−rec1	
LM0	130.1	107.2	−	33.5
LM2	89.1	67.0	65.4	31.8
LM3	88.0	67.3	63.3	31.7

Combined text−selection and MDE



LM	Perplexity				% WER (dev)	% WER (eval)
	dev−ref	dev−rec0	dev−rec2	dev−rec3		
LM0	130.1	107.2	−	−	33.5	36.9
LM1	121.8	100.7	−	−	32.8	−
LM4	86.3	64.8	64.5	−	31.7	35.8
LM5	86.3	−	64.5	−	−	−
LM6	79.2	−	54.3	55.2	31.2	35.7
LM7	79.0	−		55.2	−	−
LM8	78.2	−		51.3	31.2	35.7

5. Summary & conclusion

- Text−selection and MDE both improve perplexity and word error−rate for a Japanese lecture speech transcription task
- In conjunction with each other: further gains
- Training corpus small & task highly constrained
- Promising for larger & more diverse language model corpora

Presented at the International Conference on Spoken Language Processing (ICSLP) , Denver, September 2002.