Adaptation of Morpheme-based Speech Recognition for Foreign Entity Names

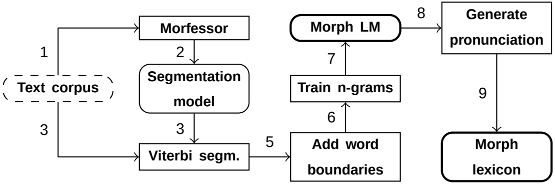
André Mansikkaniemi and Mikko Kurimo

Aalto University School of Science
Department of Information and Computer Science

October 5, 2012

Introduction

- **Statistical morph-based language models**



- **Morph n-gram examples:**
  - <w> expect ing <w> un expect ed ness <w>
  - <w> oli <w> oikea staan <w> yllättävä n <w> hyvä <w>

Introduction

- **Morph-based language models for ASR**
  - Statistical morph segmentation successfully used to tackle OOV problem in speech recognition for morphologically rich languages (Finnish, Turkish, Estonian) [1]
  - High recognition error rate still remains for foreign entity names (FENs) [2]

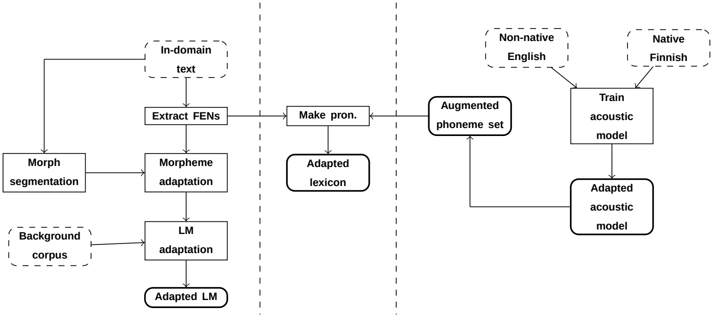
[1] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, and A. Stolcke, Morph-based speech recognition and modeling of out-of-vocabulary words across languages, ACM Trans. Speech Lang. Process., vol. 5, no. 1, pp. 1-29, 2007.
[2] T. Hirsimäki and M. Kurimo, Analysing Recognition Errors in Unlimited-Vocabulary Speech Recognition, Proc. NAACL-2009, pp. 193-196, 2009.

Introduction

- **Causes of high FEN error rate in morph-based ASR**
  - Erroneous pronunciation models
  - Out-of-domain or out-of-date background LM
  - Oversegmentation of foreign words (specific for statistical morph-based models)
    - Examples: mcafee $\rightarrow$ m + cafe + e, reading $\rightarrow$ re + a + ding
    - Makes pronunciation modeling difficult and unreliable

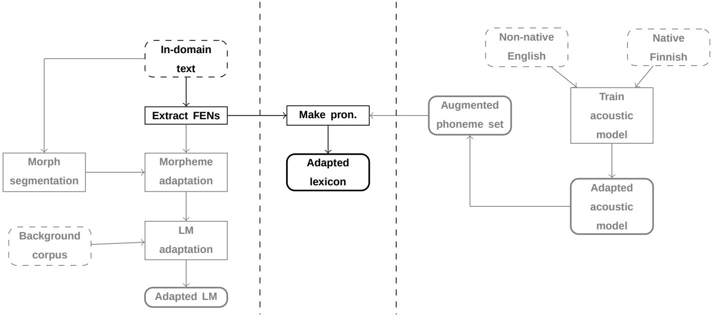
Introduction

- **Adaptation environment for improving FEN recognition**



Methods

- **Lexicon adaptation**
  - Extract foreign words from in-domain text
  - Generate pronunciation rule
  - Add to lexicon

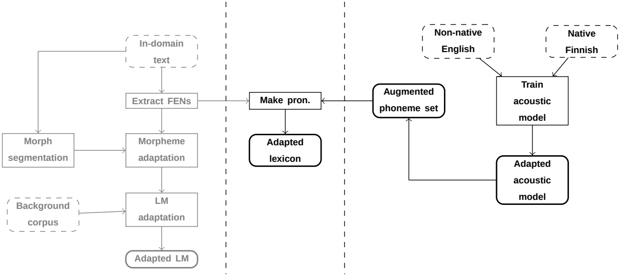


Methods

- **Acoustic model adaptation**
  - Train acoustic model with English sentences spoken by Finnish speakers
  - Augment native phoneme set with most common non-native phonemes

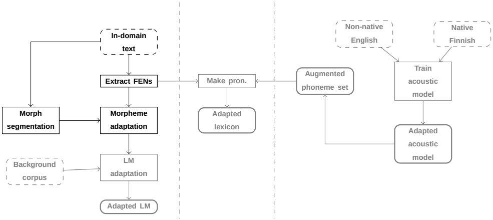
Phoneme	**Word**
CH	**ch**eese
JH	**g**eorge
SH	**sh**e
TH	**th**eta

  - Use augmented phoneme set to generate pronunciation rules for foreign words



Methods

- **Morpheme adaptation**
  - Oversegmented foreign words in in-domain text restored back in to their base forms
  - ***sta dium*** → ***stadium***
    com mon we al th → ***commonwealth***
  - Enables easier pronunciation modeling



In-domain text

Extract FENs

Morph segmentation

Morpheme adaptation

Background corpus

LM adaptation

Adapted LM

Make pron.

Adapted lexicon

Augmented phoneme set

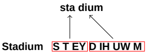
Non-native English

Native Finnish

Train acoustic model

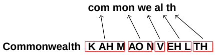
Adapted acoustic model

- **Alternative is morph-aligned pronunciation (morph pron.)**
  - Align pronunciation rule of a whole word on to separate morphs using maximum-likelihood alignment



sta dium

Stadium S T EY D IH UW M



com mon we al th

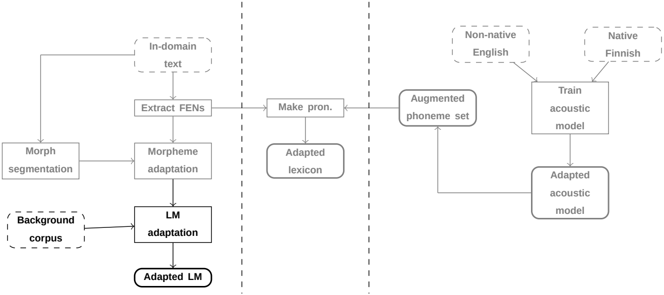
Commonwealth K AH M AO N V EH L TH

Methods

- **Language model adaptation**
  - Linear interpolation used to the adapt background LM P$_B$ (w|h) with in-domain LM P$_i$ (w|h)

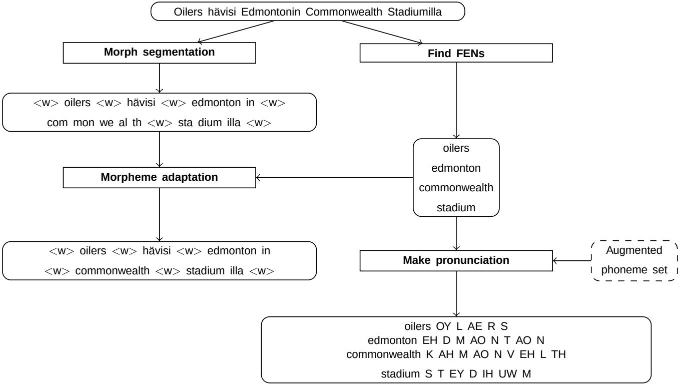
$$P_{adap_i}(w|h) = \lambda P_i(w|h) + (1-\lambda)P_B(w|h) \qquad (1)$$

  - Value of adaptation weight λ determinded beforehand



Methods

- **Example**


Oilers hävisi Edmontonin Commonwealth Stadiumilla
Morph segmentation
Find FENs
<w> oilers <w> hävisi <w> edmonton in <w>
com mon we al th <w> sta dium illa <w>
Morpheme adaptation
oilers
edmonton
commonwealth
stadium
<w> oilers <w> hävisi <w> edmonton in
<w> commonwealth <w> stadium illa <w>
Make pronunciation
Augmented
phoneme set
oilers OY L AE R S
edmonton EH D M AO N T AO N
commonwealth K AH M AO N V EH L TH
stadium S T EY D IH UW M

Experiments

- **System&Models**
  - Aalto speech recognizer [5]
  - Background text corpus of 70 million words
    - Morph segmentation model
    - Background LM (n=12, 30k morph vocabulary) trained on segmented corpus with variKN toolkit [6]
  - Audio corpus with 20h of speech (Finnish)
    - Baseline acoustic model
- **Evaluation data**
  - Finnish radio news segments in 16kHz audio
    - General news set: 32 segments, 8271 words, 4.8% FENs
    - Sports news set: 43 segments, 6466 words, 7.9% FENs
  - Spoken document retrieval set
    - General news: 1609 sentences, 4.0% FENs
    - 171 queries

[5] T. Hirsimäki, J. Pylkkönen, and M. Kurimo, Importance of High-order N-gram Models in Morph-based Speech Recognition, IEEE Trans. Audio, Speech and Lang., pp. 724-732, vol. 17, 2009.

[6] V. Siivola, T. Hirsimäki and S. Virpioja, On Growing and Pruning Kneser-Ney Smoothed N-Gram Models, IEEE Trans. Audio, Speech and Lang., Vol. 15, No. 5, 2007.

Experiments

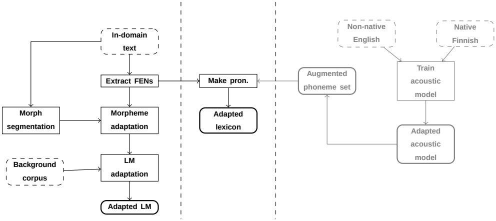
- **LM adaptation data**
  - Collected manually from the Web
  - On average 2-3 articles per topic featured in the news segments
    - 120 000 words of text gathered for general news set
    - 60 000 words of text gathered for sports news set
    - 60 000 words of text gathered for spoken document retrieval set
- **AM adaptation data**
  - English sentences spoken by native Finnish speakers, 70 minutes of 16 kHz audio

Results - Speech recognition task

- **Lexicon, LM, and morpheme adaptation**

Baseline acoustic model					
Adaptation method		**Results**			
		General News		Sports News	
Primary	Additional	WER[%]	FENER[%]	WER[%]	FENER[%]
-		21.7	76.8	34.1	80.9
Lexicon		21.7	76.6	34.0	80.7
LIN (λ = 0.1)	-	20.5	68.0	32.1	70.0
	Lexicon	20.4	67.8	32.1	70.4
	Morpheme + Lexicon	**19.9**	**55.7**	**30.1**	**52.9**
	Lexicon (morph pron.)	20.7	57.9	31.6	54.2

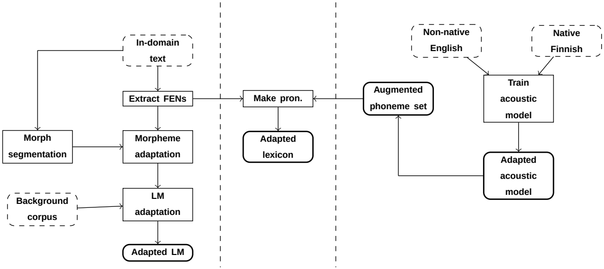
WER = Word error rate
FENER = Foreign entity name error rate



Results - Speech recognition task

- **AM adaptation with augmented phoneme set**

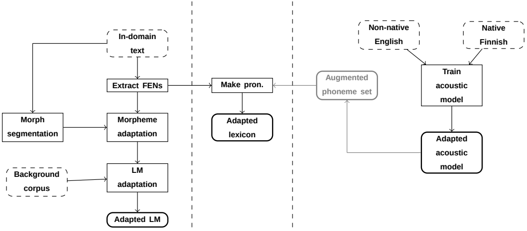
Adapted acoustic model with augmented phoneme set (CH,JH,SH,TH)					
Adaptation method		**Results**			
		General News		Sports News	
Primary	Additional	WER[%]	FENER[%]	WER[%]	FENER[%]
-		23.0	77.8	34.7	81.5
LIN (λ = 0.1)	Lexicon	22.0	64.5	31.3	61.7
	Morpheme + Lexicon	**21.6**	**56.9**	**30.6**	**53.8**



Results - Speech recognition task

- **AM adaptation with native phoneme set**
- **Non-native phonemes mapped to closest native diphone or triphone context**
  - "S" → "sh", "C" → "tsh", "D" → "dj", "T" → "th"

Adapted acoustic model with native phoneme set					
Adaptation method		**Results**			
		General News		Sports News	
Primary	Additional	WER[%]	FENER[%]	WER[%]	FENER[%]
-		21.6	77.3	33.5	80.5
LIN (λ = 0.1)	Lexicon	20.0	59.4	29.9	60.7
	Morpheme + Lexicon	**19.5**	**51.6**	**29.1**	**52.1**



Results - Speech retrieval task

- **ASR results**

Baseline acoustic model			
Adaptation method		**Results**	
		SDR eval. set	
Primary	Additional	WER[%]	FENER[%]
-		29.9	64.4
LIN (λ = 0.1)	Lexicon	**29.2**	**50.4**
	Morpheme + Lexicon	29.3	51.7

- **Ranked Utterance Retrieval results**
  - Mean Average Precision (MAP)

System	**Indexing**		
	Baseform	Morph	Combined
Baseline	0.4643	0.6296	0.6861
LIN + Lexicon	0.4651	0.6317	**0.6915**

Conclusions

- **Adaptation framework improves recognition of foreign words**
- **Positive effect on FEN recognition**
  - LM adaptation
  - Lexicon + Morpheme adaptation
  - Morph-aligned pronunciation
  - AM adaptation (native phoneme set)
- **Future work**
  - Fully unsupervised adaptation framework (partially implemented [7])
  - Adaptation of acronyms

[7] André Mansikkaniemi and Mikko Kurimo. Unsupervised vocabulary adaptation for morph-based language models. In Proceedings of the NAACL 2012 Workshop on the Future of Language Modeling for HLT. ACL, June 2012.