Carnegie Mellon.
SCHOOL OF COMPUTER SCIENCE

Learning Out-of-Vocabulary Words in Automatic Speech Recognition

Long Qin

Committee:

Alexander I Rudnicky, CMU (Chair)

Alan W Black, CMU

Florian Metze, CMU

Mark Dredze, JHU

6/6/2013

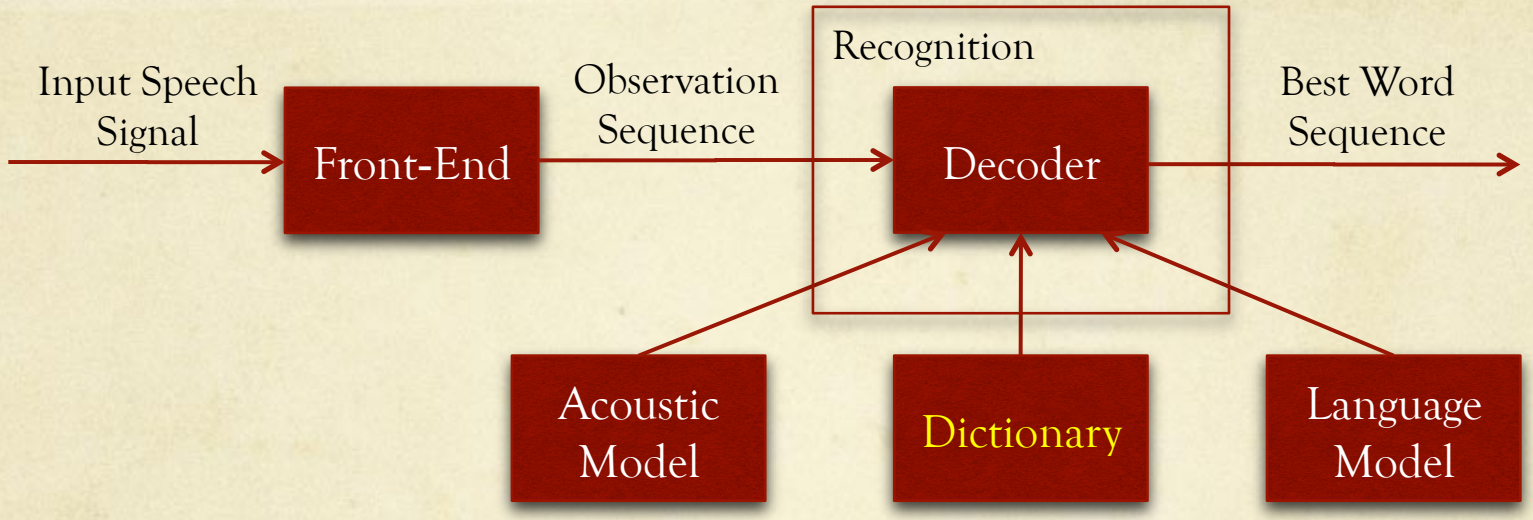
Outline

1. The Out-of-Vocabulary (OOV) word problem
2. The OOV word learning framework
   a) System overview
   b) OOV word detection
   c) OOV word clustering
   d) OOV word recovery
3. Conclusion and future work

Outline

1. The Out-of-Vocabulary (OOV) word problem
2. The OOV word learning framework
   a) System overview
   b) OOV word detection
   c) OOV word clustering
   d) OOV word recovery
3. Conclusion and future work

Automatic speech recognition (ASR)



IV (in-vocabulary) words

a	AH
a(2)	EY
abandon	AH B AE N D AH N
⋮	⋮
zurich	Z UH R IH K

OOV (out-of-vocabulary) words:
words not in this list

The OOV word problem

REF:	associated	inns	known	as	***AIRCOA***	
HYP:	associated	inns	and	is	a	tele

- ASR systems mis-recognize OOV word as IV word(s)
- OOV words degrade the recognition accuracy of surrounding IV words
- OOV words are content words, such as names or locations, which are crucial to the success of many speech recognition applications
- ASR systems which can detect and recover OOV words are of great interest

Related work

- OOV word detection
  - Find the mismatch between the phone and word recognition result
  - Consider OOV word detection as a binary classification task
  - Apply a hybrid lexicon and language model during decoding
- OOV word recovery
  - Apply phoneme-to-grapheme conversion or finite state transducer
  - Use an information retrieval and key word spotting system
  - Estimate rough language model scores from semantic similar IV words
- Convert OOV word into IV word
  - Recognize the same OOV word as IV word when it appears in the future
  - An ASR system can learn new words and operate on an open vocabulary

[Hayamizu et al., 1993; Klakow et al., 1999; Sun et al., 2003; Bisani & Ney, 2005; Hannemann et al., 2010; Parada et al., 2010; Lecorve et al., 2011]

Thesis Statement

OOV words can be automatically detected, clustered and recovered in an integrated learning framework. Given the ability to add new words, a speech recognition system can operate with an open vocabulary.

Outline

1. The OOV word problem
2. The OOV word learning framework
   a) System Overview
   b) OOV word detection
   c) OOV word clustering
   d) OOV word recovery
3. Conclusion and future work

OOV word detection

- OOV word detection is to find the appearance of OOV word in an utterance

…	associated	inns	known	as	*AIRCOA*	…

OOV word detection result

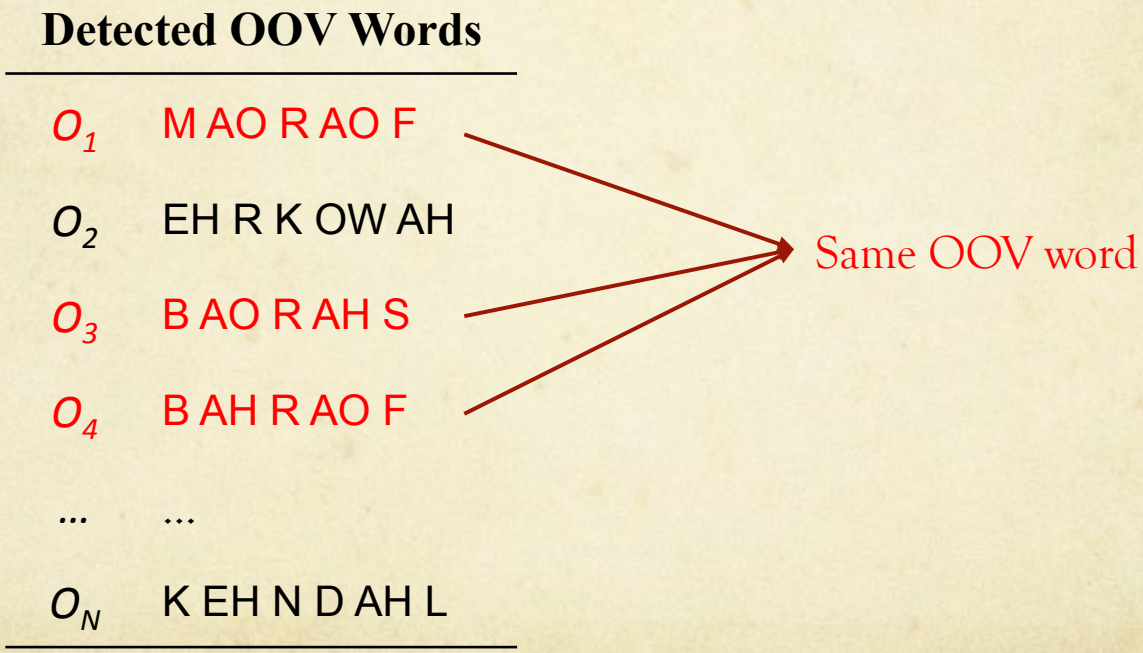
…	associated	inns	known	as	EH R K OW AH	…

From the detection results of all testing speech

Detected OOV Words	
O_1	M AO R AO F
O_2	EH R K OW AH
…	…
O_N	K EH N D AH L

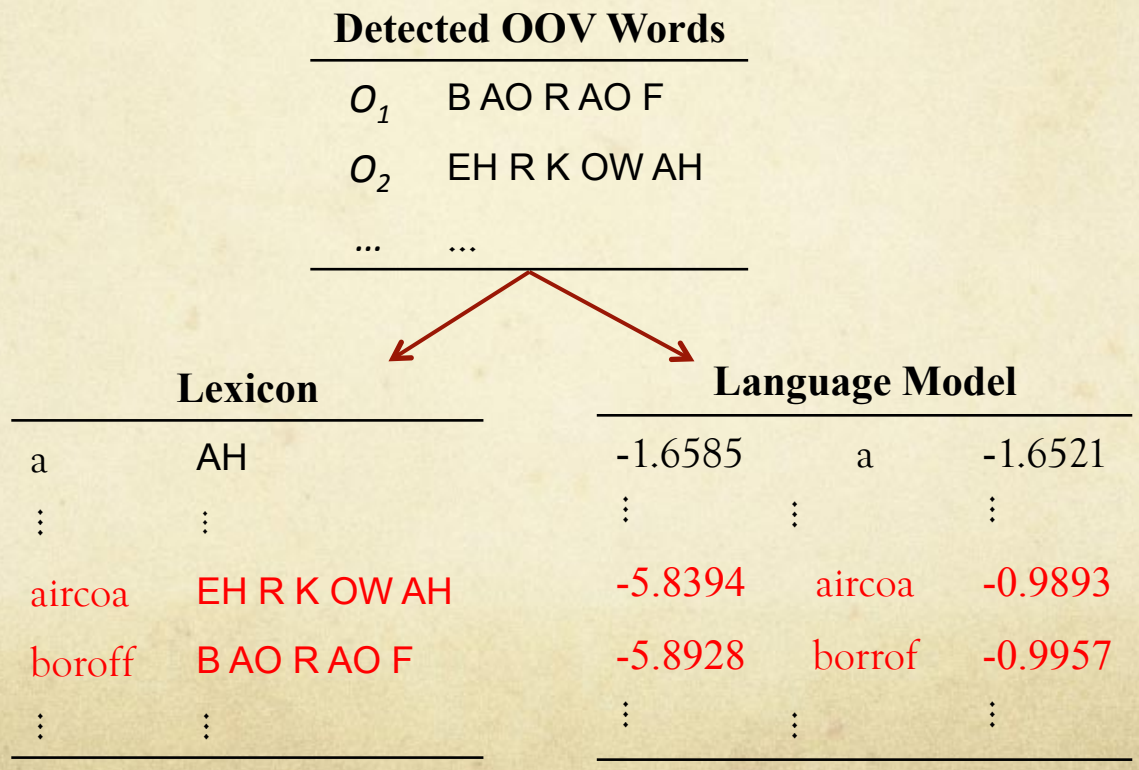
OOV word clustering

- OOV word clustering is to find the multiple instances of an OOV word from the OOV word detection result



OOV word recovery

- OOV word recovery is to recover the written form and language model (LM) scores of detected OOV words, then integrate them into the lexicon and LM



The OOV word learning framework

OOV Word Detection → OOV Word Clustering → OOV Word Recovery

Detected OOV Words

O_1	M AO R AO F
O_2	EH R K OW AH
O_3	B AO R AH S
O_4	B AH R AO F
...	...
O_N	K EH N D AH L

Detected OOV Words

O_1	B AO R AO F
O_2	EH R K OW AH
...	...
O_M	K EH N D AH L

Lexicon

a	AH
⋮	⋮
aircoa	EH R K OW AH
boroff	B AO R AO F
⋮	⋮

Language Model

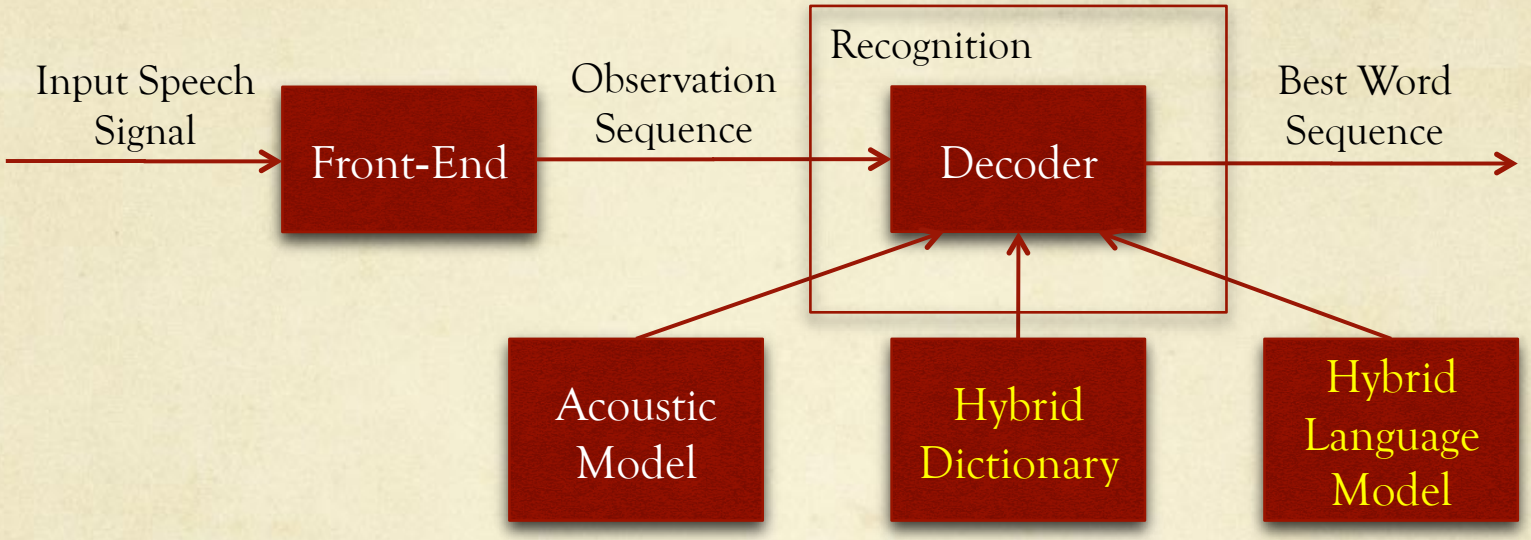
-1.6585	a	-1.6521
⋮	⋮	⋮
-5.8394	aircoa	-0.9893
-5.8928	borrof	-0.9957
⋮	⋮	⋮

OOV word detection

OOV Word Detection

Detected OOV Words	
O_1	M AO R AO F
O_2	EH R K OW AH
O_3	B AO R AH S
O_4	B AH R AO F
...	...
O_N	K EH N D AH L

The hybrid system



Mixture of words and sub-lexical units

a	AH
⋮	⋮
EH_L	EH L
⋮	⋮

⋮	⋮	⋮
-1.7965	thanks	for
-3.8746	thanks	EH_L
⋮	⋮	⋮

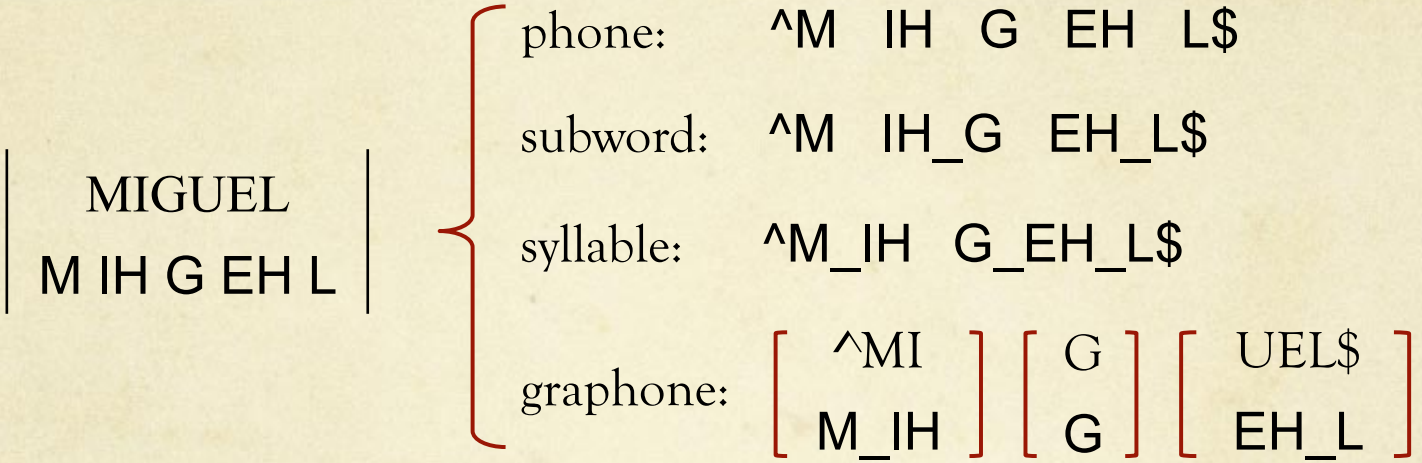
detection result

detected OOV word

associated	inns	known	as	EH R K OW AH

Sub-lexical units

- Build hybrid systems using different types of sub-lexical units



	Pros	Cons
Subword	simple and robust	lack linguistic restrictions
Syllable	maintain phonetic restrictions	produce long rare units
Graphone	model both letters and phones	large number of units

[Qin et al., 2011]

Combining multiple systems' outputs

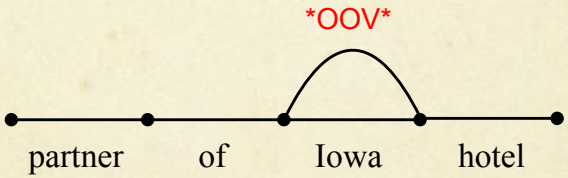
Syllable Hybrid System:	partner	of	EH R K OW AH	hotel
Subword Hybrid System:	partner	of	EH R OW AH	hotel
Graphone Hybrid System:	partner	of	Iowa	hotel

*Convert OOV tokens to *OOV**

Syllable Hybrid System:	partner	of	*OOV*	hotel
Subword Hybrid System:	partner	of	*OOV*	hotel
Graphone Hybrid System:	partner	of	Iowa	hotel

Combination

Word Transition Network



Rescoring

Best Result:	partner	of	*OOV*	hotel

[Qin et al., 2012]

Combining multiple types of sub-lexical units

- Utilize multiple types of units in one system, so that different units can complement each other

… PRESIDENT MIGUEL DE LA MADRID'S SERIOUSNESS …
… BOARD OF SAN MIGUEL CORPORATION …

For each appearance of MIGUEL, we stochastically select one type of units

… PRESIDENT ^M_IH G_EH_L$ DE LA MADRID'S SERIOUSNESS …
… BOARD OF SAN ^MI:M_IH G:G UEL$:EH_L CORPORATION …

One OOV word can be modeled by multiple types of sub-lexical units!

[Qin & Rudnicky, 2012]

Datasets

	WSJ	BN	SWB
Vocabulary Size	20k	20k	10k
Dev OOV	319 (2.1%)	204 (2.0%)	204 (1.7%)
Eval OOV	200 (2.2%)	255 (2.0%)	209 (1.7%)
Test OOV	260 (2.1%) 136	381 (2.8%) 91	383 (1.8%) 253

- Dev data is used to tune parameters in training
- Eval data is used to learn OOV words
- Test data is used to evaluate how many recovered OOV words can be recognized

OOV word detection experiments

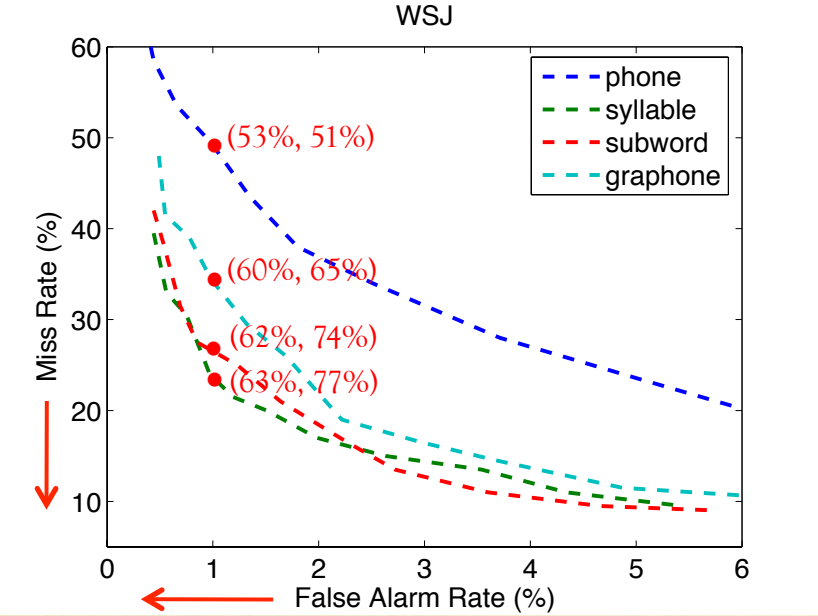
- Evaluation metrics:

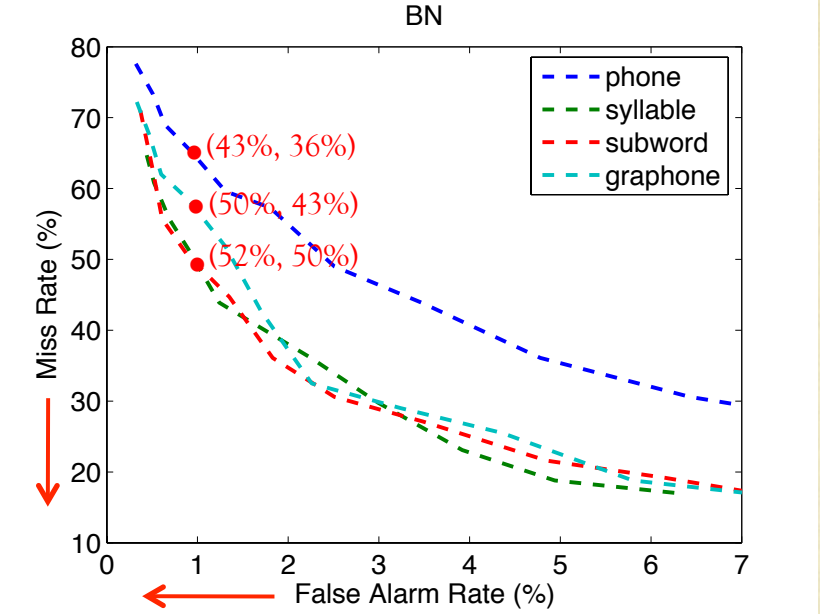
$$\text{Miss Rate (MR)} = \frac{\#\text{OOVs in reference} - \#\text{IVs detected}}{\#\text{OOVs in reference}} * 100\%$$

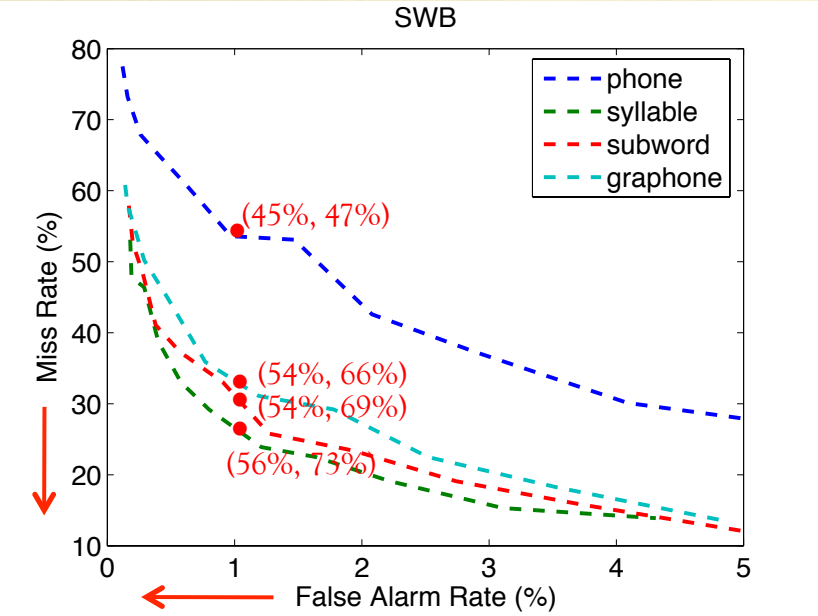
$$\text{False Alarm Rate (FAR)} = \frac{\#\text{OOVs reported} - \#\text{OOVs detected}}{\#\text{IVs in reference}} * 100\%$$

- OOV cost
  - Control how likely to decode OOV word
  - Adjusted from 0 to 2.5 with a step size of 0.25
  - Draw MR-FAR curve to select operation point for specific application

The OOV word detection results



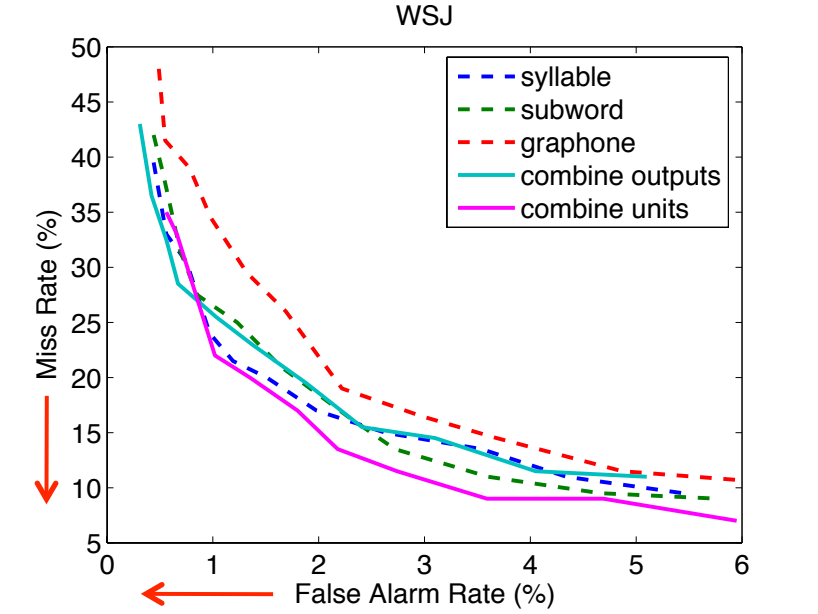


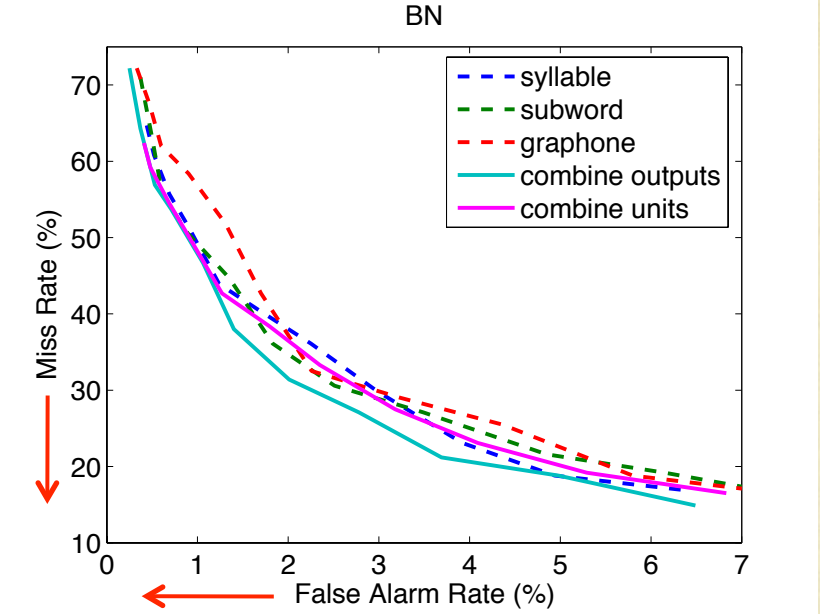


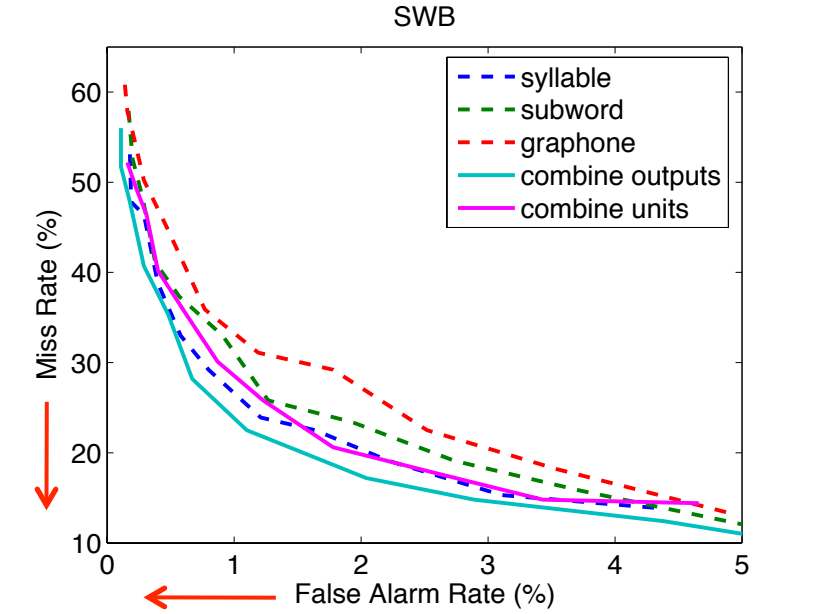
- Complex sub-lexical units perform better than simple phone units
- Perform better in the WSJ and SWB tasks than in the BN task
- On average, detect up to 70% OOV words with up to 60% precision

(Precision, Recall)

The system combination results

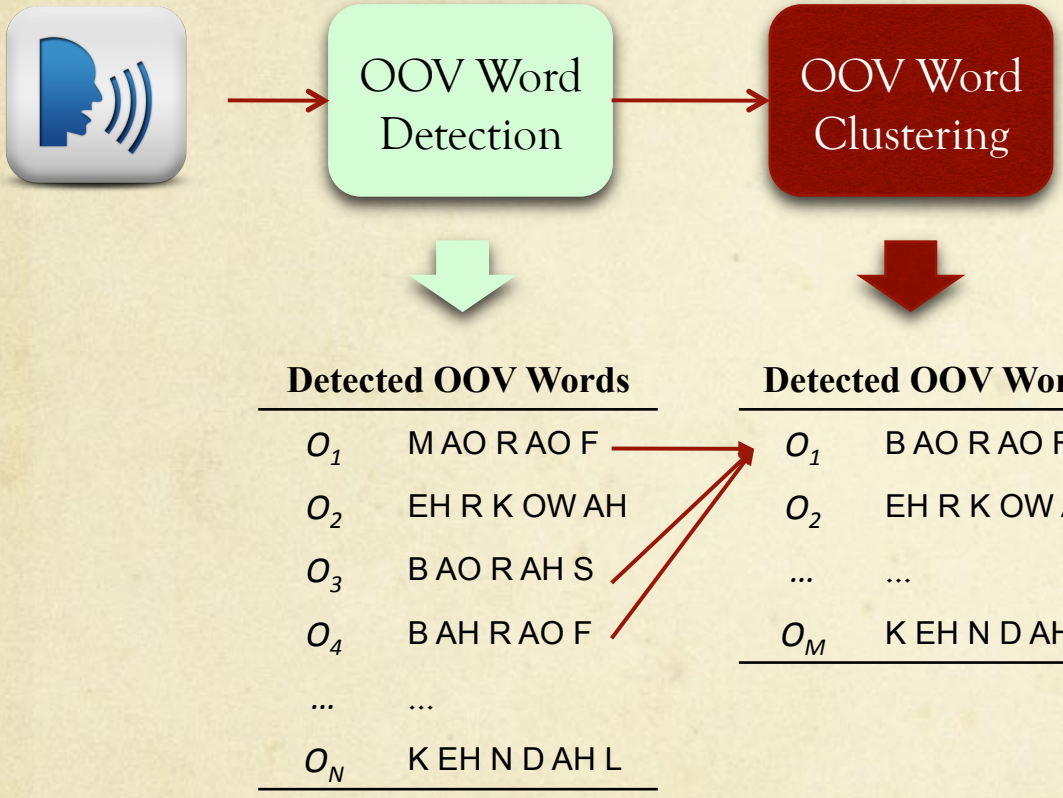






- Two combined systems perform differently across different tasks
- Better combined system performs better than individual systems

OOV word recovery



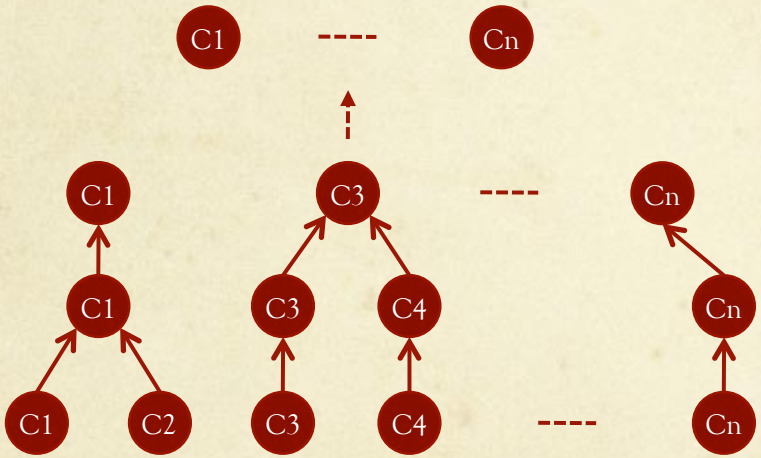
Detect up to 70% OOV words

Recurrent OOV words

- OOV words can appear more than once in a conversation or over a period of time
  - Find multiple instances of an OOV word in the detection result
- Multiple instances of an OOV word are valuable for estimating
  - Pronunciation
  - Part-of-Speech (POS) tag
  - Language model (LM) scores

A bottom-up clustering process

- Finding multiple instances of the same OOV word through bottom-up clustering



$$D(C_i, C_j) = \varpi_1 d_1 + \varpi_2 d_2 + \varpi_3 d_3$$

d_1: phonetic distance

d_2: acoustic distance

d_3: contextual distance

Collecting features from hybrid system output

OOV	Phonetic (Decoded Phones)	Acoustic (Posterior Vectors)	Contextual (Surrounding IV Words)
O_1	S EH L T S	[0.00 ... 0.17]	... from in O_1 major Dietz ...
O_2	K AE D IY	[0.01 ... 0.24]	... people's party O_2 moved into ...
O_3	W AO L IY	[0.02 ... 0.01]	... the rule of O_3 ball ...

- Phonetic distance
  - Modified edit distance
- Acoustic distance
  - Dynamic time warping (DTW) distance
- Contextual distance
  - Local contextual distance works like a LM
  - Global contextual distance resembles a topic model

OOV word clustering experiments

- Hybrid system outputs

	WSJ	BN	SWB
F1 (%)	69	55	71

- Number of recurrent OOV words

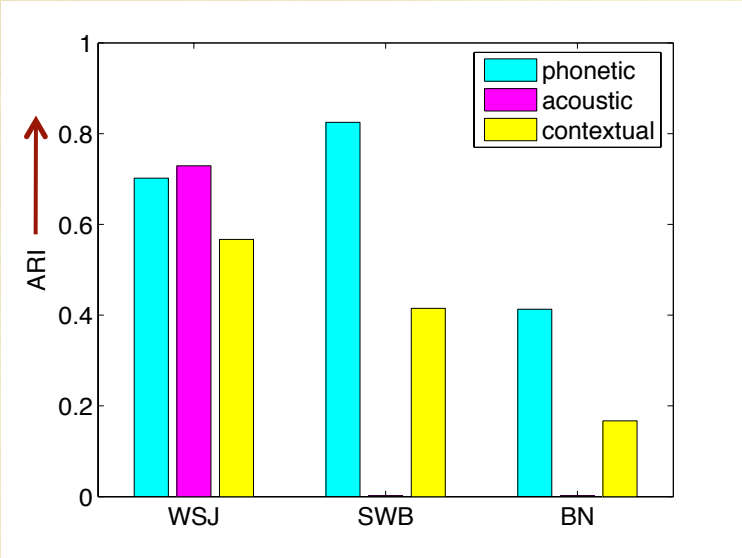
	WSJ	BN	SWB
Count	68 (29%)	109 (31%)	52 (22%)

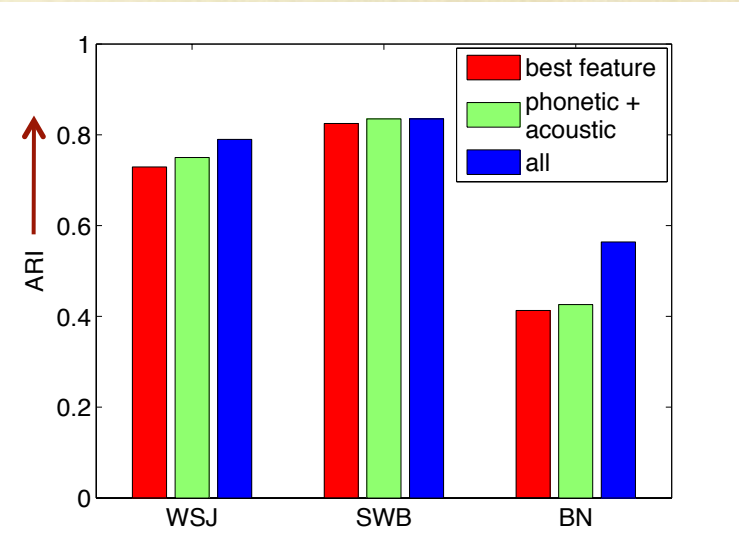
- Evaluation metrics: adjusted Rand index (ARI)

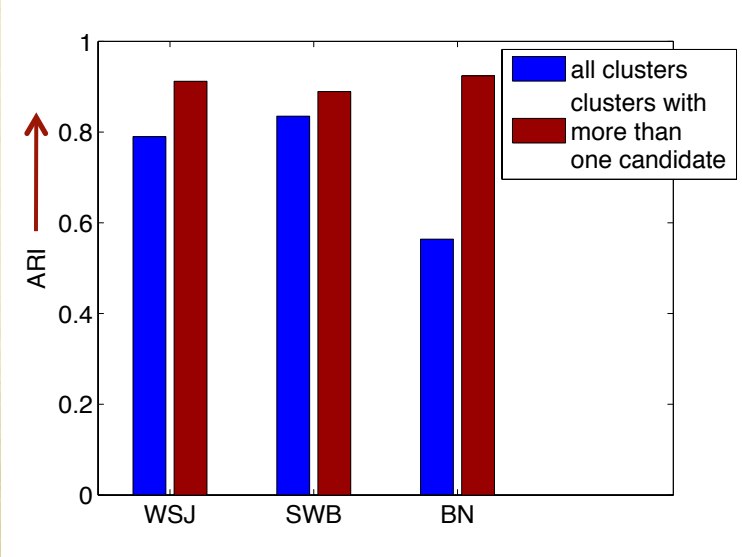
$$RI = \frac{TP + TN}{TP + FP + TN + FN}$$

  - Adjusted for the chance of a clustering
  - Bounded between [-1, 1], 0 for random clustering
  - If without clustering, ARI is close to 0

The OOV word clustering results







- Using one feature
  - Phonetic feature is effective in all tasks
  - Acoustic features only works in WSJ
  - Contextual feature produces positive result
- Using more features is better
- ARI is 0.9 on found recurrent OOV words (comparable to 10% or less errors)

The OOV word learning framework

OOV Word Detection → OOV Word Clustering → OOV Word Recovery

Detected OOV Words

O_1	M AO R AO F
O_2	EH R K OW AH
O_3	B AO R AH S
O_4	B AH R AO F
...	...
O_N	K EH N D AH L

Detect up to 70% OOV words

Detected OOV Words

O_1	B AO R AO F
O_2	EH R K OW AH
...	...
O_M	K EH N D AH L

ARI is up to 0.9

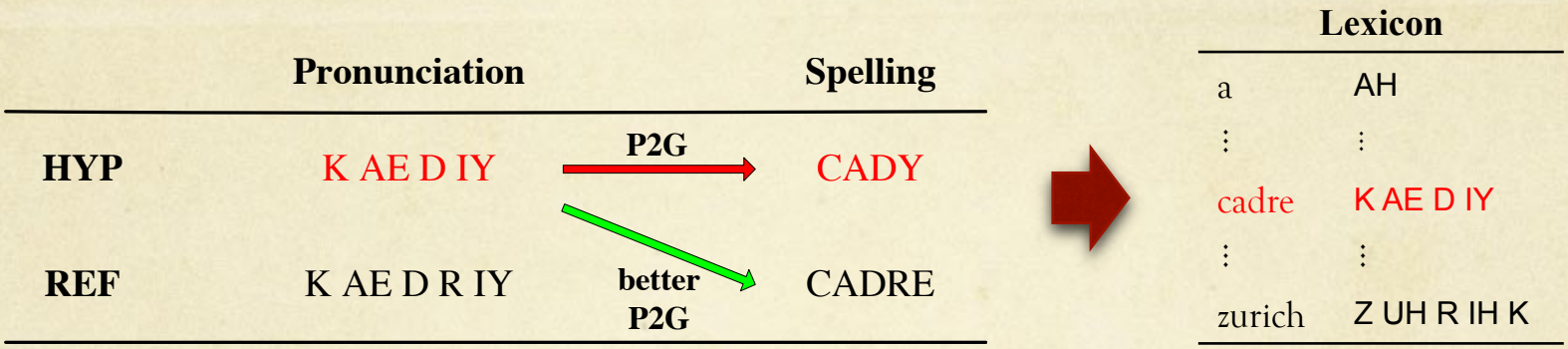
Lexicon

a	AH
⋮	⋮
aircoa	EH R K OW AH
boroff	B AO R AO F
⋮	⋮

Language Model

-1.6585	a	-1.6521
⋮	⋮	⋮
-5.8394	aircoa	-0.9893
-5.8928	borrof	-0.9957
⋮	⋮	⋮

Estimating the written form of an OOV word



- Conventional P2G model is trained from alignments between correct spelling and pronunciation
- Train better P2G model also from alignments between correct spelling and incorrect decoded pronunciation
  - Extract alignments from hybrid decoding result of training speech

Estimating language model scores of an OOV word

- Learn from IV words in the same syntactic category
  - Train a Part-of-speech (POS) class-based LM from training text data (Stanford POS tagger)
  - Estimate LM scores based on the POS label of an OOV word

HYP	partner	of	AIRCOA	hotel	partners
POS	NN	IN	NNP	NN	NNS

$$P(\text{AIRCOA} \mid \text{partner,of}) = P(\text{AIRCOA} \mid \text{NNP})P(\text{NNP} \mid \text{NN,IN})$$

- OOV words may appear in different context in future
  - Estimate possible context an OOV word may appear
  - Substitute surrounding IV words of an OOV word with other semantic similar IV words (WordNet)

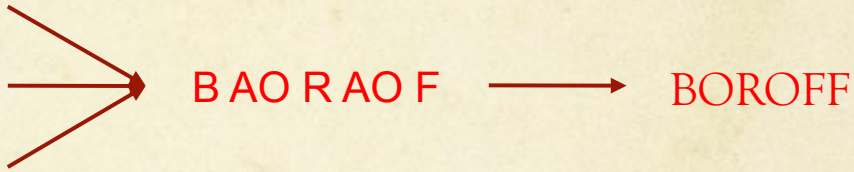
hotel ⟷ inn

New Context	partner	of	AIRCOA	***inn***	partners

Recovering recurrent OOV words

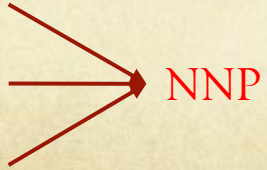
- Estimate better pronunciation

OOV	Pronunciation
O_1	M AO R AO F
O_2	B AO R AH S
O_3	B AH R AO F



- Estimate better language model scores

OOV	POS
O_1	NNP
O_2	NNP+POS
O_3	NNP



Multiple Context
... Philip BOROFF has more from ...
... I am Philip BOROFF </s>
... this is Philip BOROFF from ..

OOV word recovery experiments

- Evaluation metrics

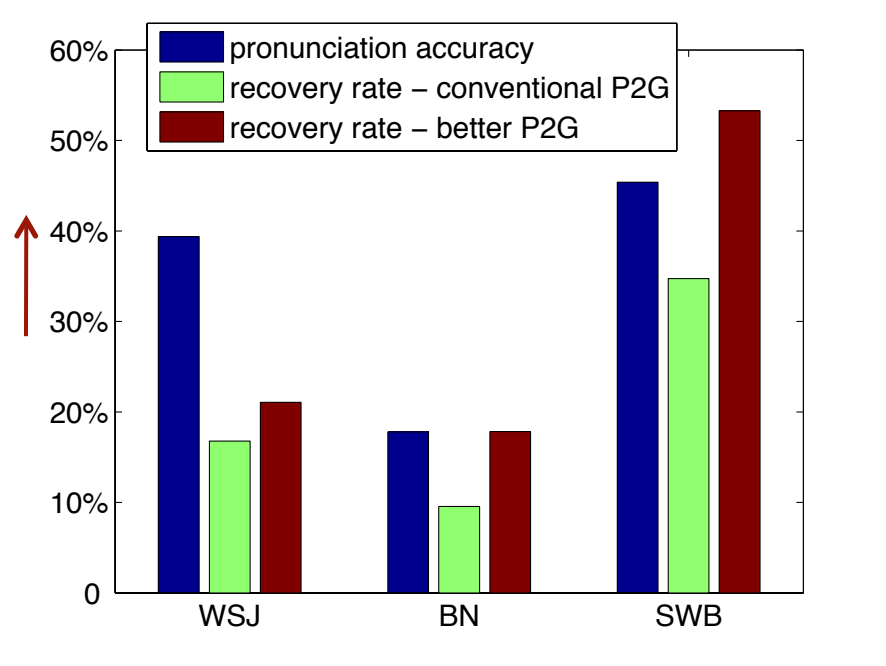
$$\text{Phone Accuracy (PA)} = \frac{\#\text{OOVs with correct pronunciation}}{\#\text{OOVs detected}} * 100\%$$

$$\text{Recovery Rate (RR)} = \frac{\#\text{OOVs recovered}}{\#\text{OOVs detected}} * 100\%$$

$$\text{Word Error Rate (WER)} = \frac{\#\text{substitution errors} + \#\text{deletion errors} + \#\text{insertion errors}}{\#\text{words in reference}} * 100\%$$

- Compare
  - The number of recovered OOV words - detected OOV words with correct written form
  - The number of recovered OOV words recognized in the 2nd pass decoding of Eval data
  - The number of recovered OOV words recognized in the 1st pass decoding of Test data

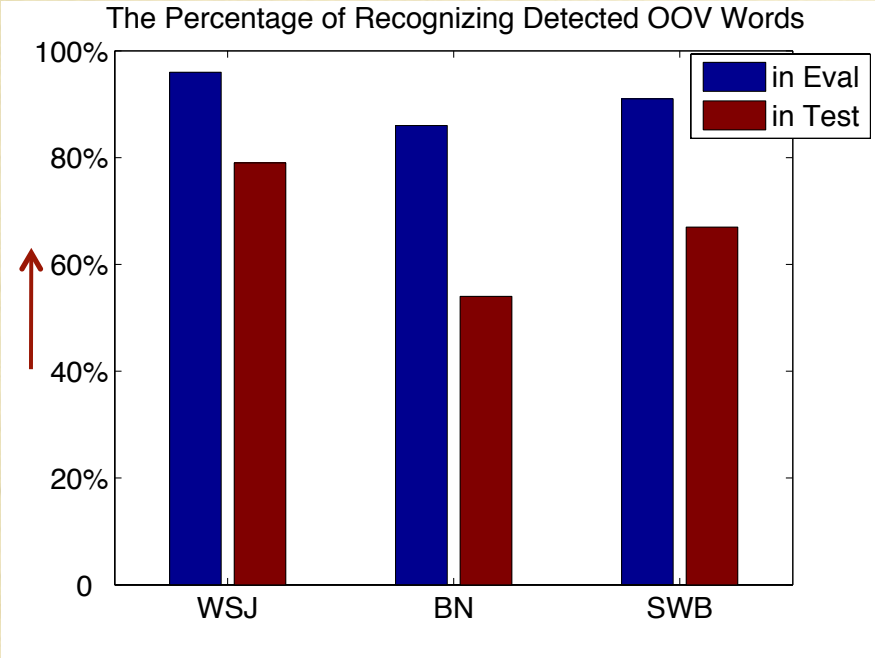
The results of estimating the written form

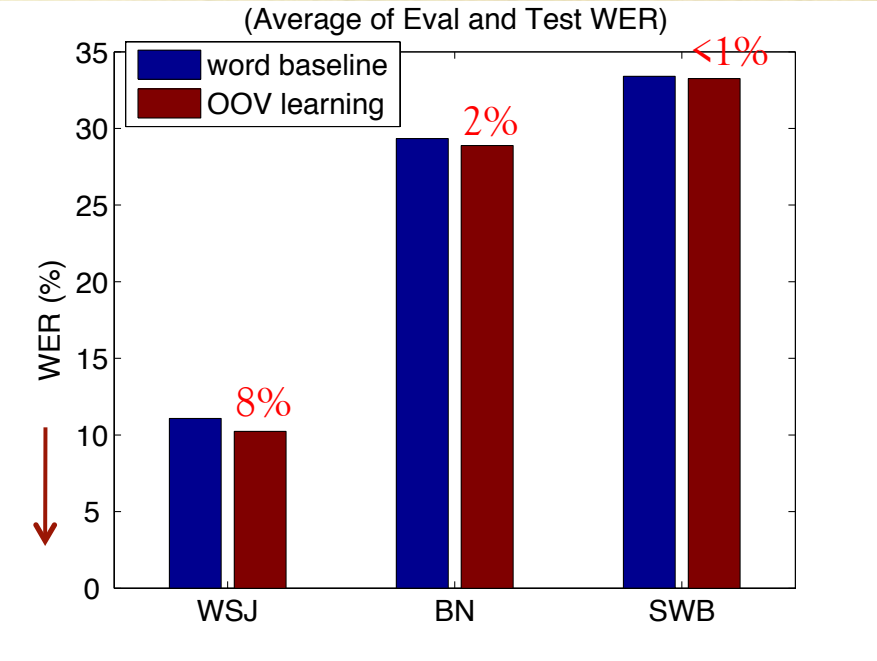


	WSJ	BN	SWB
No. OOV in Eval	200	255	209
No. recovered OOVs in Eval	90 (45%)	73 (29%)	101 (48%)
No. OOV in Test	136	91	253
No. recovered OOVs in Test	61 (45%)	39 (43%)	119 (47%)

- Significantly higher recovery rate when using the better P2G model
- Eliminate 40% OOV words after integrating recovered OOV words into the lexicon

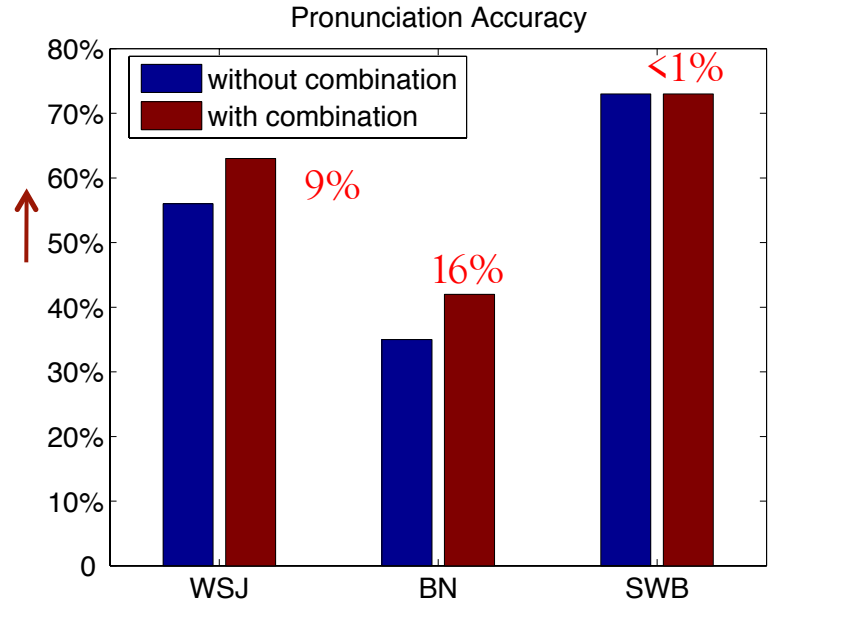
The results of estimating language model scores

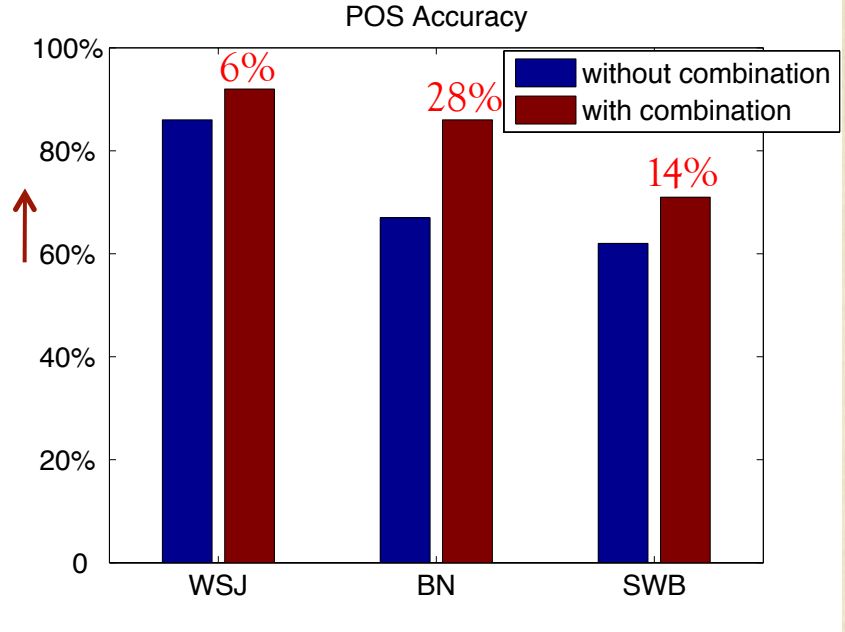




- Recognized more than 90% recovered OOV words in 2nd pass decoding of Eval data and up to 70% in 1st pass decoding of Test data
- Small improvement on WER
  - Higher WER when adding bigrams, trigrams and new context, but recognized more OOV words

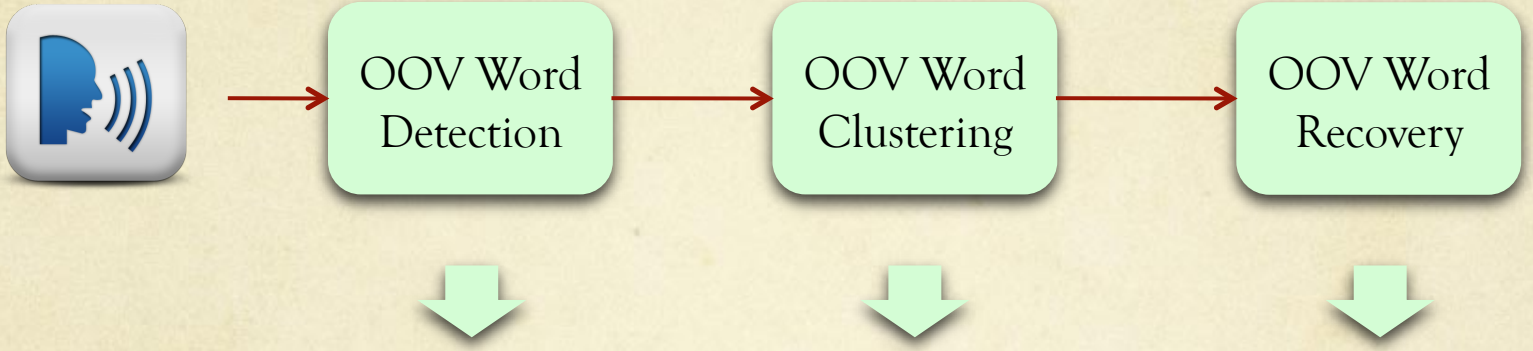
The results of recovering recurrent OOV words





- Only calculated on Recurrent OOV words
- No improvement on recognizing more recovered OOV words or WER
  - Only a few recurrent OOV words in the Eval and Test data

The overall performance



Detected up to 70% OOV words

ARI was up to 0.9

More than 40% original OOV words were integrated into new lexicon

Recognized more than 90% recovered OOVs in Eval and up to 70% recovered OOVs in Test

Overall, learned up to 40% OOV words in Eval and 30% OOV words in Test!

Outline

1. The OOV word problem
2. The OOV word learning framework
   a) OOV word detection
   b) OOV word clustering
   c) OOV word recovery
3. Conclusion and future work

Thesis contributions

- We proposed an OOV word learning framework and showed an ASR system was capable of learning new words.
  - Compared different training schemes and types of sub-lexical units for building the hybrid system. Found that flat hybrid system is better than hierarchical hybrid system, and syllable, subword and graphne units are better than simple phone units. [Qin et al., 2011]
  - Applied system combination and OOV word classifier techniques to improve the OOV word detection performance. Combined systems outperformed individual systems built with one type of units. [Qin et al., 2012; Qin and Rudnicky, 2012]
  - Identified recurrent OOV words through bottom-up clustering. And showed that multiple instances of an OOV word were valuable for improving the OOV word recovery performance. [Qin and Rudnicky, 2013]
  - Trained a better P2G model from the decoding result of training speech. And successfully estimated LM scores for OOV words from syntactic and semantic similar IV words.

Conclusion

- Hybrid system is effective at detecting OOV words. Our system can detect up to 70% OOV words with up to 60% precision.
- Bottom-up clustering finds most recurrent OOV words. Multiple instances of recurrent OOV words improve OOV word learning performance.
- Up to 90% recovered OOV words are recognized after integrating them into the recognizer's lexicon and language model.
- Overall, this OOV word learning framework can successfully learn up to 40% OOV words.

Future work

- OOV word learning curve
  - Human has learning curve to learn new words
  - How to update knowledge about learned new words?
- OOV word learning in a dialog system
  - How to learn OOV words through interactions between user and system?
  - Learn semantic information about a word

Acknowledgments

Q&A

THANKS!