Towards Using Hybrid Word and Fragment Units for Vocabulary Independent LVCSR Systems

Ariya Rastrow , Abhinav Sethy, Bhuvana Ramabhadran and Fred Jelinek

Center for Language and Speech Processing

IBM TJ Watson Research Lab

September 9, 2009

Outline

1. Introduction
   - Why Sub-Word Units?
   - Hybrid Systems
   - Experimental Setup
2. Hybrid Systems for OOV Detection
3. Improving Phone Accuracy and Robustness
4. From Sub-word units to Words
5. Summary

- The simplest answer is : Recognizing OOV terms in ASR
    - All LVCSR based systems have a closed word vocabulary
    - Recognizer replaces OOV terms with the closest match in the vocabulary
    - Neighboring words are also often misrecognized

      → Contributing to recognition errors

    - OOVs degrade the performance for later processing stages (e.g. translation,understanding, document retrieval,term detection)
    - Although OOV rate might be relatively low in state of the art ASR systems, *rare and unexpected events are information rich*
- Eventual goal is to build an open vocabulary speech recognizer

- Fragments are sub-word units (variable length phone sequences) selected automatically using statistical methods(Data-Driven)
    - See slides that follow
- Fragments have the potential to provide a good trade off between coverage and accuracy

- Hybrid System
  - Represents language as a combination of words and fragments
  - Takes advantage of both word and fragment representations yielding improved performance while providing good coverage
- LM is built for such a representation

Hybrid Language Model in detail

- Step 1: Fragment selection based on N-gram pruning
  - Convert LM training text (Exclude OOV) to phones, build N-gram (in our case 5-gram) phone LM and prune it (Entropy-based Pruning).
  - Pruning selects the set of fragments (from single phones to 5-gram phones)

Fragments → IH_N

K_L_AA_R_K

- Step 2: Converting word-based training data into Hybrid word/fragment data
  - $< s >$ THE BODY OF ZIYAD HAMDI WHO HAD BEEN SHOT WAS FOUND SOUTH OF THE CITY $< /s >$
  - need to get pronunciation for OOV terms → grapheme to phone models
    ZIYAD → Z IY AE D
    HAMDI→ HH AE M D IY
  - $< s >$ THE BODY OF Z_IY Y_AE_D HH_AE_M D_IY WHO HAD BEEN SHOT WAS FOUND SOUTH OF THE CITY $< /s >$
  - Fragment representation of OOV is obtained by left-to-right greedy search

Hybrid Language Model in detail

- Step 3: Build LM based on the Hybrid word/fragment set
    - Treat fragments as individual terms
    - After this step, Hybrid LM is built and we have a LM including both words and fragments

- The LVCSR system is based on the 2007 IBM Speech transcription system for GALE Distillation Go/No-go Evaluation
  - Acoustic Models are discriminatively trained on speaker adapted PLP features (best broadcast News acoustic models from IBM).
- The acoustic models are common for all systems in our experiments.
- The LM training text (for all systems) consists of 335M words from 8 sources of BN corpora.
  - Both word and hybrid LMs are 4-gram LMs with Kneser-Ney smoothing
- Word lexicons ranging from 10K words to 84K were selected by sorting the words based on the frequency on the acoustic training data (broadcast news Hub4) .

Continued

- The set of fragments (sub-word units) is selected as described (5-gram phone LM) on the LM training text for each vocabulary size.
    - The size of this set was fixed at roughly $20K$ for all systems. Therefore, the hybrid system includes $20K$ fragments, in addition to the words in its lexicon.
- We report the results:
    - RT-04 BN evaluation set ($45K$ words, 4.5 hours) as an in-domain test set
    - MIT lectures data set ($176K$ words, 21 hours, 20 lectures) as an out-of-domain test set

- OOV rates for different lexicon sizes

Lexicon size	10k	20k	30k	40k	60k	84k
RT-04 (%)	5.04	2.48	1.47	1.04	0.68	0.54
Lectures (%)	7.88	5.45	4.51	4.09	3.53	3.45

TABLE: OOV rates for the RT-04 set and the MIT lectures data

Outline

1 Introduction

2 Hybrid Systems for OOV Detection
- Fragment Posteriors Using Consensus
- Evaluation
- Results

3 Improving Phone Accuracy and Robustness

4 From Sub-word units to Words

5 Summary

- The idea here is that since we have used fragments in the case of OOV for building our LM, then the appearance of fragments in the ASR output indicates an OOV region
  - The simple case would be to search for the fragments in the decoder 1-best output
  - A better way is to search for the fragments in the lattice
- Fragments allow us both to detect OOVs and to represent them
  - **ASR:** TODAY TWO YOUNG GIANT PANDAS FROM CHINA ARRIVED ON A SPECIALLY R_EH_T R_OW F_IH_T IH_D FEDEX JET
  - **REF:** TODAY TWO YOUNG GIANT PANDAS FROM CHINA ARRIVED ON A SPECIALLY RETROFITTED FEDEX JET

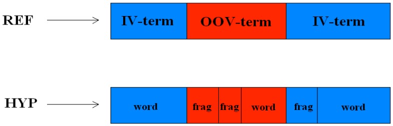
Fragment Posteriors Using Consensus

- Lattices are hard to deal with especially if you need their timings
- It would be easier to use the compact form of lattices → Confusion Networks
  - Having posterior probabilities for each hypothesis, we are able to observe the appearance of fragments and their likelihood.
  - To identify OOV regions in the confusion network we can compute an OOV score :

$$OOV_{score} = \sum_{f \in \{t_j\}} p(f|t_j)$$

where t_j is a given bin of the confusion network and $f's$ are fragments

Evaluating OOV detection



- The ASR transcript(output) is compared to the reference transcript at the *frame level* [forced alignment]
- Each frame is assigned a score equal to the OOV score of the region it belongs to [previous slide]
- Each frame is tagged as belonging to an OOV or IV region.
- *False alarm* probabilities and *miss* probabilities on the set are shown in standard detection error trade-off(DET) curves
- Entropy of bins inside confusion network is used as an OOV score for word systems

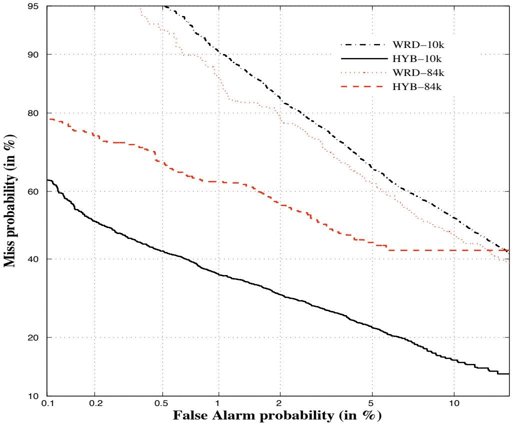


FIGURE: DET curves using hybrid and word system features

Outline

1 Introduction

2 Hybrid Systems for OOV Detection

3 Improving Phone Accuracy and Robustness
- Phone Error Rate
- Results

4 From Sub-word units to Words

5 Summary

- There are many applications in HLT which need an accurate automatic phone recognizer → e.g., Spoken term detection (STD)
    - In STD task OOV terms (queries) can not be detected and retrieved. New techniques have been proposed which are all essentially based on the phonetic search for OOV queries.
- It is a well known fact that LVCSR based systems have better phone accuracy than phone recognizer systems with phone LM
- Question: Is adding new words (enlarging the dictionary size) the only way to improve phone accuracy?
- Sub-word units are not specific to a given domain/genre and reveal the phonetic structure of the language → it is expected that applying them to out of domain data will substantially improve the phone accuracy.

- Phone Error Rate (PER) computation is done using the NIST scoring tool
- The phone sequence in the 1-Best is aligned with the reference phone sequence
- The reference phone sequence is obtained by forced-alignment to the reference transcript
  - Pronunciation of OOVs in the reference are obtained using letter to sound system.
- Oracle Phone error rate is also computed on the phonetic lattices. For this hybrid (word/fragment) lattices are converted to phonetic lattices
- In order to measure the contribution of the OOV regions to PER, $\frac{PER_{oov}}{PER}$ is computed and shown

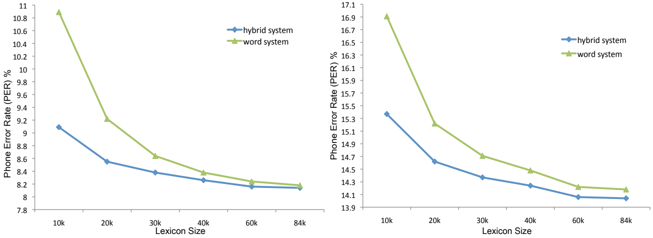


Figure: PER Results: (left) RT-04 (right) MIT Lectures

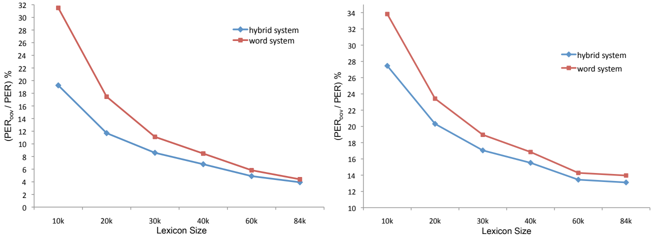


Figure: PER in OOV regions as a percentage of the overall PER: (left) RT-04 (right) MIT Lectures

Continued

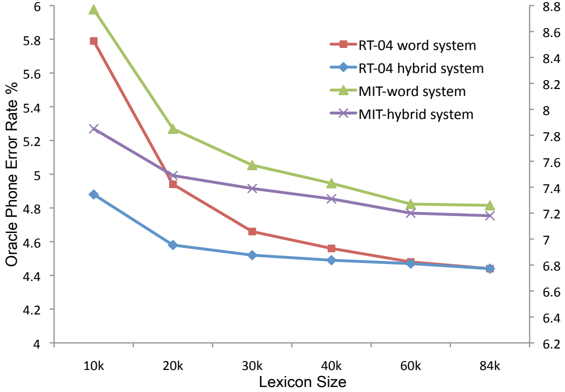


FIGURE: Oracle PER of word/hybrid systems on RT-04, shown on the left Y-axis and the MIT data set shown on the right Y-axis

Outline

1. Introduction
2. Hybrid Systems for OOV Detection
3. Improving Phone Accuracy and Robustness
4. From Sub-word units to Words
   - Results
5. Summary

- We can not expect the customer to be satisfied with the hybrid output!

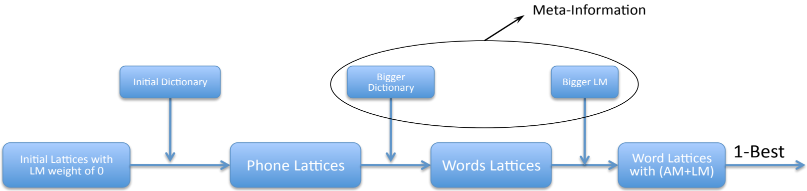
  FROM THE C. N. N. GLOBAL HEADQUARTERS IN ATLANTA I'M CAROL K_AA_S T_EH_L OW

  (COSTELLO). THANKS YOU FOR WAKING UP WITH US

- Even though the hybrid output is much better and more understandable than:

  FROM THE C. N. N. GLOBAL HEADQUARTERS IN ATLANTA I'M CAROL COX FELLOW

  (COSTELLO). THANKS YOU FOR WAKING UP WITH US



$$L \otimes d \otimes D_{inv} \otimes W$$

- In our experiments, the 84k Lexicon and LM information are used as Meta-Information

Vocab. Size	10k	20k	30k	40k	60k	84k
Hybrid (%)	15.5	14.9	14.6	14.4	14.2	14.1
Word(%)	17.1	16	15.1	14.6	14.3	14.1

TABLE: WER on the RT-04 Eval set after back-transduction in previous slide

Outline

1. Introduction
2. Hybrid Systems for OOV Detection
3. Improving Phone Accuracy and Robustness
4. From Sub-word units to Words
5. Summary

- Showed:
  - Basic method for fragment selection and building hybrid system
  - Appearance of fragments in the output is a good indicator of OOV regions (improvement over entropy of bins from word system)
  - Using fragments (along with words) improves the phone accuracy and can be helpful for STD task (for any lexicon size)
  - Hybrid system trained on a generic domain (where sufficient training data is available) can be used on domains with low resources
  - Hybrid system output is richer and is closer to the phonetic truth than the word system output

Questions/Comments