OUT-OF-VOCABULARY WORD RECOVERY USING FST-BASED SUBWORD UNIT CLUSTERING IN A HYBRID ASR SYSTEM

Ekaterina Egorova, Lukáš Burget

Brno University of Technology, Speech@FIT group, Czech Republic

e-mail: {iegorova,burget}@fit.vutbr.cz

Abstract

The paper presents a new approach to extracting useful information from out-of-vocabulary (OOV) speech regions in ASR system output. The system makes use of a hybrid decoding network with both words and sub-word units. In the decoded lattices, candidates for OOV regions are identified as sub-graphs of sub-word units. To facilitate OOV word recovery, we search for recurring OOVs by clustering the detected candidate OOVs. The metrics for clustering is based on a comparison of the sub-graphs corresponding to the OOV candidates. The proposed method discovers repeating out-of-vocabulary words and finds their graphemic representation more robustly than more conventional techniques taking into account only one best sub-word string hypotheses.

1 Goals

- to successfully detect OOV words
- to learn their acoustic model
- to find repeating OOVs
- to enhance dictionary and LM with the newly-discovered words

2 Data

- LibriSpeech ASR corpus of audiobooks: 1000 hours of 16kHz read English speech
- 3-gram ARPA LM trained on 14 500 public domain books
- 200 000 words in the dictionary
- 3-gram ARPA phonotactic language model trained on the dictionary
- 100 hours of clean data are used for system training
- a separate test set of 5 hours and 20 minutes for system performance evaluation
- discovery of OOVs is done on a bigger dataset of 360 hours

3 OOV Simulation on LibriSpeech Dataset

- OOVs are usually names and newly-coined words
- but LibriSpeech corpus majorly consists of free domain books, which are predominantly from 19th century - no "new" words
- we "reverse" the task and designate archaic and out-of-usage words as OOV words - they are not likely to be in a modern LM trained on Internet data
- archaic words and names are chosen based on Google ngram dataset of word usage statistics in books
- resulting OOV list is 1000 words long (ex. INTERPOSED, HASTENED, MADEMOISELLE, INDIGNANTLY, COUNTENANCE)
- 1.2% OOV rate
- on the 360 hours dataset the number of occurrences for the OOVs on the list ranges from 0 to 296, with the mean of 51 occurrences

4 Baseline System Description

- Kaldi baseline (nnet3 recipe): DNNs on top of the fMLLR features, using the decision tree and state alignments from the LDA+MLLT+SAT system as supervision for training
- 11.61% WER (with excluded OOVs)

5 OOV Recovery Procedure

- hybrid decoding graph is constructed by modifying G graph in a HCLG graph: phonotactic G graph is inserted instead of every OOV word (Fig. 1)
- OOV candidates are extracted from an index tree build on top of decoded lattices [Can & Saraçlar, 2011] (Fig.2)
- candidates are clustered based on their acoustic similarity (posterior on the shortest common path of the pairwise composition) - Fig. 3
- from the most-clustered candidates the best path is extracted as an input to p2g system trained on the dictionary

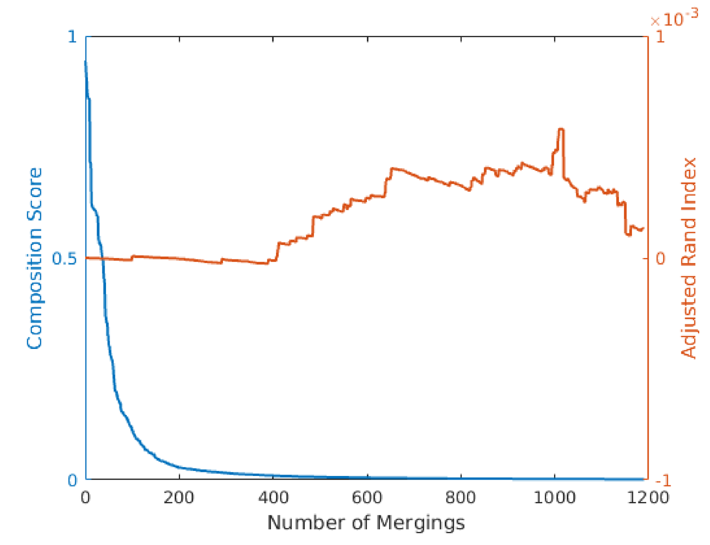

Figure 3: Composition Score and Adjusted Rand Index Score in hierarchical clustering of OOV hypotheses.

6 OOV Recovery Results

One-best clustering output (merged >2 times):

- 2 OOVs recovered correctly
- one is still recognizable
- recovery rate of 0.3%

COURAGE	K ER IH JH
VOYAGE	V OY IH JH
FLANE	F L EY N
KLEY'S	K L IY Z
SALOOSKI	S AH L UW S K IY
IMETHEUS	IH M IY TH IY AH S
ANCTIOUSLY	AE NG K SH AH S L IY

Full lattices clustering output. The number of the word's reference occurrences (where applicable) is in the brackets:

- 8 are ideally recovered words (four times as many as with one-best approach)
- 6 words that are still recognizable, although the graphemic representation is not completely right
- a name from *The Three Musketeers*
- a suffix and a phrase
- OOV recovery rate in full lattice clustering equals 1.4%; it is more than 4 times better than one-best clustering.

COURAGE (288)	K ER IH JH
VOYAGE (120)	V OY IH JH
THRACE	TH R EY S
THRONG (48)	TH R AO NG
SNESS	S N AH S
UNESE	AH N IY Z
ATHO'S	AE TH OW Z
HITHER (95)	HH IH DH ER
IGARLY (176)	IH G ER L IY
SAVAGE (182)	S AE V IH JH
WELLING	W EH L IH NG
ANXIOUS (296)	AE NG K SH AH S
BOLDLY (68)	B OW L D L IY
TRICHERY (43)	T R IH CH ER IY
DIGNITY (190)	D IH G N AH T IY
ERLOGINGS	ER L AA JH IH NG Z
ERNESSNESS	ER N AH S N AH S
ANCTIOUSLY (99)	AE NG K SH AH S L IY
CORMALIS	K AO R M AE L AH S
HITHERINTHITHER	HH IH DH ER IH N TH IH DH ER

Adding these newly-discovered OOVs to the dictionary with the learned pronunciation and to the LM as unigrams with the same probability as an OOV reduces WER from 11.77% to 11.62%

7 Conclusions

Lattice-based approach outperforms one-best approaches both in terms of OOV detection and in terms of the recovery of phonetic and graphemic representations of OOV words. There is promise of enhancing ASR user experience by bringing to her attention newly discovered words that may be added to the dictionary almost without adjustments.

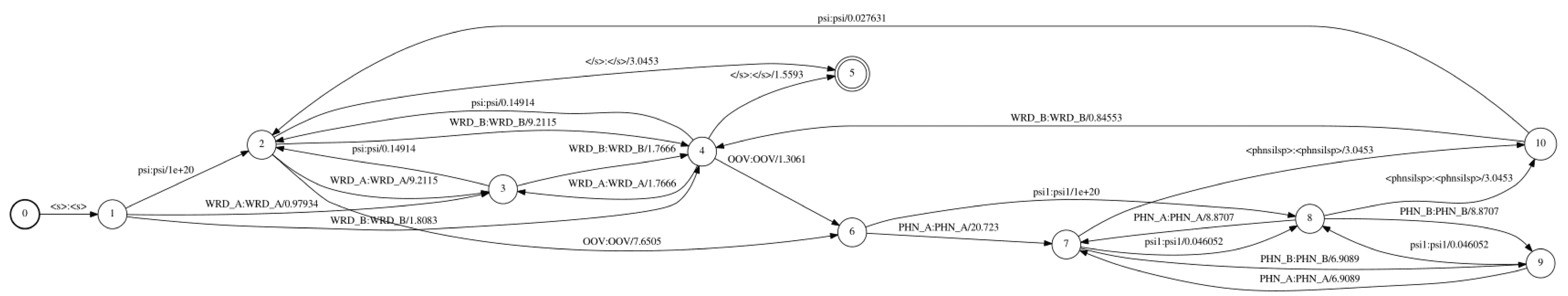

Figure 1: Hybrid Decoding Graph

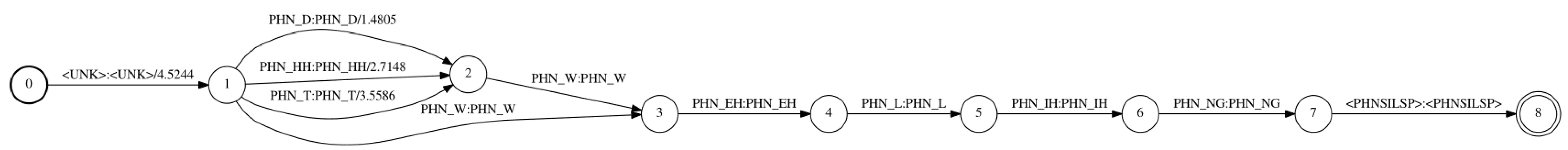

Figure 2: OOV candidate in lattice form

The Librispeech 1000 OOV list is available at **`http://www.fit.vutbr.cz/~iegorova/public/LibriSpeech_1000_OOV_list.txt`**