Recovery of Rare Words in Lecture Speech*

Stefan Kombrink[1], Mirko Hannemann[1], Lukáš Burget[1], and Hynek Heřmanský[1,2]

[1] Speech@FIT, Brno University of Technology, Czech Republic
[2] Johns Hopkins University, Baltimore, USA
{kombrink,ihannema,burget,iszoke}@fit.vutbr.cz, hynek@jhu.edu

Abstract. The vocabulary used in speech usually consists of two types of words: a limited set of common words, shared across multiple documents, and a virtually unlimited set of rare words, each of which might appear a few times only in particular documents. In most documents, however, these rare words are not seen at all. The first type of words is typically included in the language model of an automatic speech recognizer (ASR) and is thus widely referred to as in-vocabulary (IV). Words of the second type are missing in the language model and thus are called out-of-vocabulary (OOV). However, these words usually carry important information.

We use a hybrid word/sub-word recognizer to detect OOV words occurring in English talks and describe them as sequences of sub-words. We detected about one third of all OOV words, and were able to recover the correct spelling for 26.2% of all detections by using a phoneme-to-grapheme (P2G) conversion trained on the recognition dictionary. By omitting detections corresponding to recovered IV words, we were able to increase the precision of the OOV detection substantially.

1 Introduction

Since the early days of speech recognition, systems have been limited to vocabularies, which cover just the most common words. In many applications, however, the vocabulary is potentially unlimited. To deal with an unlimited vocabulary, one could either use techniques such as confidence measures [1,2], which have been developed to detect misrecognized speech due to OOV content. Another way is to use open vocabulary speech recognition systems, which enhance their vocabulary by using sub-words as e.g. in [3] and [4].

In this work, we also used a hybrid word/sub-word recognizer, where we modeled words and sub-words hierarchically as described in [5]. The word/sub-word recognizer detects OOV words, and retrieves their pronunciations as a preliminary description. Based on that, we used graphone models [6] to convert the preliminary phonetic description into a corresponding word spelling, what we refer to as *recovery of rare words*.

The focus of this work is on the recovery of rare IV and OOV words rather than on pure OOV detection. The spelling of detected sub-word sequences might be of interest due to several reasons:

* This work was partly supported by European project DIRAC (FP6-027787), by Grant Agency of Czech Republic project No. 102/08/0707, Czech Ministry of Education project No. MSM0021630528 and by BUT FIT grant No. FIT-10-S-2.

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 330–337, 2010.

- **correction:** We could substitute the sub-word sequences by the estimated spelling. By retrieving the estimated spellings of OOV words and even IV words, which got misrecognized as sub-word sequences for various reasons, we are able to correct a significant portion of them.
- **interpretability:** Non-linguists are not necessarily familiar with phonetic alphabets and find it bothersome to interpret such a description. However, they are able to guess a pronunciation from an approximate word spelling, even if seen for the first time and without knowing the meaning of the word.
- **consistency:** P2G conversion techniques are supposed to model the phonological relationship between pronunciations and spellings of words within a language. Especially if we train our P2G model on the dictionary used in the recognizer, the estimated spelling is consistent with existing words in the dictionary[1].

2 Hybrid Word/Sub-word Recognition

The utilized hybrid recognizer can be seen as a combination of two differently constrained recognizers interacting in parallel. The strongly constrained part takes into account word context by using a vocabulary of most frequent words and a general-purpose OOV word. The weakly constrained part models rare words using sub-word sequences by using a set of most frequent sub-word units. In the search for the most likely output sequence given an acoustic observation, the hybrid recognizer decides for the best path based on the overall likelihood of the composite word/sub-word model.

In the following, we show an example comparing the output of a standard word recognizer and our hybrid word/sub-word recognizer.

```
 REFERENCE: ...BACK TO BELGIUM(OOV) </s>
  WORD REC: ...BACK TO BALANCE THEM </s>
HYBRID REC: ...BACK TO <unk> b_ae_l jh_ih_ah_m <phnsilsp> </s>
```

The special words <unk> and <phnsilsp> mark transitions between the strongly and the weakly constrained parts of the recognizer. No word with an appropriate pronunciation for "BELGIUM" was found in the recognition dictionaries. The hybrid recognizer, however, was given the freedom to describe the missing word as a sequence of sub-word units, and hereby retrieving implicitly a phonetic description.

3 Partial OOV Detection

Generally speaking, the time region and the pronunciation estimated by the hybrid recognizer do not always match the reference precisely, e.g.:

```
  REFERENCE: ...GOES SUPERSTRING(OOV) THEORY WHAT...
RECOGNITION: ...GOES TO <unk> p_r_ih s_t_r_ih_ng p_iy r_iy <phnsilsp> WHAT...
```

[1] E.g. s_eh_n_t_axr recovers to CENTRE when trained on British English vs. CENTER when trained on American English.

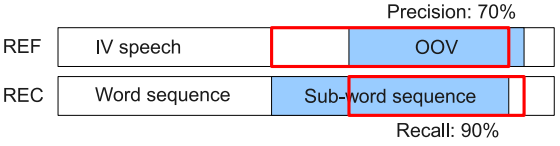

Fig. 1. Definition of Precision and Recall of a partially detected OOV Word: Precision is the percentage of OOV speech contained in the actual sub-word sequence, whereas recall is the percentage of OOV speech being detected

Hence, we decided to call sub-word sequences *partial OOV detections*. In the scope of this work, we are mainly interested in those partial detections which cover the time regions of the reference OOV word to an extent, that allows to recover the spelling from the retrieved phonetic transcription. Thus, based on the overlap between the boundaries in the reference and the recognition, we defined, how well a detected region will be suited for a word recovery task.

Figure 1 shows a typical partial OOV detection with inaccurate boundaries. The reference OOV word is being compared to the recognition of our hybrid word/sub-word decoder. Consequently, the sub-word sequence is interpreted as detected OOV, and the quality of that partial detection can be expressed precision and recall. We used the symmetric f-score to express the quality of a detected OOV word by just a single number:

$$f = \frac{2 \times precision \times recall}{precision + recall} \tag{1}$$

Detections with $f = 0$ are false alarms, whereas detections with $f = 1$ are perfect matches and thus should be most suitable for further description tasks. F-scores of partial overlaps will range between 0 and 1. In the recovery of OOV words, we aim at maximizing the number of detections and their f-scores.

4 Setup

4.1 TED Talks

TED[2] features a collection of more than 600 talks in English language about specialized topics given to a broad audience (`http://www.ted.com`). The vocabulary of many talks is highly specific (e.g. about biology, astronomy, politics . . .) and hence provide a potential source for topic-related OOV words. Manual reference transcripts are prepared in-house and being published as subtitles. Their intent, however, was to preserve correct meaning rather than to provide a word-by-word transcript as often used in the performance evaluation of speech recognition systems. We manually selected 45 talks (10 hours) of various lengths between 5 and 25 minutes.

4.2 Recognizer Setup

Our meeting recognizer developed for NIST Rich Transcription 2007 (RT07) evaluations within the AMI/AMIDA project [7] served as a baseline for reporting word

[2] With permission, TED.com.

Table 1. Word Error Rates using the Baseline Meeting Speech Recognizer and OOV rates using the word/sub-word Recognizer

Talks	Transcript (TED)	Transcript (manual)	OOV rate
All	29.5% WER	-	2.9%
Talk 1	60% WER	41% WER	4.5%
Talk 2	13% WER	13% WER	2.2%

accuracies. It used a 50k language model tuned on lectures, fast speaker adaptations (VTLN, CMLLR) and one-pass bi-gram lattice decoding. Decoding is done on PLP and posterior features processed using HLDA and CVN+CMN. Finally, the bi-gram lattices were expanded to 4-grams. The acoustic models were trained (SAT) on ca. 200 hours of meeting data recorded using independent headset microphones (IHM).

The hybrid recognizer used for partial OOV word detection was derived from this baseline system. The setup was reduced to the first pass (bi-gram decoding only) and the word recognition network was replaced by a hybrid word/sub-word recognition network.

4.3 Hybrid Language Model

The multigram sub-word language model [8] consisted of 3,977 phone and multiphone units trained on the AMI RT06 50k dictionary. The word language model was an open-set Katz-backoff language model trained on a total of 60M words from meetings (AMI/AMIDA, conversational telephone speech (CTS) and broadcast news data (BBC). The interpolation weights for the eight subsets were tuned using seven TED talks. The vocabulary size was fixed to 36k words by frequency cut-off 2 on meeting and CTS data. Due to the large amount of BBC data, this yielded in an uni-gram probability of approximately 1.5% for OOV words and a sensible number of OOV bi-grams. The hybrid language model consisting of a word and a sub-word model was combined in form of weighted finite state transducers by the use of the OpenFST toolkit[3].

4.4 OOV Transcripts

After mapping the reference transcript to unified UK spellings, we ran a forced-alignment to obtain a precise timing of all OOV words. The average OOV rate and word error rate (WER) on all TED talks is shown in the first row of Table 1. To get an impression, how much the available transcripts differ from an ideal word-by-word transcript, we manually transcribed two talks (see last two rows of Table 1). While the WER of talk 1 was notably reduced when using the manual transcript, we found, that the timing of the OOV words did not differ considerably from the TED transcripts.

The minimum and maximum OOV rate on TED talks was 0.3% and 5.1%. Against our assumption, we could not find strong correlation between WER and OOV rate. When checking the forced-aligned transcripts, we found untranscribed speech in some talks, and a few speakers making heavy use of repetitions and hesitations, which were not transcribed at all. This deteriorated word accuracy in many talks, but hardly affected the correct timing of OOV words.

[3] `http://www.openfst.org`

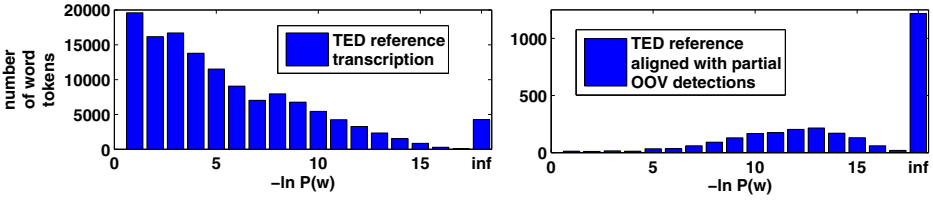

Fig. 2. Word frequency distribution of TED transcripts vs. all partial detections. Word tokens are binned by the negative log likelihood estimate of the language model.

Table 2. Correctness of all recognized words in dependence of their frequency

Frequency Bin	1	2–3	4–7	8–15	16–31	32–63	64–127	128–255	256–511	≥ 512
Correctness (%)	58.8	66.3	68.8	70.2	70.5	71.0	74.4	68.8	67.8	63.2

5 Analysis of Recognition Results

We investigated how the use of the hybrid word/sub-word model excels the recognition of words compared to a word-only recognizer. Therefore, we examine all reference words in TED depending on their estimated and real frequency. Figure 2 compares the word frequency distribution between all words in the TED data and those words, that caused partial OOV detections. It can be seen, that the majority of reference words overlapped with sub-word sequences were actually expected rarely, which is why the word language model estimated rather low likelihoods.

Next, we measured a per-word correctness[4] across all talks. Table 2 shows correctness per exponentially spaced frequency bin. It can be seen, that rarely and frequently recognized words were less often correct than words in the middle frequency ranges. To conclude, hybrid recognition improves recognition for words in the lower frequency regions, where the per-word correctness of the word recognizer is lower.

6 Results of Partial OOV Detection

Furthermore, we examined the quality of partial OOV detection in all TED talks containing 3,789 OOV words. Running on the fixed operating point provided by the one-best recognition output of the hybrid recognizer, where sub-word sequences were interpreted as partial detections, we obtained 2,898 partial OOV detections. By doing so, the system reached a precision of 40.9% and a recall of 31.4% (1,188 hits and 1,710 false alarms, see also table 3).

Figure 3 shows the precision, recall and f-score of all partial OOV detection tokens sorted independently by score. A high number of false alarms are shown in the right with scores of 0. But to the left, a fair number of partial detection tokens overlap almost perfectly.

[4] A recognized word has been considered correct iff there was an overlap in time with the reference word covering all phonemes of the reference word, and both reference and recognized words showed identical spelling.

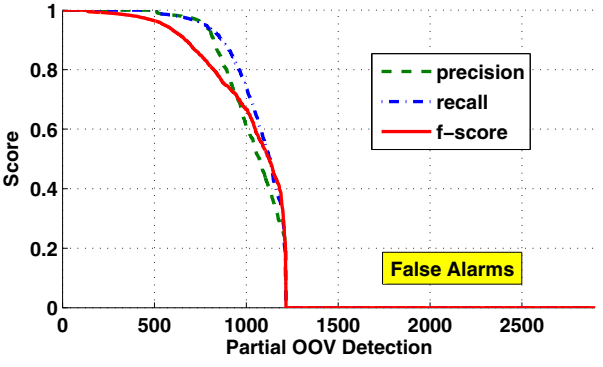

Fig. 3. Score distributions over partial OOV detections sorted by descending score

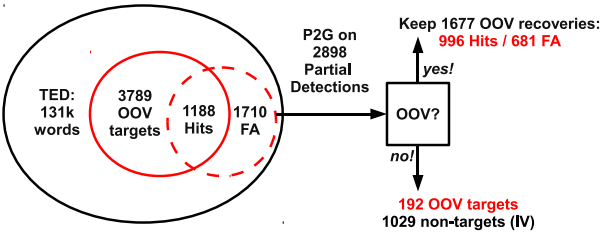

Fig. 4. Improving the Precision of partial OOV detection using P2G Conversion

7 Recovery of Rare Words

Partial OOV detections with high f-scores qualify for word recovery. Using Phoneme-to-Grapheme (P2G) conversion, we successfully recovered the correct spelling of detections from the phonetic description inherent in the corresponding sub-word sequences.

7.1 Setup

We trained a joint multigram P2G model up to 8-grams on the 36k decoding dictionary using Sequitur[5]. Co-alignments of length one and zero between phonemes and characters have been used. We kept 10% of all words for evaluation, where we obtained 21% WER and 5% CER for the generated pronunciations.

7.2 Results

During OOV recovery, we ran P2G on all partial detections, and divided those into two sets according to whether the obtained spelling was an OOV[6] or IV. Figure 4 shows,

[5] `http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html`

[6] I.e. a spelling which was not contained in our dictionary used for recognition.

Table 3. OOV Detection and Recovery Correction Statistics of Rare Words

3,789 targets/127k non-targets	Total	Hits	FAs	Precision	Recall	Corrected
all OOV detections	2,898	1,188	1,710	40.9%	31.4%	760 (26.2%)
recovered to OOV	1,677	996	681	59.4%	26.3%	277 (9.6%)
recovered to IV	1,221	192	1,029			483 (16.6%)

how recovery of partial OOV detections helped to filter out false alarms by omitting all IV recoveries. The following list shows the most frequent OOV recoveries:

12 × MYSOLIUM (MYCELIUM), O'RE (mainly false alarms)
11 × GRAVITATIONAL
8 × REPLICATOR
6 × SELEUM (MYCELIUM), PANDEMIC, GE (mainly false alarms), EXTINCTION, COURTICAL (CORTICAL)

The most frequent recoveries which converted to IV words could be categorized as follows:

- rare words: ORGAN, SEMEN, PSYCHO, PARA, PIPELINES, ...
- pre/suffixes: RE, IN, PRE, PRO, CON, ...
- fillers in between misrecognized words: N, M, S, SE, PH, SH, HA, ...

Table 3 shows the subset statistics in detail: The number of partial detections recovering to OOV was 1677, out of which 996 tokens were hits and 681 false alarms. This corresponded to an increase in OOV detection precision from 40.9% to 59.4% for the sake of recall, which in return decreased from 31.4% to 26.3%.

Furthermore it is shown, how recovery of rare words helped to correct the speech recognition output: approximately one quarter of the remaining hits (277 tokens of 211 types) were recovered to the correct spelling. This corresponds to about 7.3% of the total number of OOV words. In addition, 483 rare IV words of 365 types were recovered to the correct spelling. Altogether, this yielded in a successful recovery of 26.2% of all partial OOV detections, and 0.58% absolute reduction (2% relative) of the overall word error. Furthermore, many detections were found to recover into a readable form which would reveal the true meaning to a human reader. A demonstration of OOV word recovery using ten hours of Fisher telephone calls is available at `http://www.lectures.cz/_ted`.

8 Conclusion

In our experiments, we showed how hybrid word/sub-word recognition in combination with phoneme-to-grapheme conversion is able to recover rare words. The recovery task also motivated the introduction of partial detections and measuring their quality (f-score). Our experiments resulted in improvements of word accuracy and OOV detection performance. We suggest the hybrid word/sub-word recognition as a mean to improve the ASR accuracy especially on rare, information-rich words. Finally, the recovery of partial, potentially reoccurring OOV words, which get detected with a low f-score only, remains an interesting issue for future research.

References

1. Burget, L., et al.: Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVs. In: ICASSP (2008)
2. Jiang, H.: Confidence Measures for Speech Recognition: A Survey. Speech Communication 45(4), 455–470 (2005)
3. Bisani, M., Ney, H.: Open Vocabulary Speech Recognition with Flat Hybrid Models. In: Ninth European Conference on Speech Communication and Technology (2005)
4. Yazgan, A., et al.: Hybrid Language Models for out of Vocabulary Word Detection in Large Vocabulary Conversational Speech Recognition. In: ICASSP (2004)
5. Szoke, I., Fapso, M., Burget, L., Černocký, J.: Hybrid Word-Subword Decoding for Spoken Term Detection. In: Proc. SSCS 2008: Speech Search Workshop at SIGIR (2008)
6. Bisani, M., Ney, H.: Joint-Sequence Models for Grapheme-to-Phoneme Conversion. Speech Communication 50(5), 434–451 (2008)
7. Hain, T., et al.: The 2007 AMI(DA) System for Meeting Transcription. Multimodal Technologies for Perception of Humans, 414–428 (2008)
8. Deligne, et al.: Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams. In: ICASSP, Detroit, MI, pp. 169–172 (1995)
9. Hazen, T. J., Bazzi, I.: A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring. In: IEEE Intl. Conference on Acoustics, Speech and Signal Processing (2001)