SYSTEM COMBINATION FOR OUT-OF-VOCABULARY WORD DETECTION

Long Qin, Ming Sun, Alexander Rudnicky

Language Technologies Institute, School of Computer Science, Carnegie Mellon University

ABSTRACT

This paper presents a method to improve the out-of-vocabulary (OOV) word detection performance by combining multiple speech recognition systems' outputs. Three different fragment-word hybrid systems, the phone, subword, and graphone systems, were built for detecting OOV words. Then outputs from each individual system were combined using ROVER. Two combination metrics were explored in ROVER, voting by word frequency and voting by both word frequency and word confidence score. The experimental results show that the OOV word detection performance of the ROVER system with confidence scores is better than the ROVER system with only word frequency, as well as any of the individual hybrid systems.

Index Terms— OOV word detection, hybrid model, ROVER, confidence score

1. INTRODUCTION

Most speech recognition systems are closed-vocabulary recognizers and do not accommodate out-of-vocabulary (OOV) words. But in many applications, e.g. *voice search* or *spoken dialog systems*, OOV words are usually content words such as names and locations which embed crucial information to the success of these tasks. Speech recognition systems in which OOV words can be detected are therefore of great interest.

The fragment-word hybrid speech recognition system applies a hybrid language model (LM) during decoding to explicitly represent OOV words with phones, subwords, graphones, or generic word models [1][2][3][4]. Since different hybrid models had been individually proposed, in our previous work we compared OOV word detection and recovery performance of the phone, subword and graphone hybrid systems [5]. We found that the subword and graphone hybrid systems performed better than the phone hybrid system.

Combining outputs from multiple decoders has been used in many spoken language processing tasks. For example, Fiscus proposed ROVER which adopts a word voting scheme to improve the speech recognition performance [6]. Gales et al. investigated cross-site combination using cross-adaptation and ROVER for machine translation [7]. And Natori et al. studied the use of syllable transition network derived from multiple recognizers' outputs for spoken term detection [8].

In this paper, we investigated system combination techniques to improve the OOV word detection performance. ROVER was used to produce a combined OOV word detection hypothesis by voting among multiple hybrid systems' outputs. Rastrow et al. reported that adding confidence scores in the hybrid system improved the OOV word detection performance [9]. Besides the baseline ROVER method, which only measures word frequency in the voting module, we also studied the effectiveness of incorporating word confidence scores. The proposed system combination methods were tested on the Wall Street Journal (WSJ) and Broadcast News (BN) datasets.

The remainder of this paper is organized as follows. Section 2 describes the details of the hybrid system and system combination using ROVER. Sections 3 and 4 discuss our experiments and results. Concluding remarks are provided in Section 5.

2. METHOD

2.1. OOV word detection using a hybrid system

In the hybrid system, a fragment-word hybrid LM was applied during decoding to detect the presence of OOV words. We trained an open-vocabulary word LM from a large text corpus and a closed-vocabulary fragment LM from the pronunciations of in-vocabulary (IV) words. When training the word LM, all OOV words were matched to the same unknown token "$\langle unk \rangle$". Then by combining the word LM with the fragment LM, a single hybrid LM was generated. For example, the unigram probability of a fragment in the hybrid LM is calculated as

$$P_H(f_i) = P_W(\langle unk \rangle) \cdot P_F(f_i) \cdot C_{OOV}, \qquad (1)$$

where $P_W(\langle unk \rangle)$ is the unigram probability of the unknown token in the word LM, $P_F(f_i)$ is the unigram probability of the fragment in the fragment LM, and C_{OOV} is the cost of entering an OOV word during decoding. Similarly, we can compute N-gram probabilities in the hybrid LM.

2.2. Fragments

For the current study, we built three different hybrid systems, which are the phone, subword and graphone systems.

Of these, the phone and subword systems model only the phoneme level, while the graphone system also incorporates an orthography level.

2.2.1. Phone

In the phone hybrid system, an N-gram phone LM was trained and combined with a word LM to generate the hybrid LM. Then during decoding, OOV words were represented by phone sequences. For example, our system recognized the OOV word "ashland" as "AE SH AH N".

2.2.2. Subword

Subwords, such as "AH_N" and "EY_SH_AH_N", are iteratively trained phone sequences of variable lengths [2]. First, we added all phones to a subword inventory to ensure full coverage of all possible OOV words. In each iteration, the most frequent subword bigram was merged and added to the subword inventory. Its occurrences in the training corpus were also concatenated into one single entry. This transformed training data was then used in the next iteration. The training ended when a target number of subword units was reached.

2.2.3. Graphone

A graphone is a grapheme-phoneme pair of English letters and phones. For example, one possible representation of the word "speech" is

$$\text{speech} = \begin{pmatrix} \text{s} \\ \text{S} \end{pmatrix} \begin{pmatrix} \text{pee} \\ \text{P IY} \end{pmatrix} \begin{pmatrix} \text{ch} \\ \text{CH} \end{pmatrix}.$$

In our system, a trigram joint-sequence model was trained from the in-vocabulary (IV) dictionary and used to segment IV words into graphone sequences [10]. Then a graphone LM was trained and merged with a word LM to build the hybrid LM. A graphone can have a minimum and maximum number of letters and phones. Here, we used the same range for both letters and phones, where the minimum was set to 1 and the maximum was varied from 2 to 4.

2.3. System combination using ROVER

ROVER was originally developed at NIST to produce composite speech recognition system output when the outputs of multiple recognizers are available. In many cases, the composite recognition output has lower word error rate (WER) than any of the individual recognizers. In ROVER, the multiple recognizers' outputs are first combined into a single, minimal cost word transition network (WTN) via iterative applications of dynamic programming (DP) alignments. Then, the resulting WTN is re-scored and searched to find the optimal word sequence. The general rescoring formula is

$$Score(w_i) = \alpha \cdot \frac{N(w_i)}{\sum_w N(w_i)} + (1-\alpha) \cdot C(w_i), \qquad (2)$$

where $N(w_i)$ is the count of word w at the i-th alignment in the WTN, $C(w_i)$ is the confidence score of w_i, and α is the weight used to balance the word frequency and the confidence score. Another parameter $C(@)$ is used to set the confidence score of the NULL transition arc. For details of ROVER, please refer to [6].

The confidence score $C(w_i)$ is estimated from the confidence of word w_i in each individual system. Because we used ROVER for OOV word detection instead of speech recognition, we were more concerned about where the OOV word occurs in the transcription than what is the correct pronunciation of that word. When calculating the confidence score for w_i in the lattice of the j-th hybrid system, depending on whether w_i is an IV or OOV word, we summed over the posterior probabilities of all IV or OOV words in that region.

$$Conf_j(w_i) = \sum_{k \in [s_i, e_i]} \begin{cases} P(IV_k) & w_i \text{ is IV} \\ P(OOV_k) & w_i \text{ is OOV} \end{cases} \qquad (3)$$

where s_i and e_i are the start and end time of w_i, and $P(IV_k)$ is the posterior probability of an IV word in that region, while $P(OOV_k)$ is the posterior probability of an OOV word. $Conf_j(w_i)$ is then normalized by the sum of posterior probabilities of all words in that alignment to make sure $Conf_j(w_i) \in [0, 1]$. There are two ways to compute $C(w_i)$ from the individual confidence score $Conf_j(w_i)$, i.e., the average and the maximum of individual scores. In our experiments, we found that the performance of those two methods was essentially the same, although the latter one occasionally performed better.

In this paper, two ROVER systems with different rescoring modules were tested. In the baseline system, α in Eqn. 2 was set to 1, the optimal word sequence was found by only considering the frequency of word occurrences in each alignment from the WTN. In the second ROVER system, α was set to a value between 0 and 1, both the word occurrences and word confidence scores were measured when rescoring the WTN. Since different hybrid systems usually generate different fragment sequences for the same OOV word, ROVER cannot be applied directly. Therefore, we converted all fragment sequences in the recognizer output into the same OOV token "***OOV***". IV words were not changed so as to have a better alignment for building the transition network. Multiple outputs of individual systems were always aligned to the one with the best performance. The optimal values of α and $C(@)$ were searched from 0 to 1 with a step size of 0.2 using the grid search.

3. EXPERIMENT SETUP

3.1. Dataset

We tested our system on the Wall Street Journal (WSJ) Nov.92 20k evaluation task and the Broadcast News (BN) HUB4-96

20k F0 evaluation task. The WSJ0 and the BN 92-96 text corpora were used to train the word LM. In particular, the 20k most frequent words were choosed as vocabulary, yielding an OOV rate of 2% for the WSJ task and 4% for the BN task. An open-vocabulary 20k-word LM was trained for each task. The recognition dictionary was generated by looking up pronunciations for IV words in CMUdict (v.0.7a). The fragment LM was trained from the recognition dictionary without weighting each word by frequency. Then we built the bigram hybrid LMs for the WSJ and BN tasks, respectively. The acoustic models were trained from the WSJ-SI284 data and HUB4-96 BN data. The SPHINX3 decoder was used for recognition. The word error rate (WER) using the 20k-word bigram LM was 12.21% and 30.79% on the WSJ and BN task, which is comparable to the results of other groups.

3.2. Evaluation metrics

We measured the WER lower bound of each system by assuming that the detected OOV words could eventually be correctly recovered and recognized. This WER lower bound reveals how good the recognition performance would be if we recovered the orthography forms of detected OOV words.

We also used the *miss rate* and *false alarm (FA) rate* defined below to evaluate the OOV word detection performance.

$$Miss = \frac{\#\text{OOVs in reference} - \#\text{OOVs detected}}{\#\text{OOVs in reference}} \times 100\% \quad (4)$$

$$FA = \frac{\#\text{OOVs reported} - \#\text{OOVs detected}}{\#\text{IVs in reference}} \times 100\% \quad (5)$$

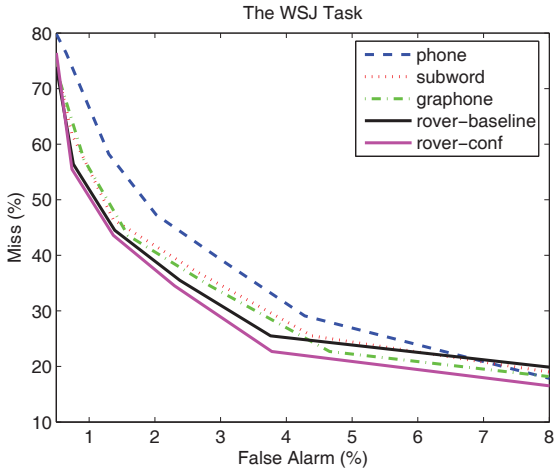
We calculated the *miss rate* and *false alarm rate* at the word level, which measures both the presence and positions of OOV words in an utterance. This is because, in practical applications, knowing where OOV words are located is more valuable than simply knowing the fact that OOV word(s) exist in an utterance.

4. RESULTS

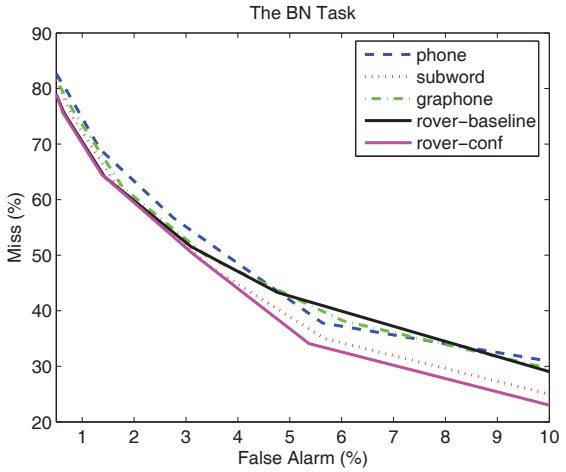
In our system, the optimal number of subwords and optimal graphone length were determined by testing on development data. For the WSJ task, the 500-subword system and the length 3 graphone system performed better than others. While for the BN task, we selected the 200-subword system and the length 2 graphone system. To draw the FA-Miss curve, when generating hybrid LMs, we adjusted the OOV cost C_{OOV} from 0 to 2.5 with a step size of 0.5. For ROVER, we combined outputs from multiple recognizers using LMs with the same OOV cost.

4.1. The baseline ROVER system

In the baseline ROVER system, when rescoring the transition network, only word frequency was considered. ROVER will identify an OOV word in a region if at least two systems reported OOV words in that alignment. Fig. 1 shows the OOV word detection results of the WSJ and BN tasks. We can find that the ROVER system with word frequency does not always outperform all individual systems. In fact, if two systems act similarly, the ROVER system usually tends to follow the performance of those two systems. For example, when the FA rate is low, the subword line and graphone line are very close and both lower than the phone line. In this case, the baseline ROVER system also beats the phone system, and even slightly better than the other two individual systems. But when the FA rate is high, sometimes two worse systems perform similarly. Then the ROVER system usually gets a poor performance and cannot win the best individual system. This fact can also be observed from Fig. 2, where the WER lower bound of the baseline ROVER system also follows the majority of three systems. Simply voting by word frequency isn't good enough

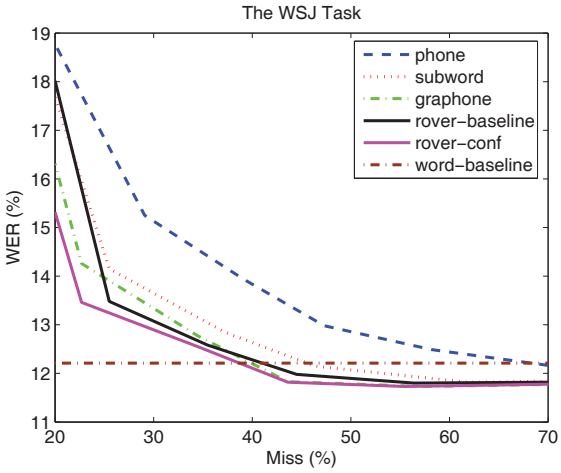


(a)

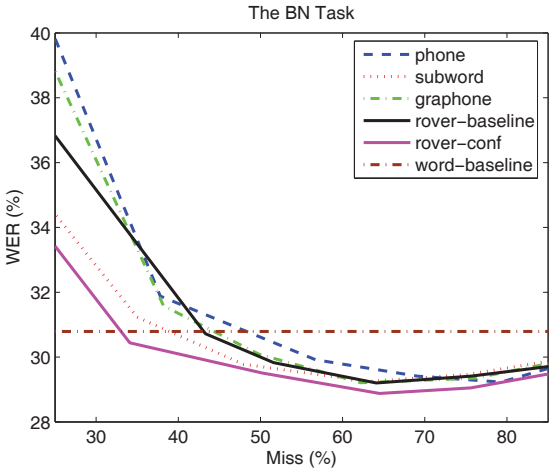


(b)

Fig. 1. The OOV word detection results



(a)



(b)

Fig. 2. The WER lower bound

for OOV word detection. Therefore, we incorporated word confidence score into the voting module.

4.2. ROVER with confidence score

In the second ROVER system, to rescore the transition network, both the word frequency and the word confidence score were used. The parameter α and $C(@)$ were determined using grid search on development sets. For both the WSJ and BN tasks, the optimal values for α and $C(@)$ are all 0.8. As presented in Fig. 1, different from the baseline ROVER system, the ROVER system with confidence score always outperforms all individual systems. We can also learn this from the WER lower bound of each system in Fig. 2, where the ROVER system with confidence score is much better than the other systems. Furthermore, the recognition performance of the proposed ROVER system (rover-conf) is much better than the word recognition system (word-baseline). This improvement in WER is due to two factors: 1) less errors were made on IV words, 2) more OOV words were correctly detected.

5. CONCLUSION AND FUTURE WORK

In this work, we studied system combination using ROVER for OOV word detection. Three hybrid systems with different fragment types were built—the phone, subword, and graphone systems. Two ROVER systems, the baseline ROVER system and the ROVER system with confidence score, were compared. From our experimental results, we found that system combination using only word frequency tends to follow the performance of the majority of systems. As a result, it doesn't work well if the performance of most individual systems is not satisfactory. On the other hand, when considering both word frequency and word confidence score, the ROVER system can outperform all individual hybrid systems.

In the future, we would like to investigate the effectiveness of incorporating other features, such as context features, into the rescoring computation. Additional evidence, which might be present in multiple N-best lists and lattices, may also yield additional accuracy in OOV word detection.

6. REFERENCES

[1] I. Bazzi, "Modelling out-of-vocabulary words for robust speech recognition," *Ph.D thesis*, MIT, 2002.

[2] D. Klakow, G. Rose, and X. Aubert, "OOV-detection in large vocabulary system using automatically defined word-fragments as fillers," *Proc. Eurospeech-1999*, pp. 49-52, 1999.

[3] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," *Proc. Interspeech-2005*, pp. 725-728, 2005.

[4] T. Schaaf, "Detection of OOV words using generalized word models and a semantic class language model," *Proc. Eurospeech-2001*, pp. 2581-2584, 2001.

[5] L. Qin, M. Sun, and A. Rudnicky, "OOV detection and recovery using hybrid models with different fragments," *Proc. Interspeech-2011*, pp. 1913-1916, 2011.

[6] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," *Proc. ASRU-1997*, pp. 347-354, 1997.

[7] M. J. F. Gales, X. Liu, et. al, "Speech recognition system combination for machine translation," *Proc. ICASSP-2007*, pp. 1277-1280, 2007.

[8] S. Natori, H. Nishizaki, and Y. Sekiguchi, "Japanese spoken term detection using syllable transition network derived from multiple speech recognizers' outputs," *Proc. Interspeech-2010*, pp. 681-684, 2010.

[9] A. Rastrow, A. Sethy, and B. Ramabhadran, "A new method for OOV detection using hybrid word/fragment system," *Proc. ICASSP-2009*, pp. 3953-3956, 2009.

[10] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, pp. 434-451, 2008.