A NEW METHOD FOR OOV DETECTION USING HYBRID WORD/FRAGMENT SYSTEM

Ariya Rastrow[1], *Abhinav Sethy*[2] *and Bhuvana Ramabhadran*[2]

(1) Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
(2) IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

ABSTRACT

In this paper, we propose a new method for detecting regions with out-of-vocabulary (OOV) words in the output of a large vocabulary continuous speech recognition (LVCSR) system. The proposed method uses a hybrid system combining words and data-driven variable length sub word units. With the use of a single feature, the posterior probability of sub word units, this method outperforms existing methods published in the literature. We also presents a recipe to discriminatively train a hybrid language model to improve OOV detection rate. Results are presented on the RT04 broadcast news task.

Index Terms— OOV, out-of-vocabulary, hybrid ASR system, discriminative training

1. INTRODUCTION

It is well known that out of vocabulary (OOV) words are an important source of error in the current large vocabulary continuous speech recognition (LVCSR) systems. The presence of OOVs can often cause mis-recognition of neighboring words. These errors propagate into the subsequent processing stages such as translation, understanding, document retrieval and term detection. Although, challenges within OOV modeling and detection are not new, issues with OOV words have traditionally been given less attention due to the fact that OOVs are rare and therefore they have low impact on the overall word error rate (WER) of LVCSR systems. However, OOVs occur inevitably due to the nature of human speech which contains proper names, foreign words and new words. Therefore, reliable detection of the presence and location of the OOV words can be used to improve the performance of real world applications of automatic speech recognition systems.

Many approaches have been proposed for OOV detection. They can be categorized into two broad groups:

1. **Filler Models** The first type of methods focuses on explicitly modeling OOVs using either filler or generic word models. Examples of such approaches can be found in [1, 2, 3]
2. **Confidence Scores** More recent approaches are focused on detecting OOVs based on some confidence measures such as acoustic scores, statistics derived from the language model and statistics derived from N-best lists (or lattices) [4, 5, 6, 7]

In general, methods using confidence scores have a better performance for OOV detection. The main weakness of this strategy is that such confidence measures are good at predicting whether the hypothesized word is correct or not, but unable to tease apart errors due to OOV words from those errors due to other phenomena such as degraded acoustic conditions. In [4, 7] the word-based system (strongly constrained with a language model) and the phone recognizer are used in parallel to address the OOV detection problem. These techniques are based on the comparison of the output of the two systems. A drawback of these approaches is that the phone recognizer suffers from high error rates making it an unreliable source for OOV detection.

In this paper, we propose a hybrid approach which directly combines words and sub-word units for OOV detection. The proposed method is based on a single feature, the posterior probability of sub word units, and outperforms existing methods. Also, based on the word/sub-word representation we have developed a technique using term weighted vectors to model the confusions inside the hybrid system and improve the OOV detection performance. Preliminary results using a hybrid LM that is discriminatively trained are also presented.

2. HYBRID WORD/FRAGMENT SYSTEM

In order to model OOV terms, we use a hybrid LVCSR system combining words and sub-word (fragments) units. Fragments are sub-word units which are variable length phone sequences and are selected automatically using statistical methods[8]. The criteria used for selecting an optimal set of fragments that provide good vocabulary coverage and discrimination between OOVs and similar-sounding in-vocabulary (IV) words are presented below. The hybrid ASR system uses the same acoustic models as a word-based LVCSR system while the language model is built from text that is tokenized into words and sub-word units.

2.1. Sub-word/Fragment Selection

Fragment[1] selection methods can be classified into two categories, namely, knowledge-driven methods that incorporate linguistic knowledge and data-driven methods [1, 8] which maximize an objective function. For fragment selection the approach suggested in [8] is used. The LM training text is converted into phones using the dictionary. All OOVs are excluded from the training set. Using this data set, an N-gram phone LM is built and pruned using a relative-entropy based method. This results in a set of fragments ranging from unigrams to N-gram phones. Some examples of fragments from our inventory include IH_N and K_L_AA_R_K.

2.2. Hybrid Language Model

The hybrid LM captures the dependencies between word and sub-word units. The vocabulary consists of a word lexicon and a sub-word unit lexicon. In order to ensure that we generate enough training data to model the fragments, the word portion of the vocabulary is limited. The LM training data is subsequently obtained by converting OOV terms in the text to their fragment representation.

[1] sub-word and fragment are used interchangeably throughout the paper

Pronunciations for the OOV terms are obtained using grapheme to phone models [9].

The set of fragments used to represent the OOVs in the LM text is selected in the following manner. A greedy search algorithm assigns the longest possible matching fragment first and iteratively uses the next longest possible fragment until the entire pronunciation of the OOV term has been represented by sub-word units. For example, consider the word, HAMDI which happens to have a pronunciation /HH/AE/M/D/IY and fragments HH_AE_M and D_IY are in the fragment inventory but HH_AE_M_D nor HH_AE_M_D_IY are not, then the fragment representation for the term would be /HH_AE_M D_IY/.

We also experimented with other techniques for tokenizing the LM text based on the degree of confusability of the fragments with the pronunciation of in-vocabulary words, i.e. selecting only those fragments that are less confusable with the words in the dictionary. This did not change the OOV performance compared to the greedy approach used in the baseline method described above.

Table (1) illustrates an example of tokenized hybrid text obtained using greedy search algorithm for tokenizing the LM text into sub-word units and words. A hybrid LM is built on the tokenized text treating each sub-word unit as an individual token.

$< s >$ THE BODY OF *ZIYAD HAMDI* WHO HAD BEEN SHOT WAS FOUND SOUTH OF THE CITY $< /s >$
$< s >$ THE BODY OF *Z_IY Y_AE_D HH_AE_M D_IY* WHO HAD BEEN SHOT WAS FOUND SOUTH OF THE CITY $< /s >$

Table 1. Tokenized Hybrid LM text

3. OOV DETECTION METHOD

Since fragments are used to represent OOVs while building the hybrid LM, the existence of these fragments in the ASR system's output can be used as a predictor of an OOV region. A simple solution to the OOV detection problem would then be reduced to a search for the fragments in the output of the ASR system. The search can be on the one-best transcripts, lattices or confusion networks. While lattices contain more information, they are harder to process and confusion networks on the other hand offer a nice trade-off between richness and compactness.

3.1. Fragment Posteriors Using Consensus

The confusion networks [10] contain the posterior probabilities of each unit in the network. This not only allows us to detect any OOV region in the confusion network by detecting the existence of fragments, but also provides a confidence measure for how likely it is. For any region in the confusion network we can compute an OOV score as given in Eqn. 1 to be the sum of the posteriors of all fragments inside that region.

$$OOV_{score} = \sum_{f \in \{t_j\}} p(f|t_j) \tag{1}$$

where t_j is the current region(bin) in the confusion network and f is the fragment inside that region.

Although sub word posteriors are very good for detecting OOV regions, we also explored the use of additional features in Eqn. 2 that contain complimentary information such as those used in the JHU workshop [4]. They include:

$$\begin{aligned} Word - Entropy &= - \sum_{w \in \{t_j\}} p(w|t_j) \log p(w|t_j) \\ Frag - Entropy &= - \sum_{f \in \{t_j\}} p(f|t_j) \log p(f|t_j) \\ LM - Score &= p_{lm}(hyp_{t_j}|hyp_{t_{j-1}}) \end{aligned} \tag{2}$$

where w is a word inside region t_j and hyp_{t_j} refers to the one-best hypothesis in the current region and $hyp_{t_{j-1}}$ refers to the one-best hypothesis in the previous region. p_{lm} is the probability of seeing hyp_j given hyp_{j-1} (bigram probability) obtained from the hybrid language model.

4. VECTOR SPACE MODELS FOR OOV DETECTION

An ASR system will make errors in both OOV and IV regions, where a set of fragments or an incorrect word is confused with spoken work. In order to model and capture these erroneous, we propose a term-weighted approach originally proposed in [11]. Consider the confusions in the regions of IV terms only. For each IV term we define a vector with a dimension equal to the total number of possible fragments. Each IV term's vector is then populated with the posterior probability corresponding to each fragment it is confused with. These confusions can be obtained from either lattices or confusion networks. Confusion networks were used in all the experiments reported in this paper. Eqn. 3 defines this vector of confusions, where F is the set of fragments and t_w is the region in the confusion networks where word w occurs in the reference in the training data set. Each q_i illustrates how confusable the word w is with fragment f_i.

$$\begin{aligned} \overline{V}(w) &= (q_1, q_2, \cdots, q_{|F|}) \\ q_i &= \frac{p(f_i|t_w)}{\sum_{f_j \in F} p(f_j|t_w)} \end{aligned} \tag{3}$$

Consider the confusion vector $\overline{V}(t_j)$

$$\begin{aligned} \overline{V}(t_j) &= (c_1, c_2, \cdots, c_{|F|}) \\ c_i &= \frac{p(f_i|t_j)}{\sum_{f_m \in F} p(f_m|t_j)} \end{aligned} \tag{4}$$

where t_j is the region in the test data's confusion network being considered. Now, we define α to capture the similarity between the confusion vector for this region, $\overline{V}(t_j)$ in the decoded network to the average of the confusion vectors, $\overline{V}_{avg}(w)$ of the one-best hypothesized word, w in this region, t_j. $\overline{V}_{avg}(w)$ is the average of $\overline{V}(w)$ over all occurrences of the word w in the training data.

$$\alpha = p(w|t_j).sim(\overline{V}_{avg}(w), \overline{V}(t_j)) \tag{5}$$

The similarity function (sim) used in this paper is the cosine similarity. For cosine similarity metric, α is a number between 0 and 1. Higher value for α indicates that the confusions we see in the test are similar to the confusions observed in the training data.

The following scores, $VM1$ and $VM2$ are used to detect the presence or absence of an OOV region.

$$VM1 = OOV_{score} - \alpha \tag{6}$$

$$VM2 = OOV_{score} \cdot (1 - \alpha) \tag{7}$$

Usually the number of fragments in the lexicon is of the order of 10K, which leads to sparsity issues for building confusion vectors.

This issue would be more severe if the type of the confusion we see in the training set is different from those seen on the test set. Projecting this vector to a lower-dimensional space will avoid this problem by reducing the dimension of the vectors without loosing any useful information. Our solution is to build these IV-term confusion vectors for the lower order N-gram phone units instead of fragments. Each dimension of these new vectors would represent n-gram phone units (For example, if $n = 1$, we will have vectors with the size of the number of phonemes). To compute this vector, every time a fragment is decoded, it is split into all possible n-gram sequences. The posterior probability of these sequences are obtained by summing the posterior probabilities of all fragments that contain this n-gram sequence and normalized using the posterior probabilities of all n-gram sequences that occur in the considered region. Since we are using a cosine similarity measure, normalization does not change the value of α. An alternative approach to capture the word-fragment confusions is to discriminatively train the hybrid language model using the technique proposed in [12] described in the next section.

5. DISCRIMINATIVE TRAINING

Discriminative training of the language model has been suggested to improve the performance of speech recognition systems[12]. The objective function captures the acoustic confusability and is formulated to minimize the word error rate. Discriminative training can help improve the LM for the purpose of better recognition by improving the separation of the correct hypothesis from competing hypotheses. We extend the same idea to the hybrid language model for improving OOV detection performance. The goal is to redistribute the hybrid language model probabilities based on the confusions in the hybrid system output, to reduce the confusability between words (IV terms) and fragments.

6. EXPERIMENTAL SETUP

6.1. Data

The RT04 Broadcast News Evaluation data was used as our test set. This set consists of roughly 45k word tokens in 4 hours. The LM training text consists of 335M words from the following data sources: 1996 CSR Hub4 Language Model data, EARS BN03 closed captions, GALE Phase 2 Distillation GNG Evaluation Supplemental Multilingual data , Hub4 acoustic model training transcripts , TDT4 closed captions, TDT4 newswire, and GALE Broadcast Conversations and GALE Broadcast News. For the discriminative training part we used the ASR acoustic training data set which includes 430 hours of speech data from the 1996 English Broadcast News Speech corpus, the 1997 English Broadcast News Speech corpus, and the TDT4 Multilingual Broadcast News Speech corpus.

To introduce enough OOVs in the evaluation data we limited our word portion of the lexicon to the 21142 most frequent (frequency greater than 5) words in the acoustic training data. This resulted in roughly 11M (3.1%) OOV tokens in the hybrid LM training set and 1127 (2.5%) OOV tokens in the evaluation set. The number of frames inside the OOV regions of the test set is 55155 (3.8%). For the fragment selection part we used a 5-gram phone language model and the hybrid LM is built with 4-gram contexts.

6.2. Evaluation

The *reference* to evaluate the performance of the OOV detection algorithm is obtained by aligning the reference transcript to the audio. The ASR transcript is compared to the reference transcript at the frame level. Each frame is assigned a score equal to the OOV score of the region. This score can be one of the following scores:

- OOV score defined in Eqn. 1
- Additional scores defined in Eqn. 2
- Compensated scores defined in Eqns. 6 and 7
- Combination of all these scores

Each frame is tagged as belonging to an OOV or IV region and this is obtained by aligning the decoder output with the reference. When a combination of scores is used, a classifier such as Maximum Entropy (MaxEnt) classifier[13] is used. False alarm probabilities and miss probabilities on the test set are shown in standard detection error trade-off (DET) curves which can be used to determine the operating point that optimally trades-off misses and false alarms for the task at hand.

7. RESULTS

Fig. 1 shows the DET curves for OOV detection using posterior probabilities of fragments and combination of that with other features described in Section 3. It is clear from Fig. 1 that the confidence measurement based on fragment posterior probability (dot-dashed line tagged as P_{frag}) in hybrid confusion networks outperforms existing methods based on confidence measures from LVCSR systems and word entropy (solid line further from the origin) of the word based system[4]. As shown by the closest solid line to origin in Fig. 1, adding other features from the confusion network ($word_{entropy}$, $frag_{entropy}$ and LM_{score}) to the posterior probability of fragments improves the detection performance in the regions where we have high false alarms. For example, accepting 10% false alarms (as an operating point) we increase detection rate from 78% to 85% (Misses from 22% to 15% respectively).

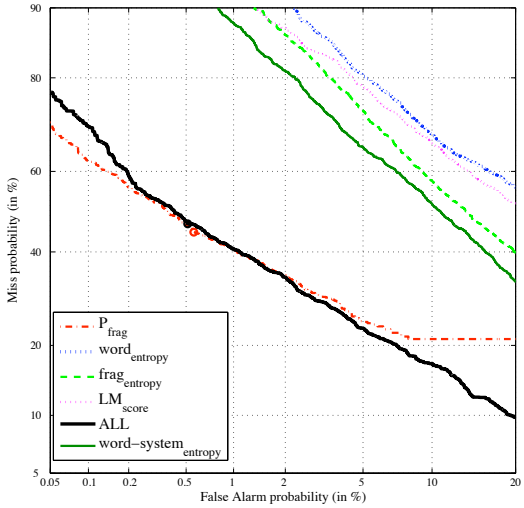

Fig. 1. OOV detection DET curves

Fig. 2 shows the DET curves for OOV detection using vector models and OOV scores defined in Section 4. Figure 2 illustrates the detection performance when using scores $VM1$ and $VM2$. Clearly, the false alarm rates have been reduced substantially using the vector models methods. This confirms the idea of trying to learn the confusions in the IV regions and reduce the posterior probability of

fragments in those regions to have better detection. In both graphs, the solid style curve shows the performance on the uni-gram based vectors and dotted style curve shows the performance on the fragment based vectors. It is clear that in both cases uni-gram based vectors have a slightly better performance as is expected from the discussion in Section 4.

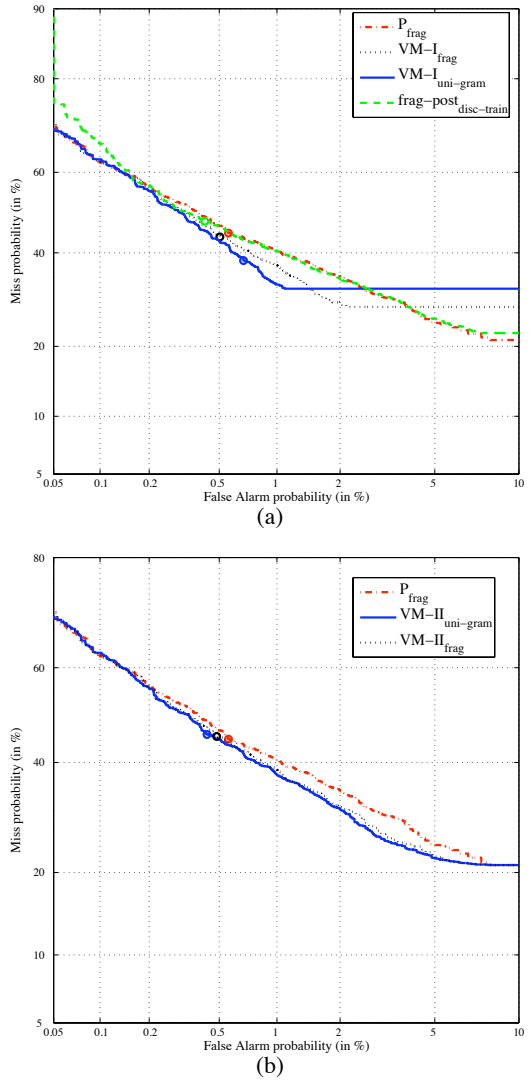

Fig. 2. OOV detection DET curves using Vector Models (a) Method I(VM1) (b) Method II(VM2)

Figure 2 also shows our results on OOV detection (dashed style curve in Fig. 2.a) using Eqn. 1 with a discriminatively trained language model as discussed in Section 5. Although we did not see any improvements in the overall OOV detection performance, we found that probabilities of fragments were significantly boosted which led to both a increase in false alarms and misses. To counter this problem, we are currently exploring updates of only those n-grams that contain at least one fragment.

8. CONCLUSION

We have presented a method for OOV detection using sub-word posterior probabilities and demonstrated how it outperforms other commonly used features in the literature. We have also proposed a new method for modeling confusions in the ASR output (in this case confusions from IV terms to fragments) and their subsequent use to significantly improve the performance of the proposed OOV detector. False alarms can be reduced from 2.5% to around 1% at 70% detection rate. Moreover, we showed the addition of other features such as word and sub-word entropy helps in improving the performance in the high false alarm regions. In the future, we plan to use the hybrid system in the Spoken Term Detection task to search for OOVs. We plan to extend the hybrid word/sub-word systems to the multi-lingual domain. By using a universal phone set, we would be able to build a set of fragments which spans and represents several languages.

9. REFERENCES

[1] I. Bazzi, *Modeling Out-of-Vocabulary Words for Robust Speech Recognition*, Ph.D. thesis, MIT, 2002.

[2] T. Schaaf, "Detection of OOV words using generalized word models and a semantic class language model," in *Proc. Eurospeech*, 2001.

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

[4] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. M. White, S. Khudanpur, H. Hermansk, and J. Cernock, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs," in *Proc. ICASSP*, 2008.

[5] H. Sun, G. Zhang, F. Zheng, and M. Xu, "Using word confidence measure for oov words detection in a spontaneous spoken dialog system," in *Proc. Eurospeech*, 2001.

[6] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," in *IEEE Trans. Speech and Audio Processing*, 2001, pp. 288–298.

[7] H. Lui and J. Blimes, "OOV detection by joint Word/Phone lattice alignment," in *Proc. of ASRU*, 2007.

[8] O. Siohan and M. Bacchiani, "Fast vocabulary-independent audio search using path-based graph indexing," in *Proc. Interspeech*, 2005, pp. 53–56.

[9] Stanley F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *Proc. Eurospeech*, 2003, pp. 2033–2036.

[10] L. Mangu, E. Brill and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. Eurospeech*, 1999.

[11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," in *Information Processing and Management*, 1988, pp. 513–523.

[12] H. J. Kuo, E. Fosler-Lussier, H. Jiang, C. Lee, "Discriminative training of language models for speech recognition," in *Proc. ICASSP*, 2002.

[13] A. Acero C. M. White, J. Dorppo and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Proc. ICASSP*, 2007.