Learning Units for Domain-Independent Out-of-Vocabulary Word Modelling

Issam Bazzi and James Glass

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139, USA
{issam, glass}@mit.edu

Abstract

This paper describes our recent work on detecting and recognizing out-of-vocabulary (OOV) words for robust speech recognition and understanding. To allow for OOV recognition within a word-based recognizer, the in-vocabulary (IV) word network is augmented with an OOV word model so that OOV words are considered simultaneously with IV words during recognition. We explore several configurations for the OOV model, the best of which utilizes a set of domain-independent, automatically derived, variable-length units. The units are created using an iterative bottom-up procedure where, at each iteration, the unit pairs with maximum mutual information are merged. When evaluating this method on a weather information domain, the false alarm rate of our baseline OOV model [1] is reduced by over 60%. For example, with an OOV detection rate of 70%, the OOV false alarm rate is reduced from 8.5% to 3.2%. At these settings the addition of the OOV model degrades the word error rate on IV data by only 0.3% absolute (3% relative).

1. Introduction

Given a finite vocabulary size, the presence of out-of-vocabulary (OOV) words is inevitable in any conversational speech recognition or understanding task. OOV and partially spoken words can be a source of both speech recognition and understanding errors. In our JUPITER weather system for example [2], the word error rate (WER) on data containing OOV words is nearly five times greater than in those containing only in-vocabulary (IV) words. While part of the WER increase is due to poor language modelling of out-of-domain queries, it is clear that OOV words cause recognition errors, and that an ability to identify OOV words would be beneficial [3].

In previous work we have examined sentence- and word-level confidence scoring to identify problematic utterances, such as those containing OOV words [4]. In this research we are exploring a different tactic by incorporating an explicit OOV word model into the word-based recognizer itself. One advantage of this approach, which we have described previously [1], is that it allows the OOV word to be predicted by a word-based language model. The OOV model is phone-based, so that an OOV word can be realized as an arbitrary sequence of phones. Currently, we use a phone bigram to provide phonotactic constraints within the OOV model. A recognizer with this configuration can therefore recognize words in the original vocabulary as well as any arbitrary new words.

This material is based upon work supported by the NSF under Grant No. IRI-9618731, and by DARPA under contract N66001-99-1-8904 monitored through NCCOSC.

In this paper we extend our previously reported work with the OOV model [1] by developing procedures which improve its performance and make it more domain-independent. Specifically, we describe a methodology we used to automatically derive a set of variable-length units for the OOV model. We also describe our work using dictionary-based methods for estimating n-grams for use within the OOV network. Both of these efforts produced significant improvements in the performance of the OOV model on our weather information task.

In the remainder of the paper we first review the recognition framework we developed for modelling OOV words. We then describe several configurations of the OOV network. We also present a method for automatically deriving a set of variable-length units using mutual information for building the OOV network. Finally, we present and discuss the results of a set of experiments in the JUPITER domain.

2. Hybrid Recognizer Framework

In devising a technique for explicitly modelling OOV words during recognition, we start with a word-based recognizer with a predefined word vocabulary. To model OOV words, we create a *generic word model* which allows for arbitrary phone sequences during recognition. One simple generic word model is a phonetic recognizer covering the set of phones in a language.

To allow for OOV words the recognizer vocabulary is augmented with a generic word model. As shown in Figure 1, the generic word model W_{OOV} is considered in parallel with all other words during recognition. Transitions between W_{OOV} and other words include a probability from the word-based language model used by the recognizer. The language model of the hybrid recognizer remains word-based, but now includes an entry for W_{OOV}. Since W_{OOV} is part of the vocabulary, the n-gram grammar treats it like any other word in the vocabulary.

In addition to the language model probability, entrance into the generic word model can be influenced by an OOV cost, C_{OOV}. This cost can be used to balance the contribution of the OOV phone grammar to the overall score of the utterance. For our experiments we varied the value of C_{OOV} to quantify receiver operating characteristic (ROC) behavior of the hybrid recognizer over a range of OOV detection and false alarm rates.

3. OOV Model Configurations

There are several requirements for the W_{OOV} model. It must be flexible enough to model the phonetic realization of any arbitrary word (with the possible exception of the active words in the vocabulary). It must also be accurate, both in its ability to

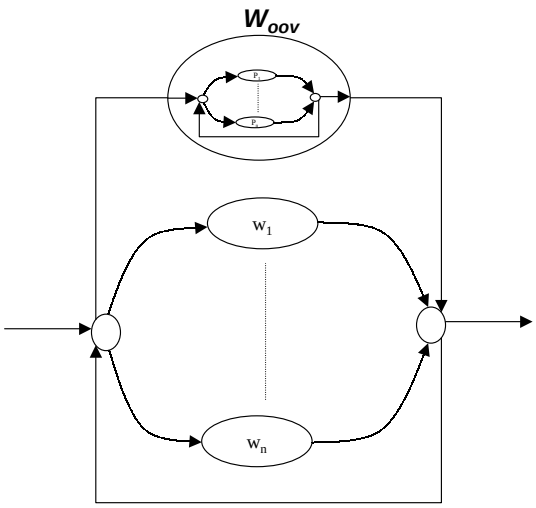

Figure 1: Recognizer configuration with generic word model.

correctly identify the phonetic sequence of an OOV word (possibly for further processing), and in its ability to discriminate between OOV words and IV words. In the following sections we describe the four OOV models we evaluate in this paper. First, we describe the baseline OOV model which is used in our experiments. Second, we describe an Oracle OOV model which was designed to measure an upper bound on OOV model performance. Third, we describe a dictionary-based OOV model which was designed to be domain-independent. Fourth, we describe a method which was used to create variable-length units for the OOV model.

3.1. The Baseline OOV Model

The baseline OOV model, which was reported in our initial work in this area [1], uses a phone n-gram to constrain allowable phonetic sequences in W_{OOV}. The phone n-gram is estimated from the same training corpus used to train the word recognizer, with words converted into their phonetic sequence. For the rest of this paper, we refer to the baseline OOV model as the corpus-based OOV model.

3.2. The Oracle OOV Model

After we obtained our first results with the corpus-based OOV model, we wanted to quantify how much the performance of W_{OOV} could be improved. We tried to answer this question by estimating the best possible performance we could achieve with our proposed framework. A good first approximation would be to build an Oracle OOV model which consisted solely of the OOV words contained in the test set. That is, the OOV network is constructed to allow for only the underlying phone sequences of these OOV words.

The Oracle OOV configuration is different from simply adding the OOV words to the recognizer vocabulary for two reasons. First, the n-gram probabilities will be those of the general OOV word as opposed to the n-gram probabilities of each word. Second, the cost of entering the OOV network C_{OOV} controls how often an OOV word is selected, which also changes the behavior of the recognizer.

3.3. A Dictionary-Based OOV Model

Although our initial work with the corpus-based OOV model worked well, there were several drawbacks to the approach which concerned us. First, since it was trained on phonetic transcriptions of sentences in the training corpus, the n-gram probabilities are influenced by the frequency of words in the corpus and will obviously favor more frequent words (e.g., *the*, *is*, and *at*). Second, in addition to modelling word-internal phonetic sequences, the n-gram would devote probability mass to cross-word sequences. Clearly, neither of these properties is desirable for modelling rare, OOV words. A third issue with the corpus-based OOV model that we disliked was the domain-dependent nature of the training corpus (i.e., we tested on different data from the same domain). We wanted to develop a more domain-independent mechanism for training the OOV model, since this would help achieve more robust performance in the long term.

To address these issues we train the OOV phone n-gram from a dictionary instead of a corpus of utterances. In this dictionary-based approach, we estimate the n-gram from phone sequences of a large domain-independent word dictionary (significantly larger than the word vocabulary of the recognizer). By using a large vocabulary, we reduce domain-dependence bias; by training on vocabulary items, we avoid modelling cross-word phonotactics, and eliminate biasing the OOV network towards frequent words (i.e., atypical OOV words).

3.4. The Mutual Information OOV Model

Although the dictionary-based OOV model constrained the W_{OOV} n-gram to model phone sequences in actual words, the topology is still quite simple. Incorporating additional structure into the model should provide more constraint, and reduce confusability with IV words. We therefore investigated an information theoretic approach to learning multi-phone units for use within the OOV model. We explored greedy methods which would measure phone co-occurance statistics in a large dictionary, and iteratively create multi-phone units which could be used to create the OOV model.

The method we adopted uses a bottom-up approach that starts with individual phones as the basic units and iteratively merges unit pairs together to form longer units. The criterion for merging a pair of units was based on the weighted mutual information of the pair; a metric used successfully for variable-length n-gram creation [5, 6]. For two units u_1 and u_2, the weighted mutual information $MI_w(u_1, u_2)$ is defined as:

$$MI_w(u_1, u_2) = p(u_1, u_2) \log \frac{p(u_1, u_2)}{p(u_1)p(u_2)}$$

Mutual information measures how much information one unit u_1 contains about the neighboring unit u_2. Note that when the two units are independent, $p(u_1, u_2) = p(u_1)p(u_2)$ and hence $MI_w(u_1, u_2) = 0$. On the other hand, the more dependent the two units are, the higher their mutual information. Since our mutual information is weighted by the joint probability $p(u_1, u_2)$, the frequency of the pair is also represented in our merging metric.

The iterative process to derive the variable-length units is applied as follows: First, we initialize the unit set to be the same as the phone set of the recognizer. At each iteration, we compute the weighted mutual information for all pairs of units that are encountered in the vocabulary. The pair (u_1, u_2) with the maximum MI_w is promoted to become a new unit. Every occurrence of the pair in the vocabulary is replaced with this new unit $u = u_1u_2$.

If this procedure is iterated indefinitely, the unit set will converge to the large vocabulary. The number of iterations we chose was decided empirically and chosen to represent a trade-off between the complexity of the OOV model and the speed of recognition.

One byproduct of our iterative process is a complete parse of all words in the vocabulary in terms of the derived units. We use the parses to estimate the OOV model n-gram parameters.

4. Experiments and Results

All the experiments for this work are within the JUPITER weather information domain [2]. A set of context-dependent diphone acoustic models were used, whose feature representation was based on the first 14 MFCCs averaged over 8 regions near hypothesized phonetic boundaries. Diphones were modeled using diagonal Gaussians with a maximum of 50 mixtures per model. The word lexicon consisted of a total of 2,009 words, many of which have multiple pronunciations. Bigram language models were used both at the word-level, as well as at the phone-level for the OOV model.

The training set used for these experiments consists of 88,755 utterances used to train both the acoustic and the language models. The test set consisted of 2,029 utterances, 314 of which contained OOV words (most of the OOV utterances had only one OOV word).

4.1. Mutual Information Results

To derive the variable-length units for the OOV model we used the LDC PRONLEX dictionary which contains 90,694 words with 99,202 unique pronunciations. Starting with an initial phone set of 62 phones, we performed 200 iterations over the unit inventory using the mutual information criterion. For computational reasons, on each iteration we created ten new units instead of just one, yielding a total of 1,977 acquired units.

Figure 2 plots the mutual information measure for the first 20 iterations. Each curve in the figure corresponds to the ordered mutual information values obtained for all existing unit pairs in a single iteration. Each curve therefore decreases monotonically when plotted against the rank of the ordered values. As one would expect, the top mutual information value (rank 0) decreases with each successive iteration. It is also interesting to observe that, on earlier iterations at least, the mutual information values drop off quickly, supporting our heuristic of merging the top ten pairs on each iteration. One can also see that as the iterations increase the curves started to level off. This behavior could possibly be used as a stopping criterion for the procedure.

In order to quantify the amount of constraint we were acquiring from the dictionary, we measured phone perplexity on the training data of the original 62 phone OOV model, the 1,977 unit MI OOV model, and a hypothetical OOV model with 99,202 units constrained by the word baseforms in the training vocabulary. The latter model would have been attained if the merging procedure had been run until each word became a recognition unit. Of course, this representation would be much larger, and would not generalize to words not in the training vocabulary. Phone transitions within a unit had a probability of 1.0, since they were deterministic. As expected, the MI OOV model significantly reduced perplexity of the original OOV model from 14.04 to 7.13. The perplexity of the 99K unit OOV model was 4.36.

When we analyzed the derived units we observed that around two thirds of the units are legal English syllables. The rest are either syllable fragments or multi-syllable phone sequences. Table 1 shows a list of some of the units obtained and the words they represent in the vocabulary. The average length of a derived unit was 3.2 phones, ranging from 1 to 9 phones.

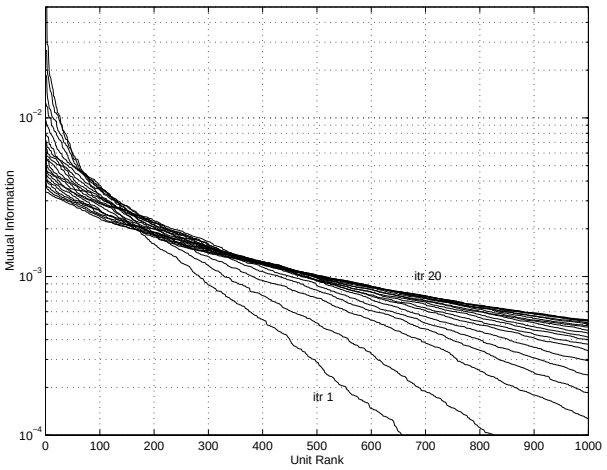

Figure 2: Ordered MI values for the first 20 iterations.

word	pronunciation
yugoslavian	y_uw g_ow s_l_aa v_iy ax_n
whisperers	w_ih s p_ax_r axr_z
shortage	sh_ao_r tf_ax jh

Table 1: Sample pronunciations with merged units.

4.2. OOV Detection Results

The behavior of the four OOV models was measured by observing the OOV detection and false alarm rates on the test set as C_{OOV} was varied. The presence or absence of an OOV word was based on the orthography of the top recognizer hypothesis. Figure 3 plots the ROC curves for the four different models. As can be seen from the figure, both the dictionary-based OOV model, and the subsequent mutual information OOV model have improved performance over the baseline corpus-based OOV model. Furthermore, the mutual information OOV model performance is approaching that attained by the Oracle OOV model.

In order to quantify the ROC behavior, a *figure of merit* (FOM), was computed which measured the area under the ROC curve. For our work we are most interested in the ROC region with low false alarm rates, since this produces a small degradation in recognition performance on IV data. For this reason we measured the FOM over the 0% to 10% false alarm rates. Note that the area is normalized by the total area in this range to produce an FOM whose optimal value is 1. For reference, a randomly guessing OOV model would produce an ROC curve which is a diagonal line (i.e., $y = x$). The FOM over the entire false alarm range would be 0.5, and the FOM over the 0% to 10% false alarm range would be 0.1. Table 2 summarizes the FOM measure for the various conditions both for the 0 to 10% range well as for the overall ROC area. All of our following discussion refers to the second set of FOM numbers (the first 10% of ROC curve).

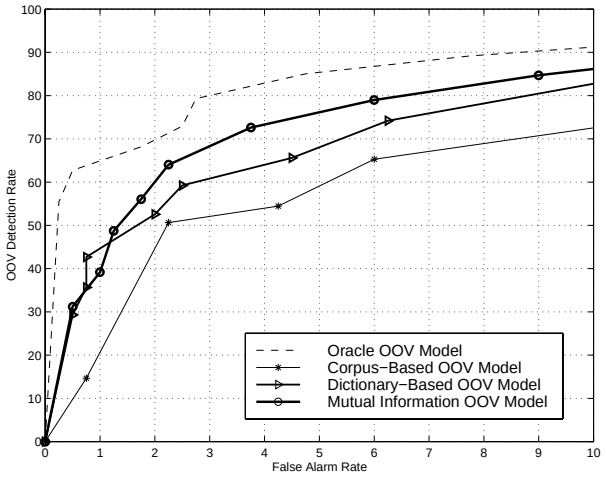

Figure 3: ROC curves for the four different OOV models.

For all four OOV models, the relationship between overall OOV false alarm rate and word error rate (WER) on IV test data is approximately linear. In the case of the dictionary-based OOV model for example, the WER increases slowly from the baseline WER of 10.9% at 0% false alarm rate to under 11.6% at 10% false alarm rate.

OOV Model	100% FOM	10% FOM
Corpus-based	0.89	0.54
Dictionary-based	0.93	0.64
Mutual Information-based	0.95	0.70
Oracle	0.97	0.80
Random	0.50	0.10

Table 2: The figure of merit performance of the OOV models.

5. Discussion

As expected, the oracle OOV network performs best in detecting OOV words by achieving an FOM of 0.80. This FOM gives an approximate upper bound on performance and gives some insight into how much we can possibly improve on our baseline FOM condition of 0.54. The Oracle OOV model we used was clearly sub-optimal; better performance would have been achieved if sentence-specific Oracle OOV models were used containing only sentence-specific OOV word(s). The joint Oracle OOV model was used since it was easier to compute, and provided at least a lower-bound on optimal performance.

With the dictionary-based OOV model, FOM improves from 0.54 for the baseline to 0.64. We were very encouraged by this result because of the domain-independent nature of the vocabulary used to train the dictionary-based OOV model n-gram. Unfortunately, we cannot quantify the individual contributions of 1) moving from continuous to isolated word n-gram training, and 2) moving from domain-dependent to domain-independent training, since the dictionary-based model differs from the corpus-based model by both of these factors. The latter factor would be difficult to quantify with the data we have at our disposal, since we do not have a large number of OOV words in our training data.

The best results were obtained using the automatically derived units using the mutual information criterion. We obtain an FOM of 0.70, which is a 30% improvement over the baseline system (of 0.54 FOM). The improvement from this method varies depending on the operating point on the ROC curve. For example, if we wish to operate at a detection rate of 70%, we notice from Figure 3 that, for this detection rate, the false alarm rate goes down from 8.5% in our baseline system to 3.2%, i.e., over 60% reduction in the false alarm rate. At these settings the addition of the OOV model degrades the WER on IV data from 10.9% to 11.2% (3% relative).

6. Conclusions & Future Work

This paper presented two new techniques for building an OOV model for use in parallel with a word-based recognizer to detect and recognize OOV words. The first technique relies on training an OOV model on a large domain-independent dictionary. The second technique uses mutual information to automatically derive a set of variable-length units from a large dictionary. These units are then used to construct the OOV model. The experimental results we achieve with these techniques significantly improve our baseline OOV model reported previously [1], and approach the results of an Oracle OOV which was used to estimate an upper bound on performance.

In future work we plan to explore different criteria, such as a normalized mutual information metric [7], or a multigram method [8], for deriving the variable-length units. We also plan on investigating the use of linguistically motivated units such as syllables. Apart from structural improvements within the OOV model, we are looking into extending our OOV framework, where multiple OOV classes are utilized to model various kinds of OOV words. Finally, we plan to measure the phonetic recognition accuracy within a detected OOV word, as we consider using a second-stage search with a large off-line dictionary to determine the identity of the OOV word itself.

7. References

[1] I. Bazzi and J. Glass, "Modelling out-of-vocabulary words for robust speech recognition," *Proc. ICSLP*, Beijing, 2000.

[2] V. Zue, et al., "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, 88(1), 2000.

[3] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic modeling for adding new words to a large vocabulary speech recognition system," *Proc. ICASSP*, Toronto, 1991.

[4] T. Hazen, T. Burianek, J. Polifroni, and S. Seneff, "Recognition confidence scoring for use in speech understanding systems," *Proc. ISCA ASR Workshop*, Paris, 2000.

[5] M. McCandless and J. Glass, "Empirical acquisition of language models for speech recognition," *Proc. ICSLP*, Yokohama, 1994.

[6] A. Gorin, G. Riccardi, and J. Wright, "How May I Help You," *Speech Communication*, 23, Oct. 1997.

[7] R. Kassel, "Automating the design of compact linguistic corpora," *Proc. ICSLP*, Yokohama, 1994.

[8] S. Deligne and F. Bimbot, "Language modeling by variable length sequences: Theoretical formulation and evaluation of multigram," *Proc. ICASSP*, Detroit, 1995.