Contextual Information Improves OOV Detection in Speech

Carolina Parada, Mark Dredze
HLTCOE
Johns Hopkins University
3400 North Charles Street,
Baltimore MD 21210, USA
carolinap@jhu.edu
mdredze@cs.jhu.edu

Denis Filimonov
HLTCOE
University of Maryland,
College Park, MD 20742 USA
den@cs.umd.edu

Frederick Jelinek
HLTCOE
Johns Hopkins University
3400 North Charles Street,
Baltimore MD 21210, USA
jelinek@jhu.edu

Abstract

Out-of-vocabulary (OOV) words represent an important source of error in large vocabulary continuous speech recognition (LVCSR) systems. These words cause recognition failures, which propagate through pipeline systems impacting the performance of downstream applications. The detection of OOV regions in the output of a LVCSR system is typically addressed as a binary classification task, where each region is independently classified using local information. In this paper, we show that jointly predicting OOV regions, and including contextual information from each region, leads to substantial improvement in OOV detection. Compared to the state-of-the-art, we reduce the missed OOV rate from 42.6% to 28.4% at 10% false alarm rate.

1 Introduction

Even with a vocabulary of one hundred thousand words, a large vocabulary continuous speech recognition (LVCSR) system encounters out-of-vocabulary (OOV) words, especially in new domains or genres. New words often include named entities, foreign words, rare and invented words. Since these words were not seen during training, the LVCSR system has no way to recognize them.

OOV words are an important source of error in LVCSR systems for three reasons. First, OOVs can never be recognized by the LVCSR system, even if repeated. Second, OOV words contribute to recognition errors in surrounding words, which propagate into to later processing stages (translation, understanding, document retrieval, etc.). Third, OOVs are often information-rich nouns – mis-recognized OOVs can have a greater impact on the understanding of the transcript than other words.

One solution is to simply increase the LVCSR system's vocabulary, but there are always new words. Additionally, increasing the vocabulary size without limit can sometimes produce higher word error rates (WER), leading to a tradeoff between recognition accuracy of frequent and rare words.

A more effective solution is to detect the presence of OOVs directly. Once identified, OOVs can be flagged for annotation and addition to the system's vocabulary, or OOV segments can be transcribed with a phone recognizer, creating an open vocabulary LVCSR system. Identified OOVs prevent error propagation in the application pipeline.

In the literature, there are two basic approaches to OOV detection: 1) *filler* models, which explicitly represent OOVs using a filler, sub-word, or generic word model (Bazzi, 2002; Schaaf, 2001; Bisani and Ney, 2005; Klakow et al., 1999; Wang, 2009); and 2) confidence estimation models, which use different confidence scores to find unreliable regions and label them as OOV (Lin et al., 2007; Burget et al., 2008; Sun et al., 2001; Wessel et al., 2001).

Recently, Rastrow et al. (2009a) presented an approach that combined confidence estimation models and filler models to improve state-of-the-art results for OOV detection. This approach and other confidence based systems (Hazen and Bazzi, 2001; Lin et al., 2007), treat OOV detection as a binary classification task; each region is independently classified using local information as IV or OOV. This work moves beyond this independence assumption

that considers regions independently for OOV detection. We treat OOV detection as a sequence labeling problem and add features based on the local lexical context of each region as well as global features from a language model using the entire utterance. Our results show that such information improves OOV detection and we obtain large reductions in error compared to the best previously reported results. Furthermore, our approach can be combined with any confidence based system.

We begin by reviewing the current state-of-the-art results for OOV detection. After describing our experimental setup, we generalize the framework to a sequence labeling problem, which includes features from the local context, lexical context, and entire utterance. Each stage yields additional improvements over the baseline system. We conclude with a review of related work.

2 Maximum Entropy OOV Detection

Our baseline system is the Maximum Entropy model with features from filler and confidence estimation models proposed by Rastrow et al. (2009a). Based on filler models, this approach models OOVs by constructing a hybrid system which combines words and sub-word units. Sub-word units, or fragments, are variable length phone sequences selected using statistical methods (Siohan and Bacchiani, 2005). The vocabulary contains a word and a fragment lexicon; fragments are used to represent OOVs in the language model text. Language model training text is obtained by replacing low frequency words (assumed OOVs) by their fragment representation. Pronunciations for OOVs are obtained using grapheme to phoneme models (Chen, 2003).

This approach also includes properties from confidence estimation systems. Using a hybrid LVCSR system, they obtain *confusion networks* (Mangu et al., 1999), compact representations of the recognizer's most likely hypotheses. For an utterance, the confusion network is composed of a sequence of *confused regions*, indicating the set of most likely word/sub-word hypotheses uttered and their posterior probabilities[1] in a specific time interval.

Figure 1 depicts a confusion network decoded by the hybrid system for a section of an utterance in our test-set. Below the network we present the reference transcription. In this example, two OOVs were uttered: "slobodan" and "milosevic" and decoded as four and three in-vocabulary words, respectively. A *confused region* (also called "bin") corresponds to a set of competing hypothesis between two nodes. The goal is to correctly label each of the "bins" as OOV or IV. Note the presence of both fragments (e.g. `s_l_ow`, `l_aa_s`) and words in some of the hypothesis bins.

For any bin of the confusion network, Rastrow et al. combine features from that region using a binary Maximum Entropy classifier (White et al., 2007). Their most effective features were:

$$\text{Fragment-Posterior} = \sum_{f \in t_j} p(f|t_j)$$

$$\text{Word-Entropy} = -\sum_{w \in t_j} p(w|t_j) \log p(w|t_j)$$

t_j is the current bin in the confusion network and f is a fragment in the hybrid dictionary.

We obtained confusion networks for a standard word based system and the hybrid system described above. We re-implemented the above features, obtaining nearly identical results to Rastrow et al. using Mallet's MaxEnt classifier (McCallum, 2002). [2] All real-valued features were normalized and quantized using the uniform-occupancy partitioning described in White et al. (2007).[3] The MaxEnt model is regularized using a Gaussian prior ($\sigma^2 = 100$), but we found results generally insensitive to σ.

3 Experimental Setup

Before we introduce and evaluate our context approach, we establish an experimental setup. We used the dataset constructed by Can et al. (2009) to evaluate Spoken Term Detection (STD) of OOVs; we refer to this corpus as `OOVCORP`. The corpus contains 100 hours of transcribed Broadcast News English speech emphasizing OOVs. There are 1290 unique OOVs in the corpus, which were selected with a minimum of 5 acoustic instances per word.

[1] $P(w_i|A)$: posterior probability of word i given the acoustics, which includes the language model and acoustic model scores, as described in (Mangu et al., 1999).

[2] Small differences are due to a change in MaxEnt library.

[3] All experiments use 50 partitions with a minimum of 100 training values per partition.

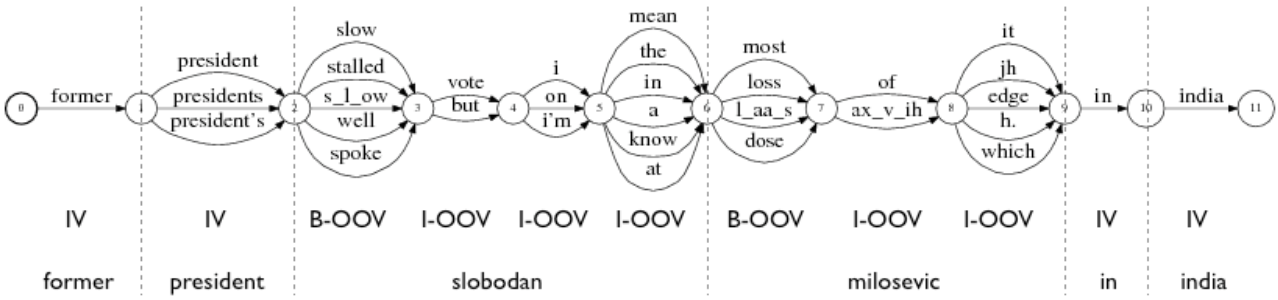

Figure 1: Example confusion network from the hybrid system with OOV regions and BIO encoding. Hypothesis are ordered by decreasing value of posterior probability. Best hypothesis is the concatenation of the top word/fragments in each bin. We omit posterior probabilities due to spacing.

Common English words were filtered out to obtain meaningful OOVs: e.g. NATALIE, PUTIN, QAEDA, HOLLOWAY. Since the corpus was designed for STD, short OOVs (less than 4 phones) were explicitly excluded. This resulted in roughly 24K (2%) OOV tokens.

For a LVCSR system we used the IBM Speech Recognition Toolkit (Soltau et al., 2005)[4] with acoustic models trained on 300 hours of HUB4 data (Fiscus et al., 1998) and excluded utterances containing OOV words as marked in OOVCORP. The language model was trained on 400M words from various text sources with a 83K word vocabulary. The LVCSR system's WER on the standard RT04 BN test set was 19.4%. Excluded utterances were divided into 5 hours of training and 95 hours of test data for the OOV detector. Both train and test sets have a 2% OOV rate. We used this split for all experiments. Note that the OOV training set is different from the LVCSR training set.

In addition to a word-based LVCSR system, we use a hybrid LVCSR system, combining word and sub-word (fragments) units. Combined word/sub-word systems have improved OOV Spoken Term Detection performance (Mamou et al., 2007; Parada et al., 2009), better phone error rates, especially in OOV regions (Rastrow et al., 2009b), and state-of-the-art performance for OOV detection. Our hybrid system's lexicon has 83K words and 20K fragments derived using Rastrow et al. (2009a). The 1290 excluded words are OOVs to both the word and hybrid systems.

Note that our experiments use a different dataset than Rastrow et. al., but we have a larger vocabulary (83K vs 20K), which is closer to most modern LVCSR system vocabularies; the resulting OOVs are more challenging but more realistic.

3.1 Evaluation

Confusion networks are obtained from both the word and hybrid LVCSR systems. In order to evaluate the performance of the OOV detector, we align the reference transcript to the audio. The LVCSR transcript is compared to the reference transcript at the confused region level, so each confused region is tagged as either OOV or IV. The OOV detector assigns a score/probability for IV/OOV to each of these regions.

Previous research reported OOV detection accuracy on all test data. However, once an OOV word has been observed in the training data for the OOV detector, even if it never appeared in the LVCSR training data, it is no longer truly OOV. The features used in previous approaches did not necessarily provide an advantage on observed versus unobserved OOVs, but our features do yield an advantage. Therefore, in the sections that follow we report unobserved OOV accuracy: OOV words that do not appear in either the OOV detector's or the LVCSR's training data. While this penalizes our results, it is a more informative metric of true system performance.

We present results using standard detection error tradeoff (DET) curves (Martin et al., 1997). DET

[4]We use the IBM system with speaker adaptive training based on maximum likelihood with no discriminative training.

curves measure tradeoffs between misses and false alarms and can be used to determine the optimal operating point of a system. The x-axis varies the false alarm rate (false positive) and the y-axis varies the miss (false negative) rate; lower curves are better.

4 From MaxEnt to CRFs

As a classification algorithm, Maximum Entropy assigns a label to each region independently. However, OOV words tend to be recognized as two or more IV words, hence OOV regions tend to co-occur. In the example of Figure 1, the OOV word "slobodan" was recognized as four IV words: "slow vote i mean". This suggests that sequence models, which jointly assign all labels in a sequence, may be more appropriate. Therefore, we begin incorporating context by moving from classification to sequence models.

MaxEnt classification models the target label as $p(y_i|\mathbf{x}_i)$, where y_i is a discrete variable representing the ith label ("IV" or "OOV") and $\mathbf{x}_i$ is a feature vector representing information for position i. The conditional distribution for y_i takes the form

$$p(y_i|\mathbf{x}_i) = \frac{1}{Z(\mathbf{x}_i)} \exp(\sum_{k=1}^{K} \lambda_k f_k(y_i, \mathbf{x}_i)) \ ,$$

$Z(\mathbf{x}_i)$ is a normalization term and $f(y_i, \mathbf{x}_i)$ is a vector of K features, such as those defined in Section 2. The model is trained discriminatively: parameters λ are chosen to maximize conditional data likelihood.

Conditional Random Fields (CRF) (Lafferty et al., 2001) generalize MaxEnt models to sequence tasks. While having the same model structure as Hidden Markov Models (HMMs), CRFs are trained discriminatively and can use large numbers of correlated features. Their primary advantage over MaxEnt models is their ability to find an optimal labeling for the entire sequence rather than greedy local decisions. CRFs have been used successfully used in numerous text processing tasks and while less popular in speech, still applied successfully, such as sentence boundary detection (Liu et al., 2005).

A CRF models the entire label sequence $\mathbf{y}$ as:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\lambda F(\mathbf{y}, \mathbf{x})) \ ,$$

where $F(\mathbf{y}, \mathbf{x})$ is a global feature vector for input sequence $\mathbf{x}$ and label sequence $\mathbf{y}$ and $Z(\mathbf{x})$ is a normalization term.[5]

5 Context for OOV Detection

We begin by including a minimal amount of local context in making OOV decisions: the predicted labels for adjacent confused regions (bins). This information helps when OOV bins occur in close proximity, such as successive OOV bins. This is indeed the case: in the OOV detector training data only 48% of OOV sequences contained a single bin; sequences were of length 2 (40%), 3 (9%) and 4 (2%). We found similar results in the test data. Therefore, we expect that even a minimal amount of context based on the labels of adjacent bins will help.

A natural way of incorporating contextual information is through a CRF, which introduces dependencies between each label and its neighbors. If a neighboring bin is likely an OOV, it increases the chance that the current bin is OOV.

In sequence models, another technique for capturing contextual dependence is the label encoding scheme. In information extraction, where sequences of adjacent tokens are likely to receive the same tag, the beginning of each sequence receives a different tag from words that continue the sequence. For example, the first token in a person name is labeled `B-PER` and all subsequent tokens are labeled `I-PER`. This is commonly referred to as BIO encoding (beginning, inside, outside). We applied this encoding technique to our task, labeling bins as either `IV` (in vocabulary), `B-OOV` (begin OOV) and `I-OOV` (inside OOV), as illustrated in Figure 1. This encoding allows the algorithm to identify features which might be more indicative of the beginning of an OOV sequence. We found that this encoding achieved a superior performance to a simple `IV`/`OOV` encoding. We therefore utilize the BIO encoding in all CRF experiments.

Another means of introducing context is through the order of the CRF model. A first order model ($n = 1$) adds dependencies only between neighboring labels, whereas an n order model creates dependencies between labels up to a distance of n positions. Higher order models capture length of label

[5] CRF experiments used the CRF++ package http://crfpp.sourceforge.net/

regions (up to length n). We experiment with both a first order and a second order CRF. Higher order models did not provide any improvements.

In order to establish a comparative baseline, we first present results using the same features from the system described in Section 2 (Word-Entropy and Fragment-Posterior). All real-valued features were normalized and quantized using the uniform-occupancy partitioning described in White et al. (2007).[6] Quantization of real valued features is standard for log-linear models as it allows the model to take advantage of non-linear characteristics of feature values and is better handled by the regularization term. As in White et. al. we found it improved performance.

Figure 2 depicts DET curves for OOV detection for the MaxEnt baseline and first and second order CRFs with BIO encoding on unobserved OOVs in the test data. We generated predictions at different false alarm rates by varying a probability threshold. For MaxEnt we used the predicted label probability and for CRFs the marginal probability of each bin's label. While the first order CRF achieves nearly identical performance to the MaxEnt baseline, the second order CRF shows a clear improvement. The second order model has a 5% absolute improvement at 10% false alarm rate, despite using the identical features as the MaxEnt baseline. Even a small amount of context as expressed through local labeling decisions improves OOV detection.

The quantization of the features yields quantized prediction scores, resulting in the non-smooth curves for the MaxEnt and 1st order CRF results. However, when using a second order CRF the OOV score varies more smoothly since more features (context labels) are considered in the prediction of the current label.

6 Local Lexical Context

A popular approach in sequence tagging, such as information extraction or part of speech tagging, is to include features based on local lexical content and context. In detecting a name, both the lexical form "John" and the preceding lexical context "Mr." provide clues that "John" is a name. While we do not know the actual lexical items in the speech sequence, the speech recognizer output can be used as a best guess. In the example of Figure 1, the words "former president" are good indicators that the following word is either the word "of" or a name, and hence a potential OOV. Combining this lexical context with hypothesized words can help label the subsequent regions as OOVs (note that none of the hypothesized words in the third bin are "of", names, or nouns).

[6] All experiments use 50 partitions with a minimum of 100 training values per partition.

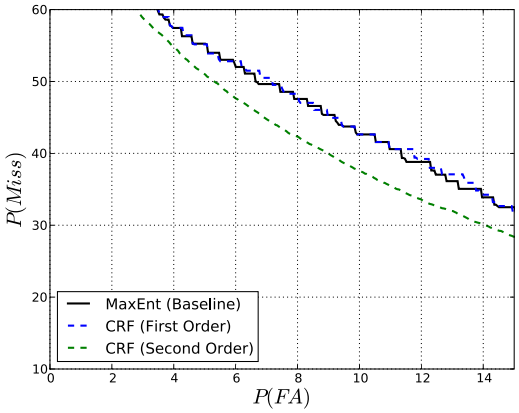

Figure 2: DET curves for OOV detection using a Maximum Entropy (MaxEnt) classifier and contextual information using a 1st order and 2nd order CRF. All models use the same baseline features (Section 2).

Words from the LVCSR decoding of the sentence are used in the CRF OOV detector. For each bin in the confusion network, we select the word with the highest probability (best hypothesis). We then add the best hypothesis word as a feature of the form: `current_word=X`. These features capture how the LVCSR system incorrectly recognizes OOV words. However, since detection is measured on unobserved OOVs, these features alone may not help.

Instead, we turn to lexical context, which includes correctly recognized IV words. We evaluate the following sets of features derived from lexical context:

- Current bin's best hypothesis. (Current-Word)
- Unigrams and bigrams from the best hypothesis in a window of 5 words around current bin. This feature ignores the best hypothesis in the current bin, i.e., `word[-2],word[-1]` is included, but `word[-1],word[0]` is not. (Context-Bigrams)

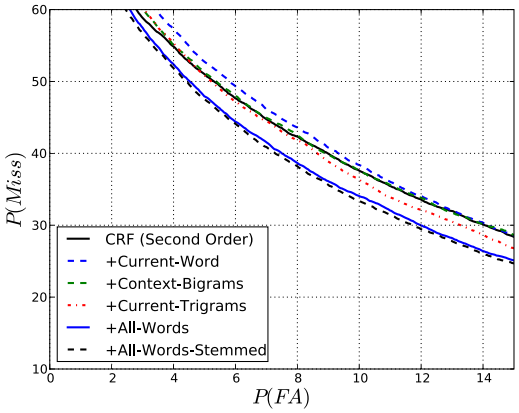

Figure 3: A second order CRF (Section 5) and additional features including including word identities from current and neighboring bins (Section 6).

- Unigrams, bigrams, and trigrams in a window of 5 words around and including current bin. (Current-Trigrams)
- All of the above features. (All-Words)
- All above features and their stems.[7] (All-Words-Stemmed)

We added these features to the second order CRF with BIO encoding and baseline features (Figure 3). As expected, the current words did not improve performance on unobserved OOVs. When the current words are combined with the lexical context and their lemmas, they give a significant boost in performance: a 4.2% absolute improvement at 10% false alarm rate over the previous CRF system, and 9.3% over the MaxEnt baseline. Interestingly, only combining context and current word gives a substantial gain. This indicates that OOVs tend to occur with certain distributional characteristics that are independent of the OOV word uttered (since we consider only unobserved OOVs), perhaps because OOVs tend to be named entities, foreign words, or rare nouns. The importance of distributional features is well known for named entity recognition and part of speech tagging (Pereira et al., 1993). Other features such as sub-strings or baseline features (Word-Entropy, Fragment-Posterior) from neighboring bins did not provide further improvement.

[7]To obtain stemmed words, we use the CPAN package: http://search.cpan.org/~snowhare/Lingua-Stem-0.83.

7 Global Utterance Context

We now include features that incorporate information from the entire utterance. The probability of an utterance as computed by a language model is often used as a measure of fluency of the utterance. We also observe that OOV words tend to take very specific syntactic roles (more than half of them are proper nouns), which means the surrounding context will have predictive lexical and *syntactic* properties. Therefore, we use a syntactic language model.

7.1 Language Models

We evaluated both a standard trigram language model and a syntactic language model (Filimonov and Harper, 2009a). The syntactic model estimates the joint probability of the word and its syntactic tag based on the preceding words and tags. The probability of an utterance w_1^n of length n is computed by summing over all latent syntactic tag assignments:

$$p(utt) = p(w_1^n) = \sum_{t_1 \ldots t_n} \prod_{i-1}^{n} p(w_i, t_i | w_1^{i-1}, t_1^{i-1}) \quad (1)$$

where w_i and t_i are the word and tag at position i, and w_1^{i-1} and t_1^{i-1} are sequences of words and tags of length $i-1$ starting a position 1. The model is restricted to a trigram context, i.e., $p(w_i, t_i | w_{i-2}^{i-1}, t_{i-2}^{i-1})$; experiments that increased the order yielded no improvement.

We trained the language model on 130 million words from Hub4 CSR 1996 (Garofolo et al., 1996). The corpus was parsed using a modified Berkeley parser (Huang and Harper, 2009) and tags extracted from parse trees incorporated the word's POS, the label of its immediate parent, and the relative position of the word among its siblings. [8] The parser required separated contractions and possessives, but we recombined those words after parsing to match the LVCSR tokenization, merging their tags. Since we are considering OOV detection, the language model was restricted to LVCSR system's vocabulary.

[8]The *parent* tagset of Filimonov and Harper (2009a).

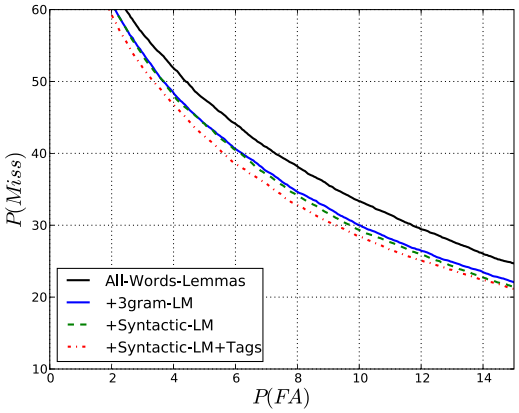

Figure 4: Features from a language model added to the best CRF from Section 6 (All-Words-Stemmed).

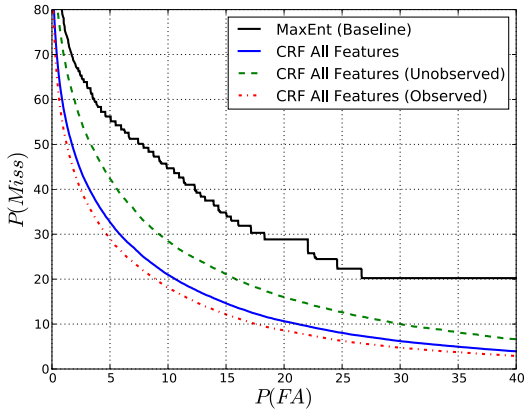

Figure 5: A CRF with all context features compared to the state-of-the-art MaxEnt baseline. Results for the CRF are shown for unobserved, observed and both OOVs.

We also used the standard trigram LM for reference. It was trained on the same data and with the same vocabulary using the SRILM toolkit. We used interpolated modified KN discounting.

7.2 Language Model Features

We designed features based on the entire utterance using the language model to measure how the utterance is effected by the current token: whether the utterance is more likely given the recognized word or some OOV word.

$$\text{Likelihood-ratio} = \log \frac{p(utt)}{p(utt|w_i = \text{unknown})}$$

$$\text{Norm-LM-score} = \frac{\log p(utt)}{length(utt)}$$

where $p(utt)$ represents the probability of the utterance using the best path hypothesis word of the LVCSR system, and $p(utt|w_i = \text{unknown})$ is the probability of the entire utterance with the current word in the LVCSR output replaced by the token <unk>, used to represent OOVs. Intuitively, when an OOV word is recognized as an IV word, the fluency of the utterance is disrupted, especially if the IV is a function word. The Likelihood-ratio is designed to show whether the utterance is more fluent (more likely) if the current word is a misrecognized OOV. [9] The second feature (Norm-LM-score) is the normalized likelihood of the utterance. An unlikely utterance biases the system to predicting OOVs.

We evaluated a CRF with these features and all lexical context features (Section 6) using both the trigram model and the joint syntactic language model (Figure 4). Each model improved performance, but the syntactic model provided the largest improvement. At 10% false alarm rate it yields a 4% absolute improvement with respect to the previous best result (All-Words-Stemmed) and 13.3% over the MaxEnt baseline. Higher order language models did not improve.

7.3 Additional Syntactic Features

We explored other syntactic features; the most effective was the 5-tag window of POS tags of the best hypothesis.[10] The additive improvement of this feature is depicted in Figure 4 labeled "+Syntactic-LM+Tags." With this feature, we achieve a small additional gain. We tried other syntactic features without added benefit, such as the most likely POS tag for <unk>in the utterance.

[9] Note that in the standard n-gram LM the feature reduces to $\log \frac{\prod_{k=i}^{i+n-1} p(w_k|w_{k-n+1}^{k-1})}{\prod_{k=i}^{i+n-1} p(w_k|w_{k-n+1}^{k-1}, w_i=\text{unknown})}$, i.e., only n n-grams actually contribute. However, in the syntactic LM, the entire utterance is affected by the change of one word through the latent states (tags) (Eq. 1), thus making it a truly global feature.

[10] The POS tags were generated by the same syntactic LM (see Section 7.1) as described in (Filimonov and Harper, 2009b). In this case, POS tags include merged tags, i.e., the vocabulary word *fred's* may be tagged as NNP-POS or NNP-VBZ.

8 Final System

Figure 5 summarizes all of the context features in a single second order BIO encoded CRF. Results are shown for state-of-the-art MaxEnt (Rastrow et al., 2009a) as well as for the CRF on unobserved, observed and combined OOVs. For unobserved OOVs our final system achieves a 14.2% absolute improvement at 10% FA rate. The absolute improvement on all OOVs was 23.7%. This result includes *observed* OOVs: words that are OOV for the LVCSR but are encountered in the OOV detector's training data. MaxEnt achieved similar performance for observed and unobserved OOVs so we only include a single combined result.

Note that the MaxEnt curve flattens at 26% false alarms, while the CRF continues to decrease. The elbow in the MaxEnt curve corresponds to the probability threshold at which no other labeled OOV region has a non-zero OOV score (regions with zero entropy and no fragments). In this case, the CRF model can still rely on the context to predict a non-zero OOV score. This continued decrease in the DET curve helps applications where misses are more heavily penalized than false alarms.

9 Related Work

Most approaches to OOV detection in speech can be categorized as filler models or confidence estimation models. Filler models vary in three dimensions: 1) The type of filler units used: variable-length phoneme units (as the baseline system) vs joint letter sound sub-words; 2) Method used to derive units: data-driven (Bazzi and Glass, 2001) or linguistically motivated (Choueiter, 2009); 3) The method for incorporating the LVCSR system: hierarchical (Bazzi, 2002) or flat models (Bisani and Ney, 2005). Our approach can be integrated with any of these systems.

We have shown that combining the presence of sub-word units with other measures of confidence can provided significant improvements, and other proposed local confidence measures could be included in our system as well. Lin et al. (2007) uses joint word/phone lattice alignments and classifies high local miss-alignment regions as OOVs. Hazen and Bazzi (2001) combines filler models with word confidence scores, such as the minimum normalized log-likelihood acoustic model score for a word and, the fraction of the N-best utterance hypotheses in which a hypothesized word appears.

Limited contextual information has been previously exploited (although maintaining independence assumptions on the labels). Burget et al. (2008) used a neural-network (NN) phone-posterior estimator as a feature for OOV detection. The network is fed with posterior probabilities from weakly-constrained (phonetic-based) and strongly-constrained (word-based) recognizers. Their system estimates frame-based scores, and interestingly, they report large improvements when using temporal context in the NN input. This context is quite limited; it refers to posterior scores from one frame on each side. Other features are considered and combined using a MaxEnt model. They attribute this gain to sampling from neighboring phonemes. Sun et al. (2001) combines a filler-based model with a confidence approach by using several acoustic features along with context based features, such as whether the next word is a filler, acoustic confidence features for next word, number of fillers in the whole sentence, etc.

None of these approaches consider OOV detection as a sequence labeling problem. The work of Liu et al. (2005) is most similar to the approach presented here, but applies a CRF to sentence boundary detection. Their features are similar to the ones used in this study, and include: n-grams of words or various tags (POS tags, automatically induced classes), prosody information, estimated posterior event probabilities using an n-gram LM.

10 Conclusion and Future Work

We have presented a novel and effective approach to improve OOV detection in the output confusion networks of a LVCSR system. Local and global contextual information is integrated with sub-word posterior probabilities obtained from a hybrid LVCSR system in a CRF to detect OOV regions effectively. At a 10% FA rate, we reduce the missed OOV rate from 42.6% to 28.4%, a 33.3% relative error reduction. Our future work will focus on additional features from the recognizer aside from the single best-hypothesis, as well as other applications of contextual sequence prediction to speech tasks.

References

Issam Bazzi and James Glass. 2001. Learning units for domain-independent out-of-vocabulary word modelling. In *Eurospeech*.

Issam Bazzi. 2002. *Modelling out-of-vocabulary words for robust speech recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

M. Bisani and H. Ney. 2005. Open vocabulary speech recognition with flag hybrid models. In *INTERSPEECH*.

L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Cernocky. 2008. Combination of strongly and weakly constrained recognizers for reliable detection of OOVS. In *ICASSP*.

Dogan Can, Erica Cooper, Abhinav Sethy, Chris White, Bhuvana Ramabhadran, and Murat Saraclar. 2009. Effect of pronounciations on OOV queries in spoken term detection. *ICASSP*.

Stanley F. Chen. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *Eurospeech*.

G. Choueiter. 2009. *Linguistically-motivated subword modeling with applications to speech recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

Denis Filimonov and Mary Harper. 2009a. A joint language model with fine-grain syntactic tags. In *EMNLP*.

Denis Filimonov and Mary Harper. 2009b. Measuring tagging performance of a joint language model. In *Proceedings of the Interspeech 2009*.

Jonathan Fiscus, John Garofolo, Mark Przybocki, William Fisher, and David Pallett, 1998. *1997 English Broadcast News Speech (HUB4)*. Linguistic Data Consortium, Philadelphia.

John Garofolo, Jonathan Fiscus, William Fisher, and David Pallett, 1996. *CSR-IV HUB4*. Linguistic Data Consortium, Philadelphia.

Timothy J. Hazen and Issam Bazzi. 2001. A comparison and combination of methods for OOV word detection and word confidence scoring. In *Proceedings of the International Conference on Acoustics*.

Zhongqiang Huang and Mary Harper. 2009. Self-Training PCFG grammars with latent annotations across languages. In *EMNLP*.

Dietrich Klakow, Georg Rose, and Xavier Aubert. 1999. OOV-detection in large vocabulary system using automatically defined word-fragments as fillers. In *Eurospeech*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

Hui Lin, J. Bilmes, D. Vergyri, and K. Kirchhoff. 2007. OOV detection by joint word/phone lattice alignment. In *ASRU*, pages 478–483, Dec.

Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using conditional random fields for sentence boundary detection in speech. In *ACL*.

Jonathan Mamou, Bhuvana Ramabhadran, and Olivier Siohan. 2007. Vocabulary independent spoken term detection. In *SIGIR*.

L. Mangu, E. Brill, and A. Stolcke. 1999. Finding consensus among words. In *Eurospeech*.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocky. 1997. The DET curve in assessment of detection task performance. In *Eurospeech*.

Andrew McCallum. 2002. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Carolina Parada, Abhinav Sethy, and Bhuvana Ramabhadran. 2009. Query-by-example spoken term detection for OOV terms. In *ASRU*.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of english words. In *ACL*.

Ariya Rastrow, Abhinav Sethy, and Bhuvana Ramabhadran. 2009a. A new method for OOV detection using hybrid word/fragment system. *ICASSP*.

Ariya Rastrow, Abhinav Sethy, Bhuvana Ramabhadran, and Fred Jelinek. 2009b. Towards using hybrid, word, and fragment units for vocabulary independent LVCSR systems. *INTERSPEECH*.

T. Schaaf. 2001. Detection of OOV words using generalized word models and a semantic class language model. In *Eurospeech*.

O. Siohan and M. Bacchiani. 2005. Fast vocabulary-independent audio search using path-based graph indexing. In *INTERSPEECH*.

H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig. 2005. The IBM 2004 conversational telephony system for rich transcription. In *ICASSP*.

H. Sun, G. Zhang, f. Zheng, and M. Xu. 2001. Using word confidence measure for OOV words detection in a spontaneous spoken dialog system. In *Eurospeech*.

Stanley Wang. 2009. Using graphone models in automatic speech recognition. Master's thesis, Massachusetts Institute of Technology.

F. Wessel, R. Schluter, K. Macherey, and H. Ney. 2001. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3).

Christopher White, Jasha Droppo, Alex Acero, and Julian Odell. 2007. Maximum entropy confidence estimation for speech recognition. In *ICASSP*.