INTERSPEECH 2009 BRIGHTON

Combined low level and high level features for Out-Of-Vocabulary Word detection

Benjamin Lecouteux[1], *Georges Linarès*[1], *Benoit Favre*[2]

[1]Laboratoire Informatique d'Avignon (LIA), University of Avignon , France
[2] ICSI, 1947 Center St, Suite 600, Berkeley, CA 94704, USA
{benjamin.lecouteux,georges.linares}@univ-avignon.fr; favre@icsi.berkeley.edu

Abstract

This paper addresses the issue of Out-Of-Vocabulary (OOV) word detection in Large Vocabulary Continuous Speech Recognition (LVCSR) systems. We propose a method inspired by confidence measures, that consists in analyzing the recognition system outputs in order to automatically detect errors due to OOV words. This method combines various features based on acoustic, linguistic, decoding graph and semantics. We evaluate separately each feature and we estimate their complementarity. Experiments are conducted on a large French broadcast news corpus from the ESTER evaluation campaign. Results show good performance in real conditions: the method obtains an OOV word detection rate of 43%-90% with 2.5%-17.5% of false detection.

Index Terms: OOV word detection, confidence measures, speech recognition

1. Introduction

Large Vocabulary Continuous Speech Recognition (LVCSR) systems are restricted by the limited size of the lexicon, since Out-Of-Vocabulary (OOV) words are frequently meaningful, and may be potentially critical for speech mining or indexing systems. The increase of lexical coverage by an extensive growth of the size of the vocabulary is not suitable because it introduces noise and it requires additional computing resources for handling huge vocabularies. Some authors proposed domain-specific methods for augmenting the lexicon. These approaches generally rely on a semantic analysis of the spoken content. [1] proposed to use local context to build web-queries allowing to retrieve missing words.

One of the main difficulty in designing efficient methods for OOV word retrieval is to automatically detect them. Most of previous work addresses this problem by integrating fillers in the language models that are supposed to absorb unexpected speech segments. These methods demonstrated good performances on small or medium vocabulary tasks but the integration of filler models requires a fine tuning of the language models, especially on large vocabulary tasks. Others proposed *a posteriori* approaches, where the intermediate Automatic Speech Recognition (ASR) system output is analyzed in order to locate the transcript areas where errors are due to OOV words. In this paper, we present such an *a posteriori* method that aims at detecting OOVs by analyzing the output of a first recognition pass. We propose various word level features that are independently evaluated and combined in a statistical classifier. Then, each word is classified as OOV or non-OOV.

In the next Section, we draw an overview of related work about OOV word detection. In Section 3, we present our method, which is derived from confidence measure estimation approaches. We focus on specific aspects of OOV detection in comparison with classical confidence measures.

In Section 4 we define the experimental framework. We describe the ESTER corpus on which experiments are conducted and we detail the experimental protocol.

In Section 5, we propose various features for OOV detection. We evaluate their relevance and their complementarity for that task. We also study the impact of Word Error Rate (WER) on our detection methods. Finally we present results on the whole test corpus in Section 5. The last Section presents conclusions and prospects.

This research is supported by the ANR (Agence Nationale de la Recherche), AVISON project.

2. Related work

The detection of OOV words has been explored in several ways; some research groups [2, 3, 4, 5] proposed to model unseen words through lexical fillers or generic word models: the goal is to cover all OOV word pronunciations, by representing them with sub-word units. Moreover, the method [5] allows to retrieve the OOV word using graphonemic sequence models. Unfortunately, these methods often absorb parts of the speech corresponding to known words, and require to be finely tuned.

High-level information (syntactical, semantic) is used for both OOV detection and confidence measure estimation. [6] proposes to detect OOV words in a spoken dialog system. Experiments are conducted in restricted domains, but show the benefits of distant context for detecting named entities. [7] proposes to use Latent Semantic Analysis (LSA) for estimating confidence measures in LVCSR systems; results show a good accuracy but relatively low recall rates.

Other works [8, 9, 10] use, in different ways, the edit distance between the phoneme lattice and the decoded words, misalignment being supposed to be more frequent in OOV words.

These methods yield good accuracy and true detection rate, but do not take directly into account semantic features or graph topology.

In other articles [11, 12, 13] OOV words are identified by using confidence measures. These approaches allow one to introduce more information sources extracted from the LVCSR system. These methods obtain better accuracy than the filler-model, but they are limited to spoken dialog systems or isolated word recognition. Moreover, they do not use robust linguistic features.

Recent approaches for confidence measure estimation use side-information extracted from the recognizer: the number of competitors at the end of a word, normalized likelihoods, linguistics features, decoding process behavior, etc. Some work addressing the use of confidence measures show the prominence of language models: [14] proposes to combine acoustic and linguistic features such as language model back-off behavior and posteriors.

3. Detecting OOV words

3.1. Principle

The proposed method consists of 3 stages. The first one extracts low level features related to acoustic and search graph topology, and high level features related to linguistics. Then, a first OOV detection hypothesis is produced by a classifier based on the boosting algorithm. Finally, a semantic module refines this detection process. The next subsections detail this 3 steps detection method.

3.2. Extracted features

Each word from the hypothesis is represented by a feature vector composed of 23 features, that are grouped into 3 classes.

OOV words probably induce acoustic distortions between the hypothesis and the best phonetic sequence. We use **acoustic features** which consist of the acoustic log-likelihood of the word, the averaged log-likelihood per frame, the difference between the word log-likelihood and the unconstrained acoustic decoding of the corresponding speech segment.

Linguistic features are based on probabilities estimated by the 3-gram language model used in the ASR system. We use the 3-gram probability, the perplexity of the word in the window, and the unigram probability. We also add the index, proposed in [14], that represents the current back-off level of the targeted word. This value is set to 3 if the current trigram occurred in the training corpus, 2 if only the current bigram occurred, 1 otherwise.

Graph features are based on the analysis of the word confusion networks. The use of these features is motivated by the idea that when an OOV occurs, the search algorithm probably explores various alternative paths that are similarly scored. Backtracking behavior and the distribution of posterior probabilities could be a good index of unknown word detection. We use the number of alternative paths in the word section, and the posterior probability. We also include values related to the distribution of posteriors probabilities in the word section: the 2 extremum values, the mean of the posteriors from the sausage section, the number of null links before and after the current section. A last parameter represents the mean duration of words, that is estimated in a 500ms window.

3.3. Feature combination and classification

We use a boosting classification algorithm in order to combine word features, as detailed in [15]. The classifier is a variant of Adaptive Boosting (Adaboost): icsiboost [1]. The algorithm consists in an exhaustive search for a linear combination of classifiers by overweighting misclassified examples. An advantage of this classifier is its ability to provide probability-like scores, which allows for an intuitive interpretation of the results.

Input vectors are estimated on a window composed by 3 consecutive words. On each word, we estimate the 23 features previously described. We finally obtain a 69 coefficient vector including previous, current and next word descriptors.

The classifier is trained on a specific training corpus, that was not included in the ASR system training. Each word from this corpus is tagged as *in* or *out* of vocabulary, according to the ASR system lexicon. The classification results in two classes for each word: OOV and non-OOV.

3.4. Latent Semantic Analysis and web 3-grams

Latent Semantic Analysis is a technique that allows to associate words that tend to co-occur within documents with a semantic relation. The assumption is that co-occurrent words within the same document are semantically linked.

[1] available on http://code.google.com/p/icsiboost/

In our system, a semantically consistent word sequence may be considered as unexpected by the ASR language model due to the limits of the n-gram language models. Therefore, we estimate an index of semantic consistency (SC) that allows to validate (or to reject) OOV detections previously performed by the classifier. This measure is not included in the classifier because of the introduced noise in non-OOV words. It is only applied on detected words, as a last filtering process.

This filter combines two measures based on the web and the French Gigaword corpus. The first estimates the probability of word co-occurences as a ratio of Google hits [1]. This measure is computed on a local window of 3 words (including the targeted one) filtered by a stop list. The second uses Latent Semantic Analysis (LSA) to estimate how much the targeted word is semantically close to the current segment [7]: for each targeted word, the LSA module selects the 100 closest words. The cardinality of the intersection of this bag of words and the current segment is normalized by the segment size, and the resulting value is linearly combined with Google-based scores. The decision threshold has been tuned on the development corpus to obtain a low recall with 100% accuracy: on the developement, 9% of false detection are then deleted.

4. Experimental framework

4.1. The LIA broadcast news system

Experiments are carried out by using the LIA broadcast news (BN) system which was used in the ESTER evaluation campaign [16]. This system relies on the HMM-based decoder developed at LIA, Speeral [17]. Speeral is an asynchronous decoder operating on a phoneme lattice; acoustic models are HMM-based, context dependent with cross word triphones. The language models are classical 3-grams estimated on about 200M words from the French newspaper *Le Monde* and from the ESTER broadcast news corpus (about 1M words). The lexicon contains 67K words. In these experiments, only one decoding pass is performed in 3x Real Time (RT).

4.2. The ESTER corpus

The ESTER corpus consists of French radio broadcasts of the Radio-France group. The training and development parts of the data set are based on the training corpus provided for the ESTER-2 (100 hours manually annotated) evaluation campaign, these data were not included in the ASR system training. The training corpus contains 15K in-vocabulary words randomly selected among 1M words and 15K OOV words extracted from the whole training corpora.

We test our approach on 7 hours of speech extracted from the ESTER test set. The number of OOV words is 982, which, since the corpus contains 70,011 words, represents an OOV rate of about 1.4%. The mean WER is 28.9: Only the first pass of decoding is performed.

4.3. Detection protocol

OOV words have been manually specified by selecting all the reference words not available in the lexicon. During the detection, if a marked OOV word overlaps with a true OOV word, the true OOV word is considered as detected. In all other cases we consider a marked word as a false detection.

5. Experiments

We performed three experiments to estimate the relevance of the features and their complementarity. Moreover, we estimate the system robustness by evaluating detection rates on a range of WER.

5.1. Relevance of the features

5.1.1. *Classification depending on each parameter*

In this Section, we draw an overview of the classification ability for each set of features. A model is trained with icsiboost for each of those sets. We use Receiver Operating Characteristic (ROC) curves to display OOV true and false detection rates. Curves are traced by varying the threshold on classifier outputs.

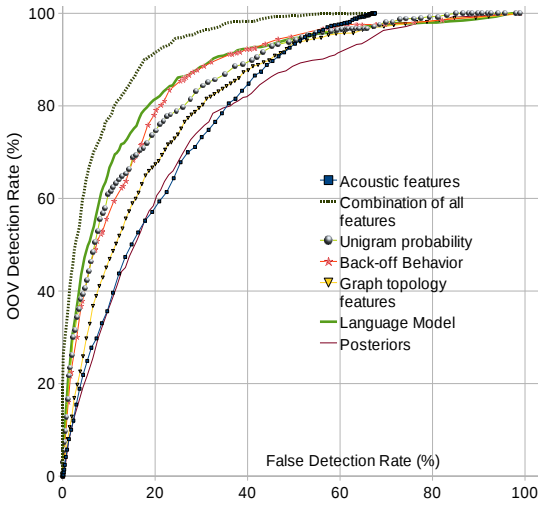


Figure 1: *Relevance of each feature*

The ROC curves, shown in Figure 1, present the output of the classifier for each separate feature compared to the combination of all features. Results show that the most relevant descriptors are the linguistic and graph-topology features, while acoustic features and posteriors seem weaker in comparison to others. Linguistic scores and language model back-off behavior are predictors of linguistic model failure (repeated fallback to unigram back-off). This failure is characteristic of incoherences due to unknown word or recognition errors. An interesting thing is the quality of the language model back-off behavior, which presents results close to language model probabilities.

Graph topology is also a interesting feature, because it is related to recognition process difficulties. A surprising result is the high discriminative capacity of the unigram probability. This may be due to the fact that the system tends to back-off, when unknown words are encountered, to small but infrequent words that acoustically match the signal. However, the combination of all the features shows a significant improvement. In spite of the feature heterogeneity, we observe a relatively high complementary.

5.1.2. *Linguistic context of OOV words*

The experiments reported in the previous Section show the relevance of the OOV word contexts (Figure 2). Our topology includes the previous and next word for each spotted word. An unexpected result is that the previous and next word seem to be more informative than the spotted word with a prevalence for the previous word. The previous word is probably more relevant, concerning linguistic model failure, than the middle and next words: we must not forget that the features of the previous word hold linguistic information about the middle word.

Yet, the combination of the three sets of feature of the words leads to a 10% relative improvement, whereas the use of a window larger than 3 words leads to worse results.

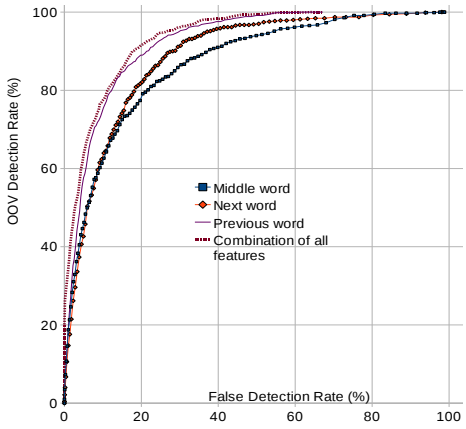
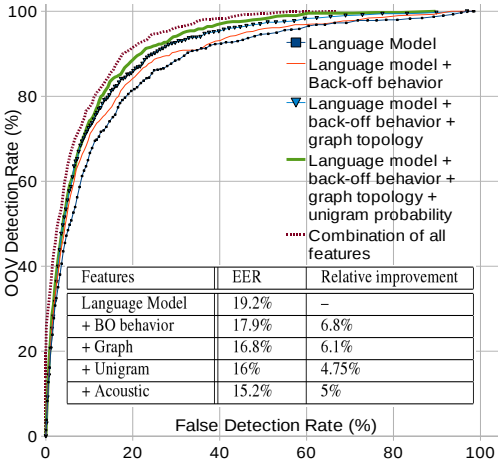


Figure 2: *Relevance of the words in an observation window*

5.2. Complementarity of features

After the overview of the feature relevance, we study their complementarity. The baseline curve uses the best observed set of features in the previous Section. Then, we enable each new set of features according to their relevance, from the most relevant to the least relevant.



Features	EER	Relative improvement
Language Model	19.2%	–
+ BO behavior	17.9%	6.8%
+ Graph	16.8%	6.1%
+ Unigram	16%	4.75%
+ Acoustic	15.2%	5%

Figure 3: *Complementarity of the features*

The ROC curves (Figure 3) show the complementarity of each set of features.

We measure the complementarity by using the Equal Error Rate (EER). The EER is the point where OOV word detection rate and false detection rate are equal. In the table of Figure 3 we show the EER evolution for each added set of features.

Each set provides some discriminative information. An unexpected result is the prominence of acoustic features. They show a significant improvement, while it exhibits the worst performance when other features are not used. However, the classifier allows one to relate different acoustic features. Sets of features are therefore complementary and mutually dependent.

5.3. OOV words and WER

These last experiments examine the behavior of OOV detection according to the WER obtained in transcription. We have sorted the test segments in 6 ranges of WER (Figure 4).

In Figure 4 we show results for the different ranges of WER. We can remark a correlation between WER and OOV rate. ROC

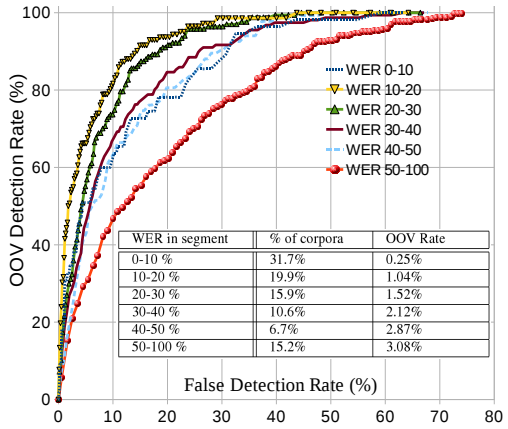


Figure 4: *Results for each WER range*

curves show three trends: an excellent classification for WER between 10% and 30%, a stable classification for WER between 0% and 10% or 30% and 50%, and a degradation beyond 50%.

The 0%-10% range is surprising because it is far from the best ranges. It corresponds to speech segments with very low OOV rate (0.25%), that may increase the detection difficulty: high lexical coverage and good WER indicates acoustic and linguistic contexts probably close to the training conditions, on which our features may be less informative.

However, between 10% and 30%, the classification rates are greater than the ones observed on other ranges. It probably corresponds to the most frequent decoding context, on which the test condition matches the training conditions and where features are significants. Beyond 50%, noise decreases the accuracy. Globally, experiments show a relatively good robustness against WER, except when the ASR system dramatically fails.

5.4. Semantic Filtering of detected OOV words

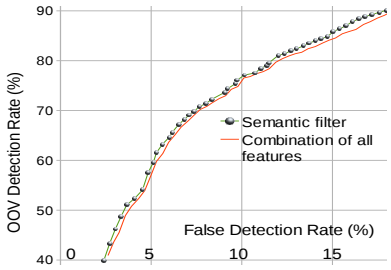


Figure 5: *Results after the semantic filter*

In the final pass, we propose to use semantic information for filtering the detected OOV words. We use the semantic module only on detected OOV words, with the purpose to refine detection on semantically coherent words. The ROC curves are presented in Figure 5. Results show that the filter reduces the EER by 4% relatively (15.2% to 14.6%), while the false detection rate decreases by about 5% relative.

6. Conclusion and future work

In this paper we presented a method for the detection of OOV words in Large Vocabulary Continuous Speech Recognition systems. Our approach consists in extracting multiple features during the speech recognition process: acoustic features, linguistic features, topology graph features. Finally, a semantic module performs a last filtering pass to reduce the false detection rates.

Our experiments showed promising results: OOV word detection seems to be homogeneous, despite the inconstant WER. The proposed method allows one to detect 43% of the unknown words with a 2.5% false acceptance rate, or 90% for 17.5% false acceptance. Experiments show that linguistic and graph-based features are the most relevant predictors. However, acoustic features associated to the others make the detection more robust. Finally, semantic filtering provides a slight but significant improvement.

Presently, these results rely only on the one-best word hypothesis; we plan to extend the semantic module in order to retrieve alternative paths in the graph, for invalidating false OOV word detections. We wish to add more semantic analysis, such as the automatic detection of named entities that usually compose a large part of unknown words.

7. References

[1] S. Oger, G. Linarès, and F. Béchet, "Local methods for on-demand out-of-vocabulary word retrieval," in *LREC*, 2008.

[2] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic detection of new words in a large vocabulary continuous speech recognition system," in *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP-90*, 1990, pp. 125–128 vol.1.

[3] I. Bazzi and J. R. Glass, "Modeling out-of-vocabularywords for robust speech recognition," in *ICSLP 2000*, 2000.

[4] G. Boulianne and P. Dumouchel, "Out-of-vocabulary word modeling using multiple lexical fillers," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding ASRU '01*, 2001, pp. 226–229.

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

[6] M. Boros, M. Aretoulaki, F. Gallwitz, E. Noth, and H. Niemann, "Semantic processing of out-of-vocabulary words in a spoken dialogue system," in *Eurospeech*, 1997.

[7] S. Cox and S. Dasmahapatra, "A semantically-based confidence measure for speech recognition," in *ICSLP*, 2000.

[8] C. White, G. Zweig, L. Burget, P. Schwarz, and H. Hermansky, "Confidence estimation, oov detection and language id using phone-to-word transduction and phone-level alignments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008*, 2008, pp. 4085–4088.

[9] H. Lin, J. Bilmes, D. Vergyri, and K. Kirchhoff, "Oov detection by joint word/phone lattice alignment," in *Proc. ASRU Automatic Speech Recognition & Understanding IEEE Workshop on*, 2007, pp. 478–483.

[10] S. Hayamizu, K. Itou, and K. Tanaka, "Detection of unknown words in large vocabulary speech recognition," in *Eurospeech*, 1993.

[11] H. Sun, G. Zhang, F. Zheng, and M. Xu, "Using word confidence measure for oov words detection in a spontaneous spoken dialog system," in *Eurospeech*, 2003.

[12] T. Cai and J. Zhu, "Oov rejection algorithm based on class-fusion support vector machine for speech recognition," in *Proc. International Conference on Machine Learning and Cybernetics*, vol. 6, 26–29 Aug. 2004, pp. 3695–3699.

[13] S. R. Young, "Recognition confidence measures:detection of misrecognitions and out-of-vocabulary words," Carnegie Mellon University, Tech. Rep., 1994.

[14] J. Mauclair, Y. Estève, S. Petit-Renaud, and P. Deléglise, "Automatic detection of well recognizedwords in automatic speech transcriptions," in *LREC*, 2006.

[15] P. J. Moreno, B. Logan, and B. Raj, "A boosting approach fo confidence scoring," in *Europseech*, 2001.

[16] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news," in *Eurospeech*, 2005.

[17] P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, and F. Béchet, "The lia's french broadcast news transcription system," in *SWIM: Lectures by Masters in Speech Processing*, 2004.