Lexical Out-of-Vocabulary Models for One-Stage Speech Interpretation

Matthias Thomae, Tibor Fabian, Robert Lieb, Günther Ruske

Institute for Human-Machine Communication
Technische Universität München, Germany
eurospeech05@thomae-privat.de, {fab,lie,rus}@mmk.ei.tum.de

Abstract

We present an approach to explicit, statistical, lexical-level out-of-vocabulary (OOV) word modeling for direct integration into the search space of a one-stage speech interpretation system. For this purpose, a generic pronunciation model for unknown words is derived from large pronunciation lexica and, optionally, word frequency knowledge. Known statistical language modeling (LM) methods are utilized to estimate different phoneme LM and apply different smoothing techniques. The resulting OOV word models are integrated with the hierarchical language model of our uniform modeling framework by declaring semantically irrelevant parts of the training utterances as unknown. Experiments were conducted with two different OOV training lexica on an airport information dialogue application, evaluating the results with both in-vocabulary (IV) and OOV-related metrics. Results for various OOV model configurations are presented, showing that OOV detection rates of 60-70% can be achieved with 1-2% falsely accepted IV words, simultaneously improving accuracy on the semantic representation.

1. Introduction

In closed-vocabulary speech recognition and understanding systems, out-of-vocabulary (OOV) words inevitably cause recognition errors because they are silently misrecognized as in-vocabulary (IV) words. Even worse, surrounding words are often also affected because of erroneous segmentation of the misrecognized OOV word. In experiments reported in [1], an OOV word caused about two misrecognized words on average. Hence, the suitability of a closed-vocabulary approach largely depends on the question if the OOV rates of the target application and the caused errors are acceptable. Due to the difficulty of the task of speech understanding, most of these systems today typically operate in narrow application domains with limited vocabulary size. In these circumstances, high rates of OOV words must be expected if users talk freely to the system. This can be the case even if great effort is taken to make all semantically relevant lexical items known to the system, if users are oblivious to domain limitations and many out-of-domain utterances occur. Moreover, the ability to detect OOV words can be vital for a spoken dialogue system in order to purposefully ask the user to rephrase parts of his utterance, instead of continuing with the misrecognized words.

In [2] we introduced our One-stage Decoder for Interpretation of Natural Speech (ODINS), which tightly integrates automatic speech recognition and natural language understanding techniques in a one-stage decoding process. As we expect high rates of unknown words for the target application fields, ODINS should be able to detect OOV words and also avoid errors at surrounding words. Therefore, explicit knowledge of OOV words should be integrated directly into the decoding process of ODINS. ODINS decodes meaning representations directly from speech with the aid of weighted transition network hierarchies (WTNH). WTNH unify acoustic-phonetic, lexical and syntactic-semantic modeling in a single modeling framework. Hence, we need to convert explicit OOV models to weighted transition networks in order to represent them within WTNH.

The explicit OOV model can be built on the acoustic level or on the lexical level. For an acoustic-level OOV model as in [1], one or several HMM models for OOV words are trained, so that a considerable amount of acoustic training data for OOV words needs to be available as an additional knowledge source. For a lexical OOV model, existing phoneme HMM are combined in a suitable way, e.g. by deriving a pronunciation model from large pronunciation lexica. A lexical OOV modeling approach similar to [3] was considered most suitable for our task. Firstly, because the OOV words occurring in the limited-domain applications for ODINS are mostly common words of the target language, and therefore are not expected to have substantially different acoustic-phonetic properties than the modeled IV words. Secondly, because large pronunciation lexica are meanwhile available for many languages, whereas it can be more difficult to obtain the knowledge source for acoustic-phonetic OOV modeling, i.e. acoustic training data. As discussed in [1], the main drawbacks of lexical OOV modeling are a tendency for over-generation and the resulting need for OOV model penalization, and high computational requirements. Whereas we address the former issue by examining the sensitivity of OOV models against penalty variations, analysis of run-time performance was not a focus of this work. Yet, some results on OOV model accuracy for reduced OOV model sizes are reported.

This work is structured as follows: Different methods to estimate lexical OOV word models from large pronunciation lexica are presented in Section 2. The integration of the resulting OOV models with our hierarchical language model [4] is discussed in Section 3. Section 4 presents suitable evaluation metrics to describe the effects of OOV modeling on system performance. In Section 5, experimental results for various system configurations are presented.

2. Phoneme Language Models

Similar to [3], we consider OOV words by a generic pronunciation model for arbitrary words. We use two main knowledge sources for model estimation, namely pronunciation lexica and word frequency lists. From these knowledge sources, statistical 'language models' (LM) of phoneme sequences for word pronunciations can be generated using standard techniques. We

This work was funded partly by the German Research Council (DFG) project Ru 301/6-2.

examine two different types of phoneme LM, namely n-gram LM (as in [3]) and so-called exact LM. In an n-gram LM, only the previous $n-1$ symbols are considered to determine the likelihood of the current symbol. Through this, an n-gram LM is able to cover arbitrary-length sequences, and to generalize from the data 'seen' during training to new, 'unseen' symbol sequences. This especially includes cutoff words which can occur frequently in natural speech. In contrast, exact LM exactly cover the symbol sequences seen during training, i.e. they are not able to generalize.

Typically, the basic LM weights are maximum-likelihood estimates, i.e. normalized counts of events from a training corpus. In data sparsity, these counts are assumed to be unreliable, and hence smoothed by reducing and re-distributing probability mass. A comparative study of a number of smoothing techniques for n-gram LM is given in [5]. Further generalization of n-gram LM is achieved by combination with lower-order n-grams through backoff. Here, we use 'canonical' Katz backoff smoothing, and modified Kneser-Ney smoothing as proposed in [5]. n-Gram LM and their transition network representations are computed with the SRILM Toolkit [6]. As context-dependent triphone HMM are used as acoustic models, phonemes must be traversed in the right order, i.e. with matching right and left contexts. In order to ensure this, unigram backoff is disallowed. We also examine removal of higher-order backoff from phoneme n-gram LM. Furthermore, we tested if discrimination between OOV and IV words can be improved by excluding IV words from the OOV training lexicon.

For exact LM, additive discounting and Good-Turing discounting (see e.g. [5]) are applied directly on the network level, i.e. we use network transitions as the basic events. Transition networks of exact LM are generated by representing the training phoneme sequences as a list of regular expressions, which is then compiled into a finite-state automaton and minimized by use of the Lextools and FSM Library toolkits [7, 8].

The German Phonolex [9] and Celex [10] pronunciation lexica were utilized as knowledge sources for phoneme LM estimation. From Phonolex, the manually checked 'core' pronunciations of 22k inflected words were used. As the phoneme sets of Phonolex and ODINS are both similar to the Verbmobil [11] definitions, only few phoneme mappings had to be performed. The German Celex database contains 52k lemmata with 366k corresponding wordforms. Phoneme set adaptation was more difficult than for Phonolex. Wordforms containing unknown phonemes were removed, leaving 314k pronunciations for OOV model estimation. Celex also contains word frequency information from different sources including spontaneous speech transcriptions. This was optionally used to weight the pronunciations accordingly. Weighting is performed by estimating phoneme LM from a modified pronunciation list, which contains as many copies of each word pronunciation as the word frequency value suggests.

3. Integration with HLM

Currently, our speech interpretation approach utilizes no explicit syntactic or morphologic knowledge, but the semantic analysis is performed directly on the word level with the aid of the so-called hierarchical language model (HLM) [4]. Therefore, the OOV word model needs to be integrated with the HLM, i.e. it must be defined at what positions OOV words may occur in utterances and which likelihood is assigned to them. For this purpose, we augment the existing symbol set with a new word symbol 'OOV'. Then, we add OOV word annotations to the speech corpus by declaring some of the previously known words as unknown, replacing their word symbol by 'OOV'. We denote these words as *known OOV words*, in contrast to the 'real' unknown words that never occur in the training set. After these steps, the HLM building process can be applied as before [4]. Later on, the generated OOV pronunciation model is added to the WTNH.

In general, candidates for known OOV words are all semantically irrelevant words within the given domain. It may nevertheless be desirable to keep some of the semantically irrelevant words in the vocabulary, because they have syntactic relevance or because their effect on confusability is larger as OOV word than as IV word. In this work, we examine two different sets of OOV word annotations on a speech corpus, yielding OOV rates of 23% and 8% on the evaluation data. In the 23%-set, all surface words (not contained in a word class or semantic concept) are declared unknown, whereas in the 8%-set, some of the most frequent syntactically relevant surface words are not declared unknown. Hence, most known OOV words appear between semantic categories, only few known OOV words occur within them. Purposefully integrating more OOV words into semantic concepts may further increase speech interpretation robustness, e.g. in order to explicitly provide for unknown word class members, but was not examined here.

Please note that we combine consecutive OOV words into a single OOV symbol. Through this, a more robust estimation of OOV model likelihood within HLM can be achieved. Furthermore, we require the OOV model to correctly detect OOV word occurrences, but not to correctly detect *how many* OOV words were uttered consecutively, and how to segment them. This simplification is performed as even humans have difficulties to correctly segment sequences of spoken words that are unknown to them. Consequently, consecutive OOV word sequences are used as the basic unit for evaluation of the OOV pronunciation model's performance. For simplicity, we still denote those sequences as OOV words in the following. In order to balance the relative weighting and thereby occurrence likelihood of IV and OOV words, penalization of OOV models is examined with two different parameters. The additive log-likelihood offset p_{oov}^{in} is applied at entering the OOV word model. The log-likelihood scaling factor λ_{oov} is applied multiplicatively to all transition scores within the OOV network, i.e. after p_{oov}^{in}.

4. Evaluation Metrics

One notion of the OOV problem is the task of detecting OOV words within IV word sequences. In this context, the detection of an OOV word is denoted as acceptance (of the OOV word), whereas an IV word is viewed as rejection. Detection tasks are evaluated by creating a *receiver-operating-characteristic* (ROC) from various operating points. Such a diagram shows the two types of errors that the OOV word detector can make, namely false acceptances (FA) and false rejections (FR). FA denotes an IV word wrongly hypothesized as OOV word, FR an OOV word wrongly hypothesized as IV word. The two types of correct operations are called correct acceptance (CA) and correct rejection (CR). The counts of FA, FR, CA and CR are computed from the mappings between reference transcriptions and hypotheses of a test set. As the evaluation of ODINS is based on the semantic tree matching scheme of [12], the OOV/IV word mappings are taken from the word level of the semantic tree match. In addition to mappings between OOV/IV words, the tree match contains insertions and deletions, i.e. an empty symbol ϵ in either reference or hypothesis. In order to

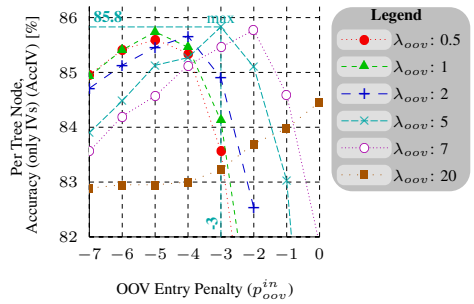

Figure 1: *Total tree node accuracy for different OOV penalties.*

consider these kinds of errors for OOV model evaluation, ϵ is treated as an IV symbol. Our ROC curves are plots of false acceptance rate (FAR) against false rejection rate (FRR), defined as:

$$FAR = \frac{FA}{CR + FA} \qquad FRR = \frac{FR}{CA + FR} \tag{1}$$

In order to capture the course of an ROC curve in a single metric, a figure-of-merit (FOM) is computed as the area over the curve. For a practical speech interpretation application, one usually tries to avoid that IV words are misinterpreted as OOV words, hence the FAR should be small. To focus the OOV model evaluation accordingly, the FOM is limited, similar to [3], to a region of interest by computing the area between $0\% \leq FAR \leq 5\%$. This area is normalized to yield a constant value range between 0 and 1, yielding the so-called 5%-FOM.

The detection of OOV words also affects the recognition of IV words, since one or several IV words (or parts of IV words) can be mistakenly recognized as OOV words or vice versa. Therefore, it is essential to include an evaluation of the IV words in an OOV model evaluation. Moreover, as we use the OOV model within a speech interpretation system, the whole semantic representation can be influenced, so that the semantic tree node accuracy Acc_n (as defined in [12]) must be taken into account instead of only the word accuracy. For a simultaneous analysis of both evaluation metrics, they are combined in a single diagram by using a common abscissa but two different ordinate axes. While the ROC curve (which only regards words) is plotted as usual with FAR on the abscissa and FRR on the ordinate, the tree node accuracy Acc_n (regarding all semantic symbols) is plotted on a second ordinate, which resides on the right border of the plot.

5. Experimental Results

Experiments were conducted on a corpus of spontaneous speech utterances collected by simulating an airport information dialogue system through a wizard-of-oz setup. The corpus is an extended version of the one used in [2, 12], containing about 2700 utterances with 15000 words from 32 subjects in total. HLM were trained on a subset of 20 subjects, evaluation and cross-validation were performed on 6 subjects' utterances each. HLM consisting of 4 hierarchy levels, namely words, word classes and 2 concept levels, were generated using a mixture of data-driven and rule-based LM techniques, as described in [4]. The acoustic models consist of the same speaker-independent tied intra-word triphone HMM with about 25k Gaussian mixture components as described in [2]. The word class level contains 10, the concept levels 40 unique symbols. As the speech corpus does not completely contain the word class contents relevant for the example application, the missing words (around 30) were added manually.

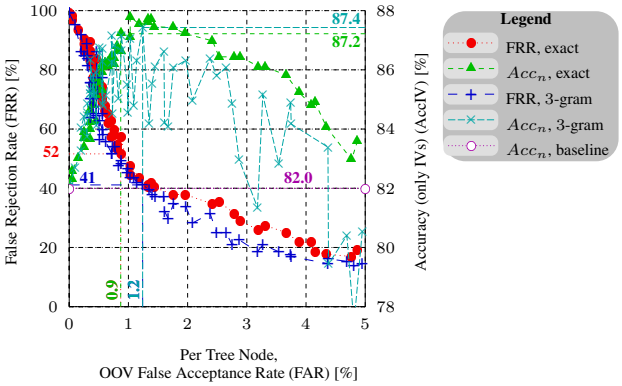

Figure 2: *Exact vs. n-gram phoneme language model.*

The 'real' OOV word rate is 2.2% and 1.5% on evaluation and cross-validation set. The word level of the baseline annotations, i.e. without explicit OOV word annotations, contains about 580 words. As explained in Section 3, two different sets of OOV word annotations are used to examine how the system reacts to varying OOV rates. For the so-called 8%-set, the known OOV word rates are $(8.6\%, 4.7\%, 7.2\%)$ on the training, cross-validation and evaluation sets, respectively. The second annotation set has $(25.2\%, 16.1\%, 21.6\%)$ known OOV words in training, cross-validation and evaluation, and is referred to as the 23%-set. From the 580 IV words of the baseline system without OOV modeling, 400 IV words remain for the 8%-set and 380 IV words for the 23%-set. The tested value ranges of OOV scaling and penalty parameters were always $p^{in}_{oov} = -7 \ldots 0$ and $\lambda_{oov} = 0.5 \ldots 20$ in order to ensure comparability. Figure 1 shows Acc_n in dependency of p^{in}_{oov} for different λ_{oov} with an additively smoothed exact OOV model on the cross-validation set. For this setup, the maximum accuracy of 85.8% is achieved at $(p^{in}_{oov}, \lambda_{oov}) = (-3, 5)$. After determining the penalty parameter setting yielding maximum accuracy for an OOV model on the cross-validation set, this setting is used to determine the accuracy on the evaluation set.

Figure 2 depicts a combined ROC-accuracy plot for two Phonolex-based OOV models using different phoneme LM techniques, namely exact and 3-gram LM. Please note that left and right y axes are scaled differently. Since cross-validation experiments showed that additive discounting yields slightly better results on exact phoneme LM than Good-Turing discounting, and that modified Kneser-Ney smoothing performs better than Katz smoothing on n-gram phoneme LM, we used those techniques for further experiments. The 3-gram OOV model of Figure 2 shows advantages in both maximum accuracy (87.4% vs. 87.2%) and 5%-FOM (66.8% vs. 61.5%) over the exact model. However, the accuracy curves reveal that the 3-gram model is rather sensitive to changes of the penalty parameters, whereas the exact model behaves more stable with respect to accuracy. Such stability differences were generally observed between exact and n-gram phoneme LM. The OOV detection performance of the 3-gram model consistently wins over the whole displayed range of the ROC curve, whereby the differences are rather small around the operating points with maximum accuracy ($FAR \approx 1\%$).

Table 1 summarizes evaluation set accuracies and figures-of-merit of a number of different OOV model configurations, along with the baseline accuracies of the closed-vocabulary sys-

tem. The results reveal that the baseline accuracies on both 8%- and 23%-sets can be outperformed clearly by both exact and n-gram phoneme LM. Although the 2%-system (i.e. the system without known OOV word annotations) seems to achieve marginally better accuracy than the best open-vocabulary systems (87.8% vs. 87.4%), comparison of these values is strictly speaking not possible due to the underlying differences in IV vocabulary size. Moreover, significant performance drops must be expected if effective OOV rates exceed 2%.

Furthermore, we tested the effects of word frequency weighting and exclusion of IV words on both exact and 3-gram phoneme LM for the 8%-set Phonolex configuration. In terms of accuracy, the models with word frequency weighting consistently outperform those without. For the exact phoneme LM, the 5%-FOM performs slightly better if no word frequency weighting is carried out. The exclusion of IV words consistently degrades both accuracy and FOM. At first, this seems surprising because it aims at clearer discriminating OOV and IV word models. Yet it can be explained with the fact that some of the known OOV words also appear as IV words within semantic concepts, so that their exclusion from the OOV training lexicon prevents their detection as OOV words. Furthermore, we built an 3-gram phoneme LM with 2-gram backoff (denoted as 3/2-gram in Table 1) and a 2-gram phoneme LM for the same Phonolex configuration. Both perform significantly worse than their 3-gram counterpart, but still well above the baseline.

For the 23%-set, there is also no clear winner between exact and 3-gram phoneme LM, as for the 8%-set of Figure 2. This time, the difference between 5%-FOM values is not that large, and the exact model displays better maximum accuracy. As expected, reducing the size of the OOV training lexicon by selecting most frequent words generally causes performance degradations, yet maximum accuracy is still well above the baseline values. The Celex-based OOV models display similar performance as the Phonolex-based ones, yet they generally require a larger number of pronunciations to achieve this. It can also be noted that the 314k-word models cannot clearly outperform the models estimated from a smaller number of pronunciations.

OOV lex.	OOV Model Type	OOV rate eval	OOV lex. size	word freq. weight	excl. IV	Acc_n	5%-FOM
none	-	2%	-	-	-	87.8%	-
none	-	8%	-	-	-	**82.0%**	-
P	exact	8%	22k	no	no	87.0%	61.9%
P	exact	8%	22k	no	yes	86.3%	60.7%
P	exact	8%	22k	yes	no	**87.2%**	**61.5%**
P	exact	8%	22k	yes	yes	86.8%	60.4%
P	3-gram	8%	22k	no	no	86.6%	66.6%
P	3-gram	8%	22k	no	yes	86.3%	65.3%
P	3-gram	8%	22k	yes	no	**87.4%**	**66.8%**
P	3-gram	8%	22k	yes	yes	86.6%	66.7%
P	3/2-gr.	8%	22k	yes	no	85.8%	64.0%
P	2-gram	8%	22k	yes	no	83.6%	64.3%
none	-	23%	-	-	-	**73.2%**	-
P	exact	23%	22k	yes	no	**86.3%**	67.3%
P	3-gram	23%	22k	yes	no	85.6%	**69.1%**
P	exact	8%	10k	yes	no	86.9%	60.1%
P	exact	8%	1k	yes	no	86.7%	59.2%
P	3-gram	8%	10k	yes	no	87.0%	63.7%
C	exact	8%	314k	yes	no	87.1%	62.7%
C	exact	8%	100k	yes	no	86.8%	61.9%
C	exact	8%	10k	yes	no	87.6%	59.8%
C	3-gram	8%	314k	yes	no	85.4%	63.2%
C	3-gram	8%	100k	yes	no	85.8%	61.5%

Table 1: *Performances based on Phonolex (P) and Celex (C).*

6. Conclusion

The OOV word problem becomes relevant if OOV rates higher than a few percent are expected, as is the case when limited-domain speech interpretation systems are confronted with natural speech. We presented a statistical, lexical-level approach to explicitely model OOV words, which is capable of correctly detecting about two thirds of OOV words at low false acceptance rates, and at the same time avoids segmentation errors affecting IV words. Consequently, semantic tree accuracy is substantially improved over the baseline closed-vocabulary system for a broad range of model configurations. The evaluations suggest that while the generalization abilities of n-gram based OOV models enable them to outperform exact models, their sensitivity against changes of the penalty parameters render them unsuitable for practical applications. Therefore, future work could aim to find a compromise between the two model types.

7. References

[1] P. Fetter, "Detection and Transcription of OOV Words," Verbmobil, Tech. Rep. 231, August 1998.

[2] M. Thomae, T. Fabian, R. Lieb, and G. Ruske, "A One-Stage Decoder for Interpretation of Natural Speech," in *Proc. NLP-KE'03*, Beijing, China, October 2003. [Online]. Available: http://www.thomae-privat.de/publications/nlpke2003.pdf

[3] I. Bazzi, "Modelling Out-of-Vocabulary Words for Robust Speech Recognition," Ph.D. dissertation, MIT Dept. of Electrical Engineering and Computer Science, June 2002.

[4] M. Thomae, T. Fabian, R. Lieb, and G. Ruske, "Hierarchical Language Models for One-Stage Speech Interpretation," in *Proc. Eurospeech*, Lisbon, Portugal, September 2005.

[5] S. Chen and J. Goodman, "An Empirical Study of Smooting Techniques for Language Modeling," CRTC, Harvard Univ., Cambridge, MA, Tech. Rep. TR-10-98, Aug. 1998.

[6] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. ICSLP*, Denver, Colorado, USA, September 2002.

[7] R. Sproat, "Lextools: a toolkit for finite-state linguistic analysis." [Online]. Available: http://www.research.att.com/sw/tools/lextools/synth.pdf

[8] M. Mohri, F. C. N. Pereira, and M. Riley, "A Rational Design for a Weighted Finite-State Transducer Library," in *WS on Implementing Automata*, 1997, pp. 144–158.

[9] "Pronunciation Lexicon PHONOLEX," Bavarian Archive for Speech Signals, Munich, Germany, June 2004, v3.11.

[10] R. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX Lexical Database (CD-ROM)," LDC, Univ. of Pennsylvania [Distributor], Philadelphia, PA, 1995.

[11] W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin, Germany: Springer, 2000.

[12] M. Thomae, T. Fabian, R. Lieb, and G. Ruske, "Tree Matching for Evaluation of Speech Interpretation Systems," in *Proc. ASRU*, St. Thomas, USVI, Nov. 2003.