USING SYNTACTIC AND CONFUSION NETWORK STRUCTURE FOR OUT-OF-VOCABULARY WORD DETECTION

Alex Marin, Tom Kwiatkowski, Mari Ostendorf, Luke Zettlemoyer

University of Washington, Seattle, Washington 98195

ABSTRACT

This paper addresses the problem of detecting words that are out-of-vocabulary (OOV) for a speech recognition system to improve automatic speech translation. The detection system leverages confidence prediction techniques given a confusion network representation and parsing with OOV word tokens to identify spans associated with true OOV words. Working in a resource-constrained domain, we achieve OOV detection F-scores of 60-66 and reduce word error rate by 12% relative to the case where OOV words are not detected.

***Index Terms*—** OOV detection, speech recognition, parsing

1. INTRODUCTION

Automatic speech recognition (ASR) systems operate with a fixed vocabulary, so systems inevitably run into problems when an out-of-vocabulary (OOV) word is encountered. In such cases, the system will recognize words or sequences of words that have similar phonetic content, as in the following examples with OOV terms underlined in the reference (Ref) and errors in boldface in the hypothesis (Hyp).

REF: what can we get at Litanfeeth
HYP: what can we get **it leaks on feet**

REF: do you know how to properly handle asbestos
HYP: do you know how to properly **and best those pistols**

As these examples illustrate, the errors often introduce strange grammatical constructions, which can sometimes result in errors on neighboring words because of the importance of the language model in the recognition search.

Handling OOV regions is important for a variety of applications, but the constraints of the applications can impact the approach. In audio indexing or spoken term detection, for example, a search term that was not in the original ASR system vocabulary can be handled by indexing a lattice expanded into subword units [1], though retrieval rates may not be as high as for in-vocabulary search terms. In human-computer interaction and speech translation, on the other hand, OOV errors often lead to failure of the interaction. If the presence of an OOV can be detected, the computer can ask the speaker to rephrase or try to learn the new word (e.g. through an interactive dialog with the speaker or a sound-to-letter mapping). In this work, we are interested in computer-mediated speech translation applications, but we focus only on the OOV region detection problem.

Early work in OOV detection used generic acoustic models with OOV word classes in the language model [2, 3, 4, 5]. Such an approach can be effective in a constrained-domain application, but it tends to have a high false detection rate in open-domain or large vocabulary systems. Better results can be obtained by using multiple sub-word fragments as "words" rather than a generic OOV word model, as explored in [6, 7]. Including sub-word fragments is particularly useful for highly inflected languages and for spoken term detection applications. Another approach is to build on word confidence estimation techniques, leveraging speech recognition word posterior probabilities with contextual features (such as local word lattice topology, 1-best neighboring words, language model features, and semantic context) and predicting OOVs instead of or in addition to errors [8, 9, 10, 11, 12].

The work proposed here extends the confidence modeling approach to explicitly leverage parsing, building on the observation that parse structure tends to be anomalous in OOV regions. Working with confusion networks provided by the ASR system, we first estimate the posterior probability (confidence) that a slot in the confusion network aligns with an OOV word in the reference transcription, then add the OOV arc to the network and choose the best path through the network using a parser that incorporates OOV tokens as words. The final confidence of the resulting OOV region can be determined using the first stage OOV confidences, the parser confidence, or a combination of these in a subsequent confidence prediction stage.

Parsers have previously been used to improve the output of the speech recognition system, most often in an n-best or lattice rescoring framework [13, 14, 15, 16]. While the hope is that the parser will improve the recognized word sequence, the goal of this work is primarily to identify OOV regions.

An important consideration in the approach described here is that both computational and data resources are limited. The recognition output is used in a speech-to-speech

This material is based upon work supported by DARPA under Contract No. HR0011-12-C-0016 (subcontract 27-001389). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

translation scenario, with dialog-based computer mediation to resolve unknown words. The recognition, parser and OOV detection together must operate in real-time on a portable platform. As a result, there is a substantial amount of pruning in both the ASR and parser hypothesis spaces. In addition, there are no in-domain hand-annotated parses and only a few hundred sentences with known OOV words. Thus, domain adaptation is used in training the parser.

2. SYSTEM DESCRIPTION

Our system takes as input a word confusion network (WCN), which is a sequence of slots with each slot comprising a list of words and their confidences. The output is a WCN annotated for OOVs. The OOV classification is performed in three stages, as illustrated in Figure 1.

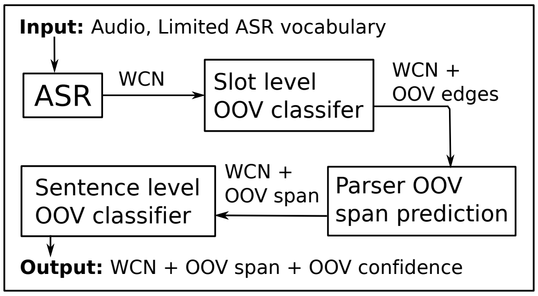
Input: Audio, Limited ASR vocabulary
ASR
WCN
Slot level OOV classifer
WCN + OOV edges
Parser OOV span prediction
WCN + OOV span
Sentence level OOV classifier
Output: WCN + OOV span + OOV confidence

Fig. 1. System diagram

In the first stage an OOV arc is added to each slot in the confusion network. These arcs are assigned a probability according to a maximum entropy classifier designed to recognize the slot level ASR response to out of vocabulary input. The second stage of OOV classification uses a statistical parser to recognize anomalous linguistic structures that are typical of the ASR system's response to an OOV word. The parser can propose OOV parse sub-trees in place of these structures as described in Section 3. The final stage of the classification process uses global features from the parser output to generate a final sentence level prediction of whether an out-of-vocabulary word was present.

2.1. Initial Slot-level Classification

The first-stage classifier predicts, for each slot in the WCN, whether that slot aligns with an OOV word. This stage adds an OOV arc $\bar{w}$ to each WCN slot, generates OOV confidence $p_i(\bar{w}|M_{sl})$ for slot i using model M_{sl}, and renormalizes all other word confidences are renormalized so that the set of posteriors sums to one.

M_{sl} is a maximum entropy (MaxEnt) slot-level classifier based on the MALLET package [17], using features that describe the distribution of arcs in the confusion network slots, summarized in Table 1. Several features are motivated by the observation that the confusion network slots with more arcs tend to correspond to erroneous regions. The *del* feature signals a slot where the ASR system assigns the highest probability to a DEL arc (i.e. a skip). The *highPost* feature is the highest confidence ASR prediction. The *del* and *highPost* features are extracted from the current slot and the previous and subsequent two slots.

Feature	Description
sentLength	length of sentence (in words)
sentPos	position of slot in sentence
mean	mean of slot arc posteriors
stdev	standard deviation of slot arc posteriors
highPost	highest posterior in slot
highLength	length of highest posterior word in slot
del	1 if highest posterior arc is DEL

Table 1. Confusion network features

In addition to the confusion network structural features, we also employ a set of features motivated by the observation that a single OOV word is often replaced by a sequence of shorter, common words. We use a set of binary features obtained by mapping the highest posterior word in the slot to a set of word classes learned from target domain data using Brown clustering [18]. The most common words in the data set are specified as their own class, under the assumption that common words may have more predictive power when used as individual features instead of within word classes. Finally, as an alternative to the parser stage, we use a set of bigram part-of-speech (POS) features, where the POS of a slot is defined as the POS of the highest confidence word in the slot.

Feature selection is performed using the criterion of mutual information with the class variable. The decision threshold and the optimal number of features are selected to optimize the slot-level F-score on a development set.

2.2. Parser-Driven Classification

We parse the confusion network output by the slot level classifier using a probabilistic context free grammar G. This generates the single most probable parse y' of a sequence of edges $\mathbf{w} = [w_0, \ldots, w_{|\mathbf{w}|-1}]$ representing a path through the confusion network. Each parse y is scored according to G, as described in Section 3, and the first stage arc posteriors.

$$y^* = \arg_{\{y,\mathbf{w}\in WCN\}} \max p(y|\mathbf{w}, G) \prod_{i=0}^{|\mathbf{w}|-1} p(w_i|M_{sl})$$

The one-best parse y^* of the WCN is used as an OOV classifier by allowing a path that assigns OOV arcs as terminals in $\mathbf{w}$. An example parse is shown in Figure 2. This parse has correctly classified slots 2 and 3 as representing an OOV region in the original string and assigned the verbal syntactic category VB, allowing a parse that is consistent with the grammar G.

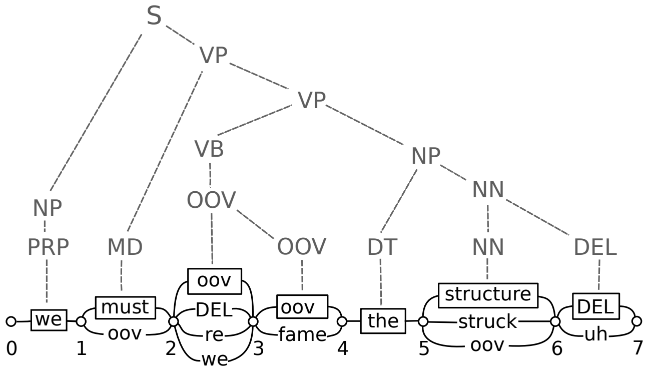

Fig. 2. Example best parse of a WCN for the utterance "We must reframe the structure", where word "reframe" is OOV.

2.3. Sentence-level Classification

One limitation of the parser-driven classification is that long-distance relationships within the utterance cannot be captured efficiently. The parser cannot therefore assign scores to global parse restructuring resulting from the inclusion of OOV subtrees. Since the parser is allowed to hypothesize any category for any sequence of non-zero probability OOV arcs in the WCN, it may overpredict OOV regions in place of low probability but correct ASR output.

We explore scoring global parse structure with a classifier that observes differences between the best parse of the WCN with and without OOV arcs. These differences are signaled using features that correspond to the difference in number of constituents of a particular type (e.g. NP) of a particular length between the two parses, aggregated over all lengths. With preliminary analysis suggesting that many of the differences manifest themselves in noun and verb phrases, we focus on specific syntactic categories: noun phrases, verb phrases, sentence (S and SBAR nodes) and the overall difference between the trees. We also look at the number of OOV non-terminals present in the OOV-aware parse tree.

Using these features, together with the region-level confidence scores obtained from the initial-stage MaxEnt classification and the parser, we perform utterance-level classification using the same maximum entropy classifier package as in the initial-stage. We use low sentence-level scores to filter out OOV regions predicted by the parser or by the initial stage MaxEnt classifier, with the intention of boosting precision while not harming recall.

3. PARSING CONFUSION NETWORKS WITH OOVS

There has been significant previous work on parsing ASR lattices [13, 14] in which each path represents a sequence of recognized words. However, to integrate with a WCN slot based OOV classifier, our parser needs to account for DEL markers used to represent ASR insertions. It must also be able to deal with spans of OOV arcs. Below we describe first modifications to an existing PCFG that supports both of these. Then we describe how these modifications are modeled probabilistically, using a model trained on in-domain data.

3.1. Parser grammar

The PCFG G has as its basis a set of unary and binary context free rewrite rules extracted from the Switchboard corpus [20]. To these we add a pair of binary grammar rules: X → X DEL and X → DEL X that collapse DEL arcs in syntactic parses. In these rules, X may match any non-terminal in G. The first of these rewrites has been used once in Figure 2.

OOV markers on confusion network arcs are treated as a new word 'oov'. These are generated from the syntactic non-terminal OOV via a Markov process integrated into the PCFG parser using the two rules OOV → oov and OOV → OOV oov. The OOV non-terminal can be generated from any other non-terminal in G. This allows the parser to model large spans in the input WCN as relating to an out-of-vocabulary word of any syntactic category. For example in Figure 2 the out-of-vocabulary word 'reframe' has been mapped, by the ASR system, onto two slots in the WCN. These are modeled by the parser as an OOV region with category VB , allowing the same syntactic parse that would have been built if the word 'reframe' had been recognized.

3.2. Parser model

The parser scores each parse using a combination of a generative parsing model trained on the Switchboard corpus and a discriminative CRF used to score OOV segments. We use the Stanford parser [19], modified to as described above, to learn a generative parsing model from counts of rules in Switchboard. We assign the non-Switchboard binary rules probability 0.1, the unary rule OOV → oov probability 1, and the unary rules X → OOV probability 0.1. We renormalize all the rule probabilities to give a generative probabilistic context-free grammar, G_S.

The Markov process used to generate OOV segments in the parse y and the rules used to collapse DEL are modeled discriminatively. These rules fire features $\phi(y)$ on the parent and child categories. The features have weights θ_{oov}. We calculate the probability of a full parse as:

$$\begin{aligned} p(y|WCN,G) = \;& p(y|\mathbf{w},G_S)p(y|\theta_{oov}) \qquad (1)\\ = \;& \prod_{a\to b\in y} p(b|a,G_S)\frac{\exp(\phi(y)\cdot\theta_{oov})}{\sum_{y'}\exp(\phi(y')\cdot\theta_{oov})} \end{aligned}$$

Where both y and y' must have yield in the word confusion network WCN. As we only use local features, this can be calculated efficiently using the inside-outside algorithm.

Domain specific parse parameters θ_{oov} are estimated using an averaged online perceptron learning algorithm [21] from in-domain training data. This allows us to adapt the model trained on the larger Switchboard corpus.

4. EXPERIMENTAL SETUP

4.1. Dataset And Recognizer Setup

We use the SRI Dynaspeak [22] speaker-independent speech recognition system. The system uses a four-gram language model trained on a mixture of in-domain TRANSTAC data and out-of-domain sources such as newswire and broadcast news corpora. The language model is trained with a 30,000 word vocabulary. Word class features are learned on the in-domain training data portion of the recognizer language model. A set of 1000 classes are obtained, augmented with the 50 most common words in the data set. The recognizer lattices are converted to confusion networks using the SRILM toolkit [23]. Confusion network statistics for the development set are included in Table 2.

For training, the reference transcripts are aligned to confusion network slots using a dynamic programming alignment algorithm implemented in SRILM. Each confusion network slot mapped to a reference word not found in the recognizer vocabulary is labeled as matching an OOV. Confusion network slots corresponding to insertions take the positive label of an adjacent OOV slot, if one exists.

Ave # slots	15.7
Ave # arcs per slot	2.4
One-Best WER	14.1
Oracle WER	7.5

Table 2. Dev set confusion network statistics

We focus on a dataset drawn from two sources containing relatively short utterances with a low formality level. Utterances drawn from TRANSTAC data contain few OOVs. Additional utterances with OOVs were recorded by SRI to match the topic and style of the TRANSTAC utterances. Development set statistics are included in Table 3. The "format" OOVs include words that are represented in the vocabulary with a different orthography, e.g. "traveling" vs. "travelling," which are not actually OOV for practical purposes.

Statistic	TRANSTAC	SRI
# Utterances	166	164
# Name OOVs	0	47
# Other OOVs	2	63
# Format OOVs	2	6
Ave utterance length	12.8	9.4

Table 3. Dev set OOV statistics

4.2. Evaluation metrics

We report word error rate (WER) along with precision and recall of out-of-vocabulary predictions. Both of these metrics are evaluated on forced alignments between the Hyp and Ref strings created using the SCLITE toolkit.[1] A hypothesized OOV is considered correct if it is aligned with an OOV tag in the reference. For example, the hypothesis below contains a single correct OOV.

REF:	we	must	OOV	the	structure
HYP:	we		OOV	the	structure

In this example, the predicted OOV region swallows one slot too many in the confusion network, resulting in a Hyp with fewer words than the reference and a word error rate of 20%. As the precision of OOV predictions calculated on the aligned strings only accounts for the position and not the length of the OOV span, it is important to consider also the WER that accounts for both.

5. EXPERIMENTS

We conducted a series of experiments to assess the contributions of the different stages, as well as contrasting conditions that omit the MaxEnt classifier or the parser. Since the parser is constrained to provide only one OOV region, we constrain the first stage MaxEnt classifier in this comparison to output only the region with the highest level score, where the region-level score is based on the highest OOV confidence slot in the region. The first-stage MaxEnt classifier is most similar to past work using acoustic and language model score-based features, and thus serves as our baseline.

5.1. Classification Results

Table 4 contains word error rate, precision, recall, and F-score of OOV prediction results for the various contrasting conditions for each different stage. Results for the three stages on the eval set are given in Table 5. The operating point in each condition is tuned for slot-level F-score on the dev set. When the parser uses a constant prior for OOV words in all slots, word error rate degrades. All other OOV classifier conditions yield a solution with a lower word error rate than the initial ASR prediction. Although the parser hurts performance given a constant OOV prior, it improves WER when seeded with the MaxEnt priors. The full three-stage system achieves an F-score of 66.2% on the development set and 59.6% on the evaluation set. [2]

The initial slot-based MaxEnt classifier recovers 70.6% of OOV regions and achieves a WER 12% better than the original ASR result. Adding the POS bigrams as features does

[1] ftp://jaguar.ncsl.nist.gov/pub/sctk-2.4.0-20091110-0958.tar.bz2

[2] These scores include the "format" OOV contribution; minor F-score differences are observed when these OOVs are removed.

Condition	WER	Precision	Recall	F-score
No OOVs	14.1	n/a	n/a	n/a
MaxEnt	12.8	50.3	70.6	58.8
MaxEnt + POS	16.7	30.6	64.7	41.5
Parser, const priors	17.0	40.8	75.0	52.9
Parser, ME priors	**12.4**	55.6	**80.2**	65.7
MaxEnt + Sent	12.5	55.8	65.5	60.3
Parser + Sent	**12.4**	**56.7**	79.3	**66.2**

Table 4. Results, dev set

Condition	WER	Precision	Recall	F-score
No OOVs	15.5	n/a	n/a	n/a
MaxEnt	14.2	47.6	66.9	55.6
Parser, ME priors	**13.6**	53.2	**67.8**	**59.6**
Parser + Sent	**13.6**	**53.6**	66.9	**59.6**

Table 5. Results, eval set

not help, due to data sparsity. The parser stage improves upon both the precision and recall of the MaxEnt classifier, more accurately identifying OOV regions in which the ASR predictions are not grammatically likely. The gains from the sentence-level filtering are less strong; we find that the filtering helps improve the first stage results, but the precision improvements over the raw parser decisions are minimal, with no F-score improvement on the evaluation set.

The precision-recall trade-off of the three classification stages is illustrated in Figure 3. The MaxEnt slot-level and sentence-level classifiers yield a range of precision-recall pairs at different classification probability thresholds. The parser-based classifier must contain an OOV subtree in the single best parse in order to predict an OOV region in the WCN. Each of these OOV subtrees has an associated likelihood under the distribution in Equation 1. We iterate over a range of thresholds on this likelihood to generate the set of precision-recall pairs in Figure 3 with the actual, zero threshold, prediction at the rightmost extremity.

Both the slot-level and sentence-level MaxEnt classifiers yield an obvious precision-recall trade-off with precision reaching 70% at low recall. The slot-level classifier, however, cannot attain high recall without a steep drop off in precision. Adding the parser stage allows the system to maintain levels of precision while correctly classifying up to 80% of OOV words in the reference-aligned strings. The parse classifier precision-recall curve shows very little variation in precision. This may be due to the ad-hoc nature in which the data points were generated by combining continuous confidences with a non-continuous binary classification. The final stage sentence classifier regains the precision recall trade-off that we expect to see while also improving precision in high recall regions, but a similar gain can be achieved simply by using the highest first stage confidence in the region predicted by the parser.

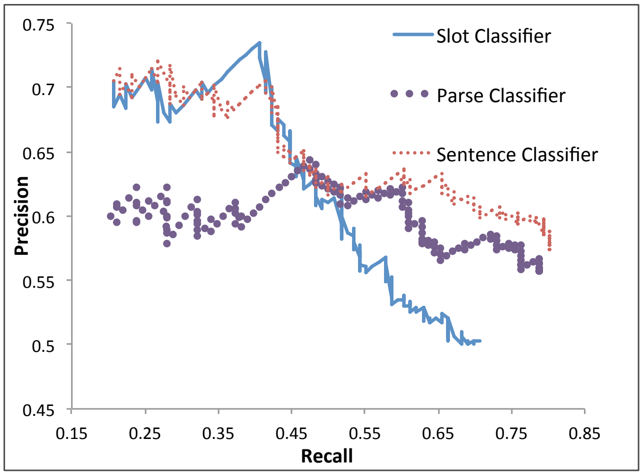

Fig. 3. Precision-recall trade-off of classifiers.

5.2. Confidence Prediction Results

While the region-level metrics allow us to determine whether the predicted OOVs are located in regions of confusion networks which align with true OOVs, they do not measure the length of predicted OOV regions. In particular, too large OOV regions will swallow words which had been recognized correctly; this phenomenon may have a large impact on downstream applications, such as machine translation.

To evaluate the extent of OOV regions, we present a word-level confidence bias plot for the first-stage classifier OOV predictions on the development set. A confidence bias plot measures how the prediction confidences (in this case, OOV posteriors) match the true distribution (in this case, the distribution of OOVs) for each particular posterior value. A perfectly straight $y = x$ line would indicate that the posteriors match the true distribution perfectly; the plot in Figure 4 shows that our predictions are overconfident, in particular in the higher posterior regions. For example, only about 20% of the words with OOV prediction posterior of 0.4 are true OOVs. The sharp drop at the highest point along the graph corresponds to a very small number of samples with high posterior value; we hypothesize that these are alignment errors.

6. DISCUSSION

In this paper, we have introduced a multi-stage system for detecting OOVs in the confusion network output of a speech recognizer, with the initial slot-level confidence prediction stage being used as a prior by a parser stage, whose output is further filtered by a sentence-level prediction with the goal of improving precision. Our results show improvement of each stage over the previous. We observed a performance tradeoff, where the MaxEnt classifier performs better in the low recall region and the parser is best at higher recall levels. Adding the sentence-level filtering is able to match, or slightly out-

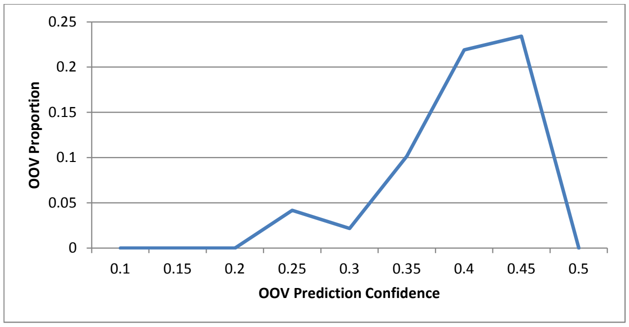

Fig. 4. Word-level confidence bias, dev set.

perform, the top performer in both cases.

We find that the features that help the sentence-level classifier the most are the first-stage MaxEnt predictions and parser predictions, as well as the aggregate NP difference and the overall difference between the OOV-enabled and non-OOV parse trees. This matches our intuition. However, the relatively low weight given to the aggregate features compared to the posterior-based features suggests that significant work is required to make better use of parse information in a post-processing setting.

One potential area for improving the contribution of the parser stage is the availability of additional target domain data. This would allow us to improve the parser by adding more features and lexicalization.

The computational resource constraints also impact the findings of this work. For example, the heavy pruning in the parser makes the parse confidence scores less useful. Given more resources, it would be of interest to investigate richer word lattices, parsing alternatives, and global parse features.

7. ACKNOWLEDGMENTS

The authors graciously thank the SRI STAR Laboratory for recording and providing some of the data used in this paper.

8. REFERENCES

[1] H. Lin et al., "Oov detection by joint word/phone lattice alignment," in *Proc. ASRU*, 2007, pp. 478–483.

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

[3] M. Boros et al., "Semantic processing of out-of-vocabulary words in a spoken dialogue system," in *Proc. Eurospeech*, 1997, pp. 1887–1890.

[4] I. Bazzi, J. Glass, and A. C. Smith, "Modeling out-of-vocabulary words for robust speech recognition," 2000.

[5] T. Schaaf, "Detection of oov words using generalized word models and a semantic class language model," in *Proc. Eurospeech*, 2001, pp. 2581–2584.

[6] A. Rastrow, A. Sethy, and B. Ramabhadran, "A new method for oov detection using hybrid word/fragment system," in *Proc. ICASSP*, 2009, pp. 3953–3956.

[7] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in *Proc. Eurospeech*, 2005, pp. 725–728.

[8] S. Young, "Detecting misrecognitions and out-of-vocabulary words," in *Proc. ICASSP*, 1994, vol. ii, pp. II.21–II.24.

[9] H. Sun et al., "Using word confidence measure for oov words detection in a spontaneous spoken dialog system," in *Proc. Eurospeech*, 2003, pp. 2713–2716.

[10] D. D. Palmer and M. Ostendorf, "Improving out-of-vocabulary name resolution," *Computer Speech and Language*, vol. 19, no. 1, pp. 107–128, 2005.

[11] B. Lecouteux, G. Linars, and B. Favre, "Combined low level and high level features for out-of-vocabulary word detection.," in *Proc. Interspeech*, 2009, pp. 1187–1190.

[12] S. Kombrink et al., "Posterior-based out of vocabulary word detection in telephone speech," in *Proc. Interspeech*, 2009, pp. 80–83.

[13] C. Chelba and F. Jelinek, "Structured language modeling," *Computer Speech and Language*, vol. 14, pp. 283–332, 2000.

[14] B. Roark, "Probabilistic top-down parsing and language modeling," *Computational Linguistics*, vol. 27, no. 2, pp. 249–276, June 2001.

[15] M. Collins, M. Saraclar, and B. Roark, "Discriminative syntactic language modeling for speech recognition," in *Proc. ACL*, 2005, pp. 507–514.

[16] J. Kahn and M. Ostendorf, "Joint reranking of parsing and word recognition with automatic segmentation," *Computer Speech and Language*, vol. 26, no. 1, pp. 1–19, 2012.

[17] A. K. McCallum, "MALLET: A machine learning for language toolkit," http://mallet.cs.umass.edu, 2002.

[18] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, Dec. 1992.

[19] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proc. ACL*, 2003, pp. 423–430.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ACL*, 1992, vol. I, pp. 517–520.

[21] M. Collins, "Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms," in *Proc. ACL/EMNLP*, 2002, pp. 1–8.

[22] J. Zheng et al., "Implementing SRI's Pashto speech-to-speech translation system on a smart phone," in *Proc. SLT*, 2010, pp. 133 –138.

[23] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.