A HIERARCHICAL APPROACH FOR BETTER ESTIMATION OF UNSEEN EVENT LIKELIHOOD IN SPEECH RECOGNITION

Imed Zitouni, Qiru Zhou, Qi Peter Li*

Bell Laboratories, Lucent Technologies, Murray Hill, NJ, USA
{zitouni, qzhou}@research.bell-labs.com
*Li Creative Technologies, Inc. New Providence, NJ, USA
qili@ieee.org

ABSTRACT

The backoff hierarchical class n-gram language models (LMs) are a generalization of the common backoff word n-gram LMs. Compared to the traditional backoff word n-gram LMs that uses (n-1)-gram to estimate the likelihood of an unseen n-gram event, backoff hierarchical class n-gram LMs uses a class hierarchy to define an appropriate context. In this paper, we study the impact of the hierarchy depth on the performance of the approach. Performance is evaluated on several databases such us switchboard, call-home and Wall Street Journal (WSJ). Results show that better improvement is achieved when a shallow word (few levels) tree is used. Experiments show up to *26%* improvement on the unseen events perplexity and up to *12%* improvement in the word error rate (WER).

Keywords: Language modeling, Backoff, n-gram models, Hierarchical class n-gram

1. INTRODUCTION

Backoff word n-gram LMs is the most commonly used approach [Katz (1987)]. When enough data is available, backoff word n-gram LMs have proved extremely useful to estimate the likelihood of n-grams (w_1,…w_n) that occur frequently. However, the estimation of the probability of low frequency and unseen n-grams is still inherently difficult. One of the approaches that can overcome the probability estimation problem of unseen n-grams event is the class n-gram LMs. The class n-gram LMs are more compact and generalize better on unseen n-grams than standard word-based LMs. Nevertheless, for large training corpus, word n-gram LMs are still better in capturing collocational relations between words.

We recently proposed an approach that find a set of classes that is general enough to better model unseen events, but specific enough to capture the ambiguous nature of words [Zitouni et al. (2002)]. This approach hierarchically clusters the vocabulary words, building a word tree. The leaves represent individual words, while the nodes define clusters, or word classes: a node contains all the words of its descendant nodes. The closer a node is to the leaves, the more specific the corresponding class is. At the top of the tree, the root cluster contains all the words in the vocabulary. The tree is used to balance generalization ability and word specificity when estimating the probability of low frequency and unseen events. The backoff hierarchical class n-gram LMs estimate the probability of an unseen event using the most specific class of the tree that guarantees a minimum number of occurrences of this event, hence allowing accurate estimation of the probability. This approach allows us to take advantage of both the power of word n-grams for frequent events and the predictive power of class n-grams for unseen or rare events.

In this paper, we report new results using different databases (Switchboard, Call-home and WSJ) as well as large vocabularies: *16,850* words on switchboard/call-home corpus, and 5000 words as well as *20,000* words on WSJ corpus. We study the influence of the hierarchy depth on the backoff hierarchical class n-gram models: with large number of levels in the class hierarchy, the model becomes less accurate than the baseline n-gram models. The idea of using classes to estimate the probability of unseen events in a backoff word n-gram model was proposed by many researchers [Miller and Alleva (1996); Samuelsson and Reichl (1999)]. The originality of our approach is the use a hierarchical clustering rather than a simple set of classes.

2. BACKOFF HIERARCHICAL CLASS N-GRAM LANGUAGE MODEL

The probability $P(w_i \mid w_{i-n+1}^{i-1})$ is estimated as follows [Zitouni et al. (2003)]:

$$P(w_i \mid w_{i-n+1}^{i-1}) = \begin{cases} \tilde{P}(w_i \mid w_{i-n+1}^{i-1} & \textit{if } N(w_{i-n+1}^{i}) > 0 \\ \alpha(w_{i-n+1}^{i-1}) P(w_i \mid F_{i-n+1}^{1}, w_{i-n+2}^{i-1}) & \textit{otherwise} \end{cases} \quad (1)$$

where F_i^j denotes the j^{th} parent (cluster) of the word w_i: $F_i^j = F^j(w_i)$.

If the event $F_{i-n+1}^{j}, w_{i-n+2}^{i}$ is not found in the training data $N(F_{i-n+1}^{j}, w_{i-n+2}^{i}) = 0$, we recursively use a more general context by going up one level in the hierarchical word clustering tree. This context is obtained by taking the parent of the first class in the hierarchy followed by the *n-2* last words:

$$P(w_i \mid F_{i-n+1}^{j}, w_{i-n+1}^{i-1}) = \begin{cases} \tilde{P}(w_i \mid F_{i-n+1}^{j}, w_{i-n+1}^{i-1} & \text{if } N(F_{i-n+1}^{j}, w_{i-n+1}^{i}) > 0 \\ \alpha(F_{i-n+1}^{j}, w_{i-n+1}^{i-1}) P(w_i \mid w_{i-n+2}^{i-1}) & \text{if } F_{i-n+1}^{j+1} \text{ is theroot} \\ \alpha(F_{i-n+1}^{j}, w_{i-n+1}^{i-1}) P(w_i \mid F_{i-n+1}^{j+1}, w_{i-n+2}^{i-1}) & \text{otherwise} \end{cases} \quad (2)$$

where $\alpha()$ denotes a normalizing constant to guarantee that all probabilities sum to 1 [Zitouni et al. (2002)]. As a result, the whole procedure provides a consistent way to compute the probability of a rare or unseen n-gram by backing-off along the classes that are defined in the hierarchical word tree. If the parent of the class F_{i-n+1}^{j} (respectively, the word w_{i-n+1}) is the class root, the context becomes the last n-2 words, which is similar to the traditional back-off word n-gram model.

3. HIERARCHICAL WORD CLUSTERING ALGORITHM

The hierarchical word-clustering algorithm proceeds in a top-down manner to cluster a vocabulary word set *V*, and is controlled by two parameters: (1) the maximum number of descendant nodes (classes) *C* allowed at each node, (2) the minimum number of words *K* in one class $O_c : (N(O_c) \geq K)$. Experiments shown in this paper is performed with a value of *C* equal to *6*. Other experiments with different values of *C* led to similar performance. Starting at the root node, which contains a single cluster representing the whole vocabulary, we build a maximum number of C clusters (classes) to define the immediate child nodes of the root node. We then continue the process recursively on each descendant node to grow the tree. The criterion used to build the word tree is based on the work of Bai et al. (1998) and uses a concept of minimum discriminative information [Zitouni et al. (2002)].

5. EXPERIMENTS ON SWITCHBOARD AND CALL-HOME DATABASES

5.1 Data description

Call-home database contains approximately *24,000* words and switchboard database contains approximately *3.5* million words. Since the size of call-home database is relatively small to train the language model, we set SWB corpus as the combination of these two databases: switchboard is used for training and call-home is used for test. For language modeling purposes, the used vocabulary contains all the words that appear more than once in the training corpus: *16,850* words (17K). The test set contains *2004* bigram unseen events and *6782* trigram unseen events.

5.2 Test perplexity result

The number of levels in the hierarchy represents the depth of the word tree. Note that according to equation 2, only the probability of unseen events is different between the word n-gram model and hierarchical class n-gram model. To show the real impact of this approach, the test perplexity presented in Figure 1 is computed *only* on the unseen events. The upper plot shows the backoff hierarchical bigram model for different number of levels in the word tree. The lower plot presents the backoff hierarchical trigram model. A number of levels equal to *0* represents the backoff word n-gram test perplexity, which is considered as the baseline.

Results show that the backoff hierarchical class n-gram models outperform the baseline n-gram models. The backoff hierarchical class bigram model shows an improvement of *6%* for the unseen event perplexity (*27517* vs. *25947*). A similar improvement is also observed with the hierarchical class trigram model (*1206* vs. *1140*). The small number of unseen events in the test corpus explains the small improvement in terms of perplexity: the number of unseen bigram events is equal to *2004*, while the number of unseen trigram events is equal to *6782*. Results also show that we do not need a large number of levels in the class hierarchy to improve upon the baseline: two or three levels are enough to achieve a good performance.

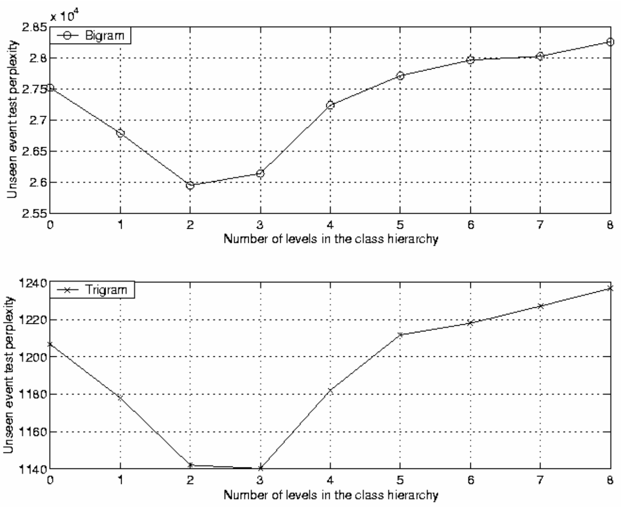

Fig. 1. Unseen events test perplexity on SWB with different number of levels in the class hierarchy

On the whole test set, the perplexity of the backoff bigram model is equal to *179.6*, compared to *177.0* obtained by the hierarchical class bigram model. The backoff trigram model perplexity is equal to *170.1*, compared to *166.1* obtained by the hierarchical class trigram model.

6. EXPERIMENTS ON WALL STREET JOURNAL DATABASE

6.1 Data description

For WSJ experiments, we divided WSJ database into two sets: the training and the test sets. For language modeling purposes, the training set contains *56* million words, and the test set contains approximately *6* million words. Two vocabulary sizes are used: a first one containing *5,000* words (5K) and a second one including *20,000* words (20K). Note that the 5K vocabulary leads to about *2%* of out-of-vocabulary words on the test data, and in that regard differs substantially from the official WSJ 5K lexicon that was designed for a closed-set evaluation (no OOV words). The number of bigram unseen events is approximately equal to *100,000* with the 5K vocabulary, compared to *300,000* with the 20K one. The number of trigram unseen events increases to approximately *900,000* for the 5K vocabulary and to approximately 1 million for the 20K one.

6.2 Test perplexity result

We recall that only the probability of unseen events differs between the word n-gram model and the hierarchical class n-gram model (cf. Equation 2). Hence, we show in Figure 2 the test perplexity on the unseen events only. Like the SWB experiments, the upper pane of Figure 2 shows the backoff hierarchical bigram model and the lower pane presents the backoff hierarchical trigram model. Two plots are shown in each pane: the first plot is based on the 5K vocabulary, while the second plot uses the 20K vocabulary.

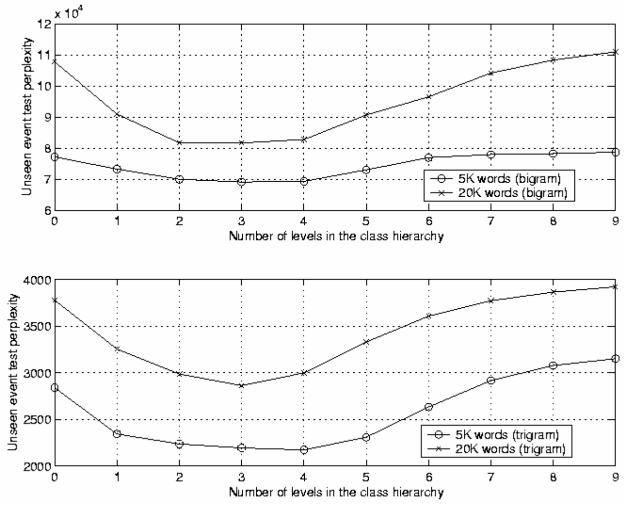

Fig. 2. Unseen event test perplexity on WSJ with different number of levels in the class hierarchy

The number of bigram unseen events is approximately equal to *100,000* with the 5K vocabulary, compared to *300,000* with the 20K one. With the 5K vocabulary, we observe an improvement of *10%* (*77198* vs. *69111*). More than *24%* improvement is reported with the 20K vocabulary (*107824* vs. *81689*). Results show that the test perplexity decreases when the number of unseen events increases. Experimental results also suggest that only few numbers of levels (e.g., 3 or 4) are required in the class hierarchy to yield improvements over the baseline. When using the 5K vocabulary, *3%* improvement of the bigram test perplexity on the whole test set is observed (*100.3* vs. *103.0*). When using the 20K vocabulary, the backoff bigram perplexity on the whole test set is equal to *216.7*, compared to *210.6* obtained by the hierarchical class bigram model.

While using trigrams, the number of unseen events increases to approximately *900,000* for the 5K vocabulary and to approximately 1 million for the 20K one. In the case of trigram, we observe a *23%* improvement of the unseen event perplexity over the baseline on the 5K vocabulary (*2840* vs. *2172*) and more than *24%* improvement on the 20K one (*3778* vs. 2862). The obtained trigram results confirm that as the number of unseen events increases, the proposed approach improves the perplexity compared to the baseline, making it a promising approach for applications using sparse data. The perplexity on the whole test set, using 5K or 20K vocabularies, improves by *6%* approximately (from *69.1* to *64.6* for the 5K vocabulary, and from *140.8* to *132.2* for the 20K one).

The depth of the hierarchy is an important point to consider when building the model. A very flat hierarchy can results in an unwanted over-generalization. On the other hand, a too deep tree can leads to poor generalization for some unseen events. We think that a shallow word (few levels) tree should give better result compared to a deep one (many levels).

7. ASR EXPERIMENTS

The word error rate (WER) on 5K has been evaluated on the 330 sentences of the si_et_05 evaluation set. The 333 sentences of the si_et_20 evaluation set were used for the 20K ASR experiment. We used tied-state triphone acoustic models built on the WSJ SI-84 database. The speech recognition experiments were performed using Bell Labs ASR system [Zhou and Chou (1997)]. We remind that the 5K vocabulary differs from the official WSJ 5K lexicon that was designed for a closed-set evaluation. In table 1, we report the WER obtained using a class hierarchy of two levels. Results show that there is no significant improvement in performance between the baseline backoff bigram model and the hierarchical class bigram model. These results can be explained by the small number of unseen bigrams in this experimental setup and therefore the lack of room for any significant improvement: only *4%* and *8%* of unseen bigrams on the 5K and 20K vocabularies respectively. However, when the trigram is used the number of unseen events increases to *27%* for the 5K vocabulary and to *34%* for the 20K vocabulary, resulting in a considerable relative improvement of the WER: *12%* and *10%* respectively.

Based on these results, we would like to raise the fact that it may be better to reduce the unseen events perplexity rather that frequent event perplexity, since ASR systems work best on frequent events. However, we are uncertain because more parameter tuning of the ASR systems is required. We did not evaluate our ASR system on SWB corpus due to the lack of an acoustic model trained on switchboard database.

	5K		**20K**	
	bigram	trigram	bigram	Trigram
Baseline	9.3%	7.6%	14.2%	12.4%
HCLM	9.0%	6.7%	13.9%	11.2%

Table 1. WER on 5K and 20K vocabularies using word bigram, word trigram, hierarchical class bigram and hierarchical class trigram (HCLM) respectively

9. CONCLUSION

We have discussed in this paper the potential of the HCLMs to estimate the likelihood of unseen events. We show that better performance is achieved with high-unseen event numbers. Compared to the traditional backoff word n-gram LMs, the originality of this approach is in the use of a class hierarchy that leads to a better estimation of the probability of unseen events. We also concluded that the depth of the hierarchy is an important point to be considered when building the model. A tree with few levels should give better result compared to a deep one. Experiments on SWB and WSJ databases show that the improvement of the test perplexity over the standard backoff approach is directly related to the total number of unseen events: *6%* improvement on SWB with *6782* trigram unseen events and *24%* improvement with 1 million trigram unseen events using the 20K vocabulary. Speech recognition results are also promising: up to *12%* improvement in the WER when the hierarchical class trigram model is used. As a future work, we will investigate much more accurate techniques in building the class word hierarchy.

References

[1] Bai S., Li H. Lin Z., Yuan B., 1998. Building Class-based Language Models with Contextual Statistics. In: Proc. ICASSP-1998.

[2] Katz S., 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Transactions on Acoustic, Speech, and Signal Processing 35 (3).

[3] Miller J., Alleva F., 1996. Evaluation of a Language Model using a Clustered Model Backoff. In: Proc. ICSLP-1996.

[4] Samuelsson C., Reichl W., 1999. A Class-Based Language Model for Large-Vocabulary Speech Recognition Extracted from Part-of-Speech Statistics. In: Proc. ICASSP-1999.

[5] Zhou Q., Chou W., 1997. An approach to continuous speech recognition based on self-adjusting decoding graph. In: Proc. ICASSP-1997.

[6] Zitouni I., Siohan O., Kuo H.-K., Lee C.-H., 2002. Backoff Hierarchical Class n-gram Language Modelling for Automatic Speech Recognition Systems. In: Proc. ICSLP-2002.

[7] Zitouni I., Siohan O., Lee C.-H., 2003. Hierarchical Class n-gram Language Models: Towards Better Estimation of Unseen Events in Speech Recognition. In: Proc. Eurospeech-2003.