Within and Across Sentence Boundary Language Model

Saeedeh Momtazi, Friedrich Faubel, Dietrich Klakow

Spoken Language Systems, Saarland University, Germany

{saeedeh.momtazi, friedrich.faubel, dietrich.klakow}@lsv.uni-saarland.de

Abstract

In this paper, we propose two different language modeling approaches, namely *skip trigram* and *across sentence boundary*, to capture the long range dependencies. The skip trigram model is able to cover more predecessor words of the present word compared to the normal trigram while the same memory space is required. The across sentence boundary model uses the word distribution of the previous sentences to calculate the unigram probability which is applied as the emission probability in the word and the class model frameworks. Our experiments on the Penn Treebank [1] show that each of our proposed models and also their combination significantly outperform the baseline for both the word and the class models and their linear interpolation. The linear interpolation of the word and the class models with the proposed skip trigram and across sentence boundary models achieves 118.4 perplexity while the best state-of-the-art language model has a perplexity of 137.2 on the same dataset.

1. Introduction

Statistical language models have been widely used in speech recognition and various natural language processing applications. The task of language modeling is finding the probability of a word given the preceding sequence of words. Most language models fall into one of these two categories: word n-grams which calculates the conditional distribution of each word given the previous $n-1$ words or class n-grams which uses the clusters of $n-1$ words to calculate the word probability. The history used in the n-gram model can cover the whole sentence; however, due to the space complexity and the data sparsity in higher order n-grams, the history is typically limited to trigrams or quadrigrams. As a results, the model can not cover all words' distribution in the entire sentence.

Different research has been done to overcome this problem. Among them, the approaches which use the syntactic information of the entire sentence achieved a promising performance in predicting the next word. Chelba and Jelinek [2] showed that building the syntactic structure incrementally while traversing the sentence left-to-right reduces the text perplexity compared to the trigram baseline. In another research by Roark [3], a probabilistic top-down parser is used to improve the language model performance. Although both models capture all words appearing within the present sentence, they are computationally expensive and a deep syntactic analysis is required for estimating the word probability. In addition, these models are not able to capture the across sentence boundary dependencies.

Such problems in the current n-gram and the structured language models motivated us to introduce a new language model for capturing a longer range dependency of the language not only for the entire sentence, but also considering the previous sentences. This model which uses pure statistical approaches, without a need for syntactic analysis, considers more words of the present sentence with the same space complexity as a normal n-gram model and even uses the words of the previous sentences to estimate the probability of the current word.

In this paper, we use the word and the class models to estimate the word probability. The *skip trigram model* proposed in this paper is used to calculate the word probability while considering long range dependencies within the sentences; The *across sentence boundary model* is used as the emission probability for both the word and the class models to find the long range dependencies across the sentence boundaries. The structure of the paper is as follows: in the next section, the skip n-gram model is described. Section 3 introduces the novel across sentence boundary model. In Section 4, we will show how our new ideas can be used together inside the adapted word or class model frameworks. The experimental results are presented in Section 5; and finally, Section 6 summaries the paper and suggests the future work.

2. The Skip n-gram Model

As mentioned, in an n-gram model, the probability of a word is estimated based on the sequence of $n-1$ previous words appeared in the context. In this paper, we use another type of n-gram, called skip n-gram, for estimating the word probability in which a longer range of dependencies within the sentences is captured. The skip n-gram model uses the same number of predecessor words as the normal n-gram model. However, this model is not limited to the words that exactly appear before the present word; i.e., the model is able to skip some of the previous words and as a result, uses a wider range of words in the sentence compared to the n-gram model while the same memory size is required. Previous studies shows that the skip bigram in the word model can reduce the perplexity [4]. However, to best knowledge of the authors this model has not been applied to the class model and also higher order n-grams. In this paper, we investigate how the interpolation of skip bigram and skip trigram improves the perplexity within the word and the class model frameworks.

Using the log linear interpolation of skip bigram and skip trigram, the probability of the word model is calculated as follows:

$$\begin{aligned} P_{Word}(w|h) = & \frac{1}{Z_\lambda(h)} \\ & \times \prod_{i=1}^{m} P(w|h_{-i})^{\lambda_i} \\ & \times \prod_{\substack{j=1 \\ i=j}}^{m} P(w|h_{-j}h_{-i})^{\lambda_{i,j}} \end{aligned} \quad (1)$$

where w is the sentence word and $h = h_{-1}, h_{-2}, ..., h_{-m}$ is the history of the previous words. λ_i and $\lambda_{i,j}$ are the interpolation

weights, and $Z_\lambda(h)$ is the normalizer in which λ is the union of all interpolation weights. If $c(w)$ is the class that w belongs to, and $c(h)$ is the cluster that each of the words in the history belong to, the class model with skip bigram and skip trigram is defined as

$$\begin{aligned} P_{Class}(w|h) = & \frac{1}{Z_\lambda(h)} P(w|c(w)) \\ & \times \prod_{i=1}^{m} P(c(w)|c(h_{-i}))^{\lambda_i} \\ & \times \prod_{\substack{j=1 \\ i=j}}^{m} P(c(w)|c(h_{-j})c(h_{-i}))^{\lambda_{i,j}} \end{aligned} \quad (2)$$

As an example, in normal trigrams, the probability of the present word is conditioned the probability of the two previous words which are adjacent to the present word; i.e., a window size three is used for the trigram model. Contrarily, in skip trigrams, we still use the two previous words but they are not limited to the adjacent words which are in a window of three. Figure 1 shows the skip bigram and skip trigram models which are distributed over a window of four words. In Figure 1(a), the first model is similar to the normal bigram model, while the second and the third models can not be captured by the normal bigram model. Figure 1(b) shows how the skip trigram model can cover a wider range of dependencies within the sentences while only the first one is captured by the normal trigram model. Of course, all models presented in this figure are covered by a normal quadrigram model; however, it requires more memory, while the skip bigram and the skip trigram models have a comparable space complexity to normal bigram and normal trigram respectively.

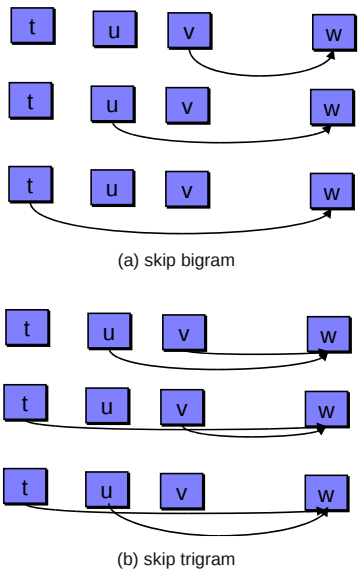

Figure 1: The skip bigram and the skip trigram models over a window of four words

3. The Across Sentence Boundary Model

As mentioned, the current n-gram models are limited to the small set of predecessor words which is not enough to capture the long range dependencies in the text. Although the proposed skip n-gram model goes through more words compared to normal n-grams, the model is still limited to the entire sentence and can not capture the dependencies across the sentence boundaries. To overcome such a problem, we propose another type of language model, called across sentence boundary, which uses a wider context than n-grams and skip n-grams.

In this approach, we consider that there is a close relation between the words in adjacent sentences; so that, the words in the previous sentences trigger the words in the current sentence. As a result, instead of estimating the word unigram from a large corpus, the model estimates the unigram of the present word based on the word distribution in the previous sentences. We hypothesis such an assumption in calculating the word probability improve the unigram model and the proposed across sentence boundary model is usable at the same position as the normal unigram model.

As elaborated in the literature, there is a close correlation between the unigram and the trigram model such that an improved unigram model can reduce the trigram perplexity in both the adapted word and the adapted class models [5]. The proposed model in [5] motivated us to use a novel across sentence boundary language model to improve the unigram model which is used in the adapted word model and applied as an emission probability in the adapted class model. The unigram probability of the current word based on the across sentence boundary model is defined as follows:

$$\begin{aligned} & P(w|S_{-1}, S_{-2}, ...) = \\ & \sum_{\substack{u \in S_{-1} \\ v \in S_{-2} \\ \cdots}} P_{SentSent}(w|u, v, ...) \times f_{S_{-1}}(u) \times f_{S_{-2}}(v) \times ... \end{aligned} \quad (3)$$

where S_{-1} and S_{-2} are the predecessor and pre-predecessor sentences; $f_{s_{-1}}(u)$ and $f_{s_{-2}}(v)$ are the relative frequencies on the previous sentences; and $P_{SentSent}(w|u, v, ..)$ models the relating words in the adjacent sentences.

To simplify the model, we only use one preceding sentence for calculating the word probability while using more sentences from the history is our future work. Although in this model only one adjacent sentence is used, the model can be trained on different ways; i.e., the across sentence boundary model can be trained while the unigram frequency of the present word is estimated based on the word itself, the other words of the same sentence, the words of the predecessor sentence, or the words of the pre-predecessor sentence.

Having the across sentence boundary estimation, the adapted word model with fast marginal adaptation uses this model as its unigram probability:

$$P_{AdaptedWord}(w|h, S_{i-1}) = \frac{1}{Z_\lambda(h)} \left(\frac{P(w|S_{i-1})}{P(w)} \right)^{\lambda_u} P(w|h) \quad (4)$$

where $P(w)$ is the normal word unigram, $P(w|S_{i-1})$ is the word unigram based on the across sentence boundary model as described in (3), and λ_u is the interpolation weight.

The same across sentence boundary estimation is applied to the adapted class model as follows:

$$P_{AdaptedClass}(w|h, S_{i-1}) = P(w|c(w), S_{i-1}) P(c(w)|c(h)) \quad (5)$$

where $P(w|c(w), S_{i-1})$ is the emission probability of w given its class and the previous sentence.

4. The Combined Model

As mentioned, the skip n-gram model is used for estimating $P(w|h)$, and the across sentence boundary model is applied for the emission probability. These two models can be used together in a single word or class model. Equations (6) and (7) present the probability of the adapted word and the adapted class models when the skip bigram and the skip trigram models are used to estimate $P(w|h)$ and $P(c(w)|c(h))$; and the across sentence boundary model is used to calculate the unigram and the emission probabilities respectively.

$$\begin{aligned} P_{AdaptedWord}(w|h, S_{i-1}) =& \frac{1}{Z_\lambda(h)} \left(\frac{P(w|S_{i-1})}{P(w)} \right)^{\lambda_u} \\ & \times \prod_{i=1}^{m} P(w|h_{-i})^{\lambda_i} \\ & \times \prod_{\substack{j=1 \\ i=j}}^{m} P(w|h_{-j}h_{-i})^{\lambda_{i,j}} \end{aligned} \tag{6}$$

$$\begin{aligned} P_{AdaptedClass}(w|h, S_{i-1}) =& \frac{1}{Z_\lambda(h)} P(w|c(w), S_{i-1}) \\ & \times \prod_{i=1}^{m} P(c(w)|c(h_{-i}))^{\lambda_i} \\ & \times \prod_{\substack{j=1 \\ i=j}}^{m} P(c(w)|c(h_{-j})c(h_{-i}))^{\lambda_{i,j}} \end{aligned} \tag{7}$$

Finally, we use the linear interpolation of the adapted word and the adapted class models while our proposed across sentence boundary and skip trigram are used inside each of them.

$$\begin{aligned} &P_{AdaptedInterpolation}(w|h, S_{i-1}) = \\ &\alpha P_{AdaptedWord}(w|h, S_{i-1}) + (1-\alpha) P_{AdaptedClass}(w|h, S_{i-1}) \end{aligned} \tag{8}$$

5. Experimental Results

5.1. Dataset

We tested the long range dependency language model on the Penn Treebank[1], an annotated corpus of American English [1]. We selected this dataset for evaluating our model since the state-of-the-art structured language modeling techniques proposed by Chelba and Jelinek [2] and Roark [3] evaluated on the same dataset and as a result we can have a complete comparison between our proposed models and the best result achieved so far.

To do the experiments, the corpus is divided into three non-overlaping subsets as the train, the development, and the test sets. Penn Treebank Sections 0-20 which consists of 925665 word tokens are used as the training data; Sections 21 and 22 which includes 73447 word tokens are considered as the development set; and the rest of the corpus, Sections 23 and 24 which consists of 81965 word tokens, is used as the test set to calculate the perplexity.

For the experiments on the class model, we used the Brown word clustering algorithm [6] as implemented in the SRILM toolkit [7] while a normal bigram which trained on Penn Treebank Sections 0-20 is used for clustering terms. All results reported in this paper in based on a vocabulary of 10000 word types which are clustered into 500 classes.

Table 1 presents the baseline and the state-of-the-art results for the structured language models. The first row of the table shows the perplexity of the text, while a normal word trigram is used as the baseline. The second and the third rows present the best results achieved so far with the structured language models proposed by Chelba and Jelinek [2] and Roark [3] on the Penn Treebank.

Table 1: Perplexity of the state-of-the-art language models

Model	Perplexity
Trigram	167.1
Chelba & Jelinek Structured LM	148.9
Roark Structured LM	137.2

5.2. Results of the Skip n-gram Model

Considering the previous studies on skip bigram [4], the main contribution of our proposed model is using skip trigram and also applying skip bigram on the class model. For skip trigram, we used the log linear interpolation of skip bigram and skip trigram over a window of four words as described in Equations (1) and (2). The results of this evaluation are presented in the third column of Table 2; while the first column shows the baseline of both the word and the class models and their linear interpolation such that the normal trigram is used. The second column presents the perplexity of skip bigram over the window of four words. The results verify the proposed skip trigram outperforms the normal trigram and the skip bigram, even though our skip trigram requires the same memory space as the normal trigram.

As we can see in this table, the linear interpolation of the word and the class models with skip trigram beats the best results reported in Table 1; which shows that even using only two words of the context without any syntactic analysis is enough to achieve a better performance than the two structured models.

Table 2: Perplexity of the baseline and the proposed skip n-gram models

Model	n-gram	Skip n-gram	
	trigram	bigram	bigram & trigram
Word	167.1	160.0	139.8
Class	179.7	190.3	161.5
Interpolation	140.4	150.9	129.7

5.3. Results of the Across Sentence Boundary Model

Before evaluating the across sentence boundary model within the word and the class models, we first evaluated the model individually and then compared it with the normal unigram model usually used as the emission probability for the adapted word and the adapted class models. The results of this model is presented in Table 3 in which the across sentence boundary model is trained on different range of context. The first row of the table is the perplexity of the normal unigram which serves as the baseline. In the second row, our proposed across sentence boundary model is trained on the present word which is an ideal

[1] http://www.cis.upenn.edu/ treebank/

case for estimating the word unigram and it is the reason this type of training outperforms the other types. The third row presents the results of across sentence boundary while the word distribution of the present sentence is used to estimate the unigram probability. We also used the predecessor sentence and the pre-predecessor sentence for calculating the unigram probability of the present word in which the results are shown in the fourth and the fifth rows of the table. Finally the last row presents the across sentence boundary perplexity while a combination of all different training types is used. The results show that all across sentence boundary models outperform the unigram model such that the combination of the different training methods beat the results achieved by each of these individual methods. This model is used for our further experiments.

Table 3: Perplexity of the across sentence boundary model while training in different ways (ASB stands for Across Sentence Boundary)

Model	Perplexity
Unigram on corpus	626.3
ASB trained on the present word itself	534.2
ASB trained on the same sentence	567.8
ASB trained on the previous sentences	558.3
ASB trained on the pre-predecessor sentences	568.2
ASB with the combined training	498.3

In the next step, the proposed across sentence boundary model is applied in the word and the class models in which the model is used as a unigram model in the adapted word model and as the emission probability in the adapted class model. The results are compared with the baseline which uses the normal unigram model for the same purpose. Table 4 presents the results of this evaluation. To have a better comparison between the across sentence boundary and the baseline model, the baseline results from Table 2 are repeated here. We can see that our new across sentence boundary model improves the baseline in which the linear interpolation of the word and the class models with the new across sentence boundary beat the best results achieved by the state-of-the-art language models presented in Table 1.

Table 4: Perplexity of the baseline and the proposed across sentence boundary model

Model	Baseline	Across Sentence Boundary
Word	167.1	149.5
Class	179.7	158.9
Interpolation	140.4	126.5

5.4. Results of the Combined Model

As the final evaluation, we combined both the across sentence boundary and the skip n-gram models together in which the across sentence boundary is used as the emission probability and the skip n-gram model is used as the main probability presented in Equations (6)-(8); and they are compared with the baseline which uses the normal unigram as the emission probability and normal trigram as the main probability. The results presented in Table 5 show that the combination of our proposed methods outperforms each of the individual methods. In addition, in the adapted word model and in the linear interpolation of the word and the class models, the combination of our proposed models beat the best results achieved by the state-of-the-art language models.

Table 5: Perplexity of the baseline and the combination of the proposed skip n-gram and the across sentence boundary models

Model	Baseline	Combined
Word	167.1	127.4
Class	179.7	143.0
Interpolation	140.4	118.4

6. Concluding Remarks

In this paper, we proposed two different approaches to expand the normal n-gram model while capturing a wider range of dependencies in the language. Although the skip trigram model has the same space complexity as normal trigram, it considers a wider range of words within the sentences to calculate the probability of the current word. Our experiments on this model showed that even though the skip trigram model only uses two words of the context, the model outperforms the structured models which use all words of the sentence and need a deep syntactic analysis. In the across sentence boundary model, the unigram probability is calculated based on the word distribution in the previous sentence. As a result, a wider context is considered when calculating the probability of the present word. The results showed the superiority of this model over the normal unigram model which usually used in the adapted word and the adapted class models. Finally, the perplexity of the combined model verified that the combination of the both techniques achieved a significant improvement compared to the best state-of-the-art language model in which the perplexity decreased from 137.2 to 118.4.

For the future work, we are planing to use different types of training for clustering the vocabulary terms. As mentioned, in the current experiments the same word clusters are used for different proposed language models; while this set of words' clusters is trained on the normal bigram and it is only suitable for the normal bigram model. We believe that a different set of word clusters should be used for each of the proposed models in which the clustering algorithm will be trained on the same way as the language model is going to be trained.

7. Acknowledgments

Saeedeh Momtazi and Friedrich Faubel are funded by the German research foundation DFG through the International Research Training Group (IRTG 715).

8. References

[1] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: the penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313 – 330, 1993.

[2] C. Chelba and F. Jelinek, "Exploiting syntactic structure for language modeling," in *Proceedings of ACL International Conference*, 1998, pp. 225 – 231.

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

[4] D. Klakow, "Log-linear interpolation of language models," in *Proceedings of ICSLP International Conference*, 1998, pp. 1695–1698.

[5] ——, "Language model adaptation for tiny adaptation corpora," in *Proceedings of IEEE Interspeech International Conference*, 2006.

[6] P. Brown, V. Pietra, P. Souza, J. Lai, and R. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[7] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the Spoken Language Processing Conference*, 2002.