First-pass decoding with n-gram approximation of RNNLM: The problem of rare words

Mittul Singh, Peter Smit, Sami Virpioja, Mikko Kurimo

Department of Signal Processing and Acoustics
Aalto University, Espoo, Finland
firstname.lastname@aalto.fi

Abstract

Recurrent Neural Network Language Models (RNNLMs) can be utilized in first-pass decoding by approximating them to *N*-gram models. Although these approximated RNNLMs have shown to improve the Word Error Rate (WER), our experiments show that the word-based *N*-gram approximation seems to be poor at predicting words that occur with low frequency. In our ongoing work, we plan to switch from words to subword units for building approximated RNNLMs to improve the rare word prediction without compromising the general WER. To support this aim, we outline the various challenges and discuss the important factors for building better RNNLM approximations for the first-pass decoding.

Index Terms: rare words, first-pass decoding, rnnlm

1. Introduction

N-gram Language Models (LMs) are preferred to Recurrent Neural Network LMs (RNNLMs) for first-pass decoding in Automatic Speech Recognition (ASR), because they depend only on a short and fixed length word history making these LMs faster to use than RNNLMs.

Nevertheless, the RNNLMs provide richer information through their distributed representation and larger context coverage. Recently, researchers have incorporated this information to first-pass decoding by approximating RNNLMs directly by N-gram LMs [1, 2] or dynamically scoring of *N*-gram LMs using RNNLMs [3, 4]. In experiments on large English-based corpora, these approximated RNNLMs have shown benefits over conventional *N*-gram models [1, 3, 2, 4] in first-pass decoding.

However, adoption of such techniques for under-resourced languages has been limited. The main reason is the approximated RNNLMs inability to express rare words (words occurring with low frequency in the training data) adequately. In this work, we study the modeling of rare words in RNNLM approximations and compare it for a conventional *N*-gram language models on rare words.

We observe that pruning and no rescoring of the approximated RNNLMs can improve the prediction of the rare words, but the ASR performance is adversely affected. This effect indicates that the overall performance of the approximated RNNLMs and the rare word prediction can be conflicting goals. We plan to overcome this problem by applying subwords instead of words in the approximated RNNLMs, and in this paper, we discuss the important factors for designing such models.

2. Approximated RNNs for Rare Words

2.1. Speech Recognition Setup

We use a similar ASR setup to [5]. In this setup, the acoustic models are trained using the Kaldi toolkit [6] on 1500 hours of Finnish audio data from three different data sets, namely, the Speecon corpus [7], the Speechdat database [8] and the parliament corpus [9].

We train the language models on the Finnish Text Collection [10]. The training set consists of 143M tokens with 4.2M unique types. As the *n*-gram LM baselines, we train the Kneser-Ney [11] smoothed trigram (**KN3**) using the VariKN toolkit [12]. For first pass-decoding, the RNNLMs, built similarly to [5], are approximated using *probability-conversion* method as described in [2]. The RNNLMs (non-approximated) are also used for subsequent rescoring in our experiments.

We approximate these RNNLMs to a trigram model (**RNN3**) and interpolate with a smoothed trigram model (**KN3+RNN3**) for first-pass application. The approximated RNNLMS are also pruned using the VariKN toolkit [12] for different pruning threshold ($p \in \{0.001, 0.1\}$).

The trained system is evaluated using Word Error Rate (WER) as a metric on a broadcast news set, obtained from the Finnish national broadcaster YLE. Different language models are also compared on rare words (W_f), where f is the training-set frequency of the words in this set. We calculate the Rare Word Prediction Rates (RWPR) by counting the correctly recognized rare words in a hypothesis transcription (H) given a reference transcription (R):

$$RWPR(f) = \frac{\sum_{(w,s):w\in s, s\in R} \mathbb{1}_{\{w\in H(s) \text{and } w\in W_f\}}}{\sum_{w_i\in R} \mathbb{1}_{\{w_i\in W_f\}}},$$

where s represents an utterance in the reference transcription and $H(s)$ is the corresponding utterance in the hypothesis set. A model that predicts a higher number of rare words correctly than other models is better and hence has a higher RWPR. Similar metric was previously used in [5] to compare models but only on Out-Of-Vocabulary words. RWPR only calculates the miss rate but not the false alarm rate of rare words in hypotheses and in future, we plan to evaluate using Term Weighted Value [13] that captures both aspects of rare word prediction.

2.2. Rare Word Prediction with Approximated RNNLMs

The results of the pruned *N*-gram approximations of RNNLMs are shown in Table 1. The pruning tries to optimize the WER by removing the most unreliable *N*-grams that might hurt the recognition. However, if the pruning threshold is too high, the WER starts to increase, because too many *N*-grams are removed. This effect is observed for only RNN3 and alleviated when interpolating with KN3.

Quite similar to regular RNNs in [14], we observe that approximated RNNs lag behind KN3 in terms of the Rare Word Prediction Rate (RWPR) for rare words, in particular of frequency 1 to 5 as shown in Figure 1. The interpolated model KN3+RNN3, pruning with threshold of 0.1, however, provides

Table 1: *WER for first-pass decoding and after rescoring with RNNLMs for different LMs from Section 2.1 are compared. RWPR for words up to frequency 5 and size in terms of number of N-grams are also displayed for the models in bold.*

Models	WER			RWPR5 (%)	Size
KN3		**16.74**		29.12	67M
Pruning Threshold	0	0.001	0.1	-	-
RNN3	23.26	**21.90**	23.77	25.33	38M
KN3+RNN3	-	16.78	**16.68**	29.32	68M
After rescoring with RNNLM					
KN3		**15.04**		28.85	67M
KN3+RNN3	-	**14.93**	14.94	28.51	80M

a slightly better RWPR than KN3. In our experiments, we also applied larger 5-gram RNNLM approximations and observed similar pruning effects.

When the first-pass KN3+RNN3 lattices have been rescored with an RNNLM [5], the RWPR becomes worse than in the results obtained without rescoring but, the overall performance improves, as shown in Table 1.

Above mentioned observations suggest that improving the overall performance and rare word prediction can be contrary goals for approximated RNNLM. To study and alleviate this effect, we investigate approximated subword-based RNNLMs for first-pass decoding.

3. Approximated subword RNNLMs for first-pass

Prior work [5, 15, 16] has shown that building LMs on subwords like characters instead of words allow better handling of rare words and improve the overall performance simultaneously. Though, non-recurrent neural LMs, with their rich distributed representation, can also be used to handle rare words but [5] uses LMs of large context sizes ($\sim$ 100 units), making the reccurent version a preferable choice for our experiments. In this section, we discuss important factors for designing such a subword RNNLM for first-pass decoding while balancing the overall performance and the rare-word prediction.

3.1. Subword LMs for Speech Recognition

In speech recognition, subword-based N-gram language models have shown impressive Out-Of-Vocabulary (OOV) detection rate improvements over the word-based models [5]. In subword-based speech recognition systems, the acoustic models are typically built using a grapheme-based lexicon. The division of words into subwords reduces the sparsity of the training data, which is particularly important for recognizing rare and even unseen words. In such scenarios, we would like to understand the limits of subword modeling during the successive recognition passes in an ASR pipeline.

3.2. Approximating RNNLMs for first-pass

Quite a few approximation techniques exist for converting RNNLMs to N-gram-based LMs [1, 2, 3, 4]. In [2], the best approximating technique, outperformed other techniques using smaller order N-grams in a speech recognition task. However, using low-order N-grams can not capture long-term information, which is one of the strengths of the RNNLMs.

Hence, techniques that approximate RNNLMs while harnessing the long-term information will be better suited for our purposes. A possible solution would be to extend the order of N-grams using the variable-size N-gram growing algorithm [12] combined into the RNNLM approximation technique. The iterative growing procedure will build long-spanning N-grams for those contexts where needed.

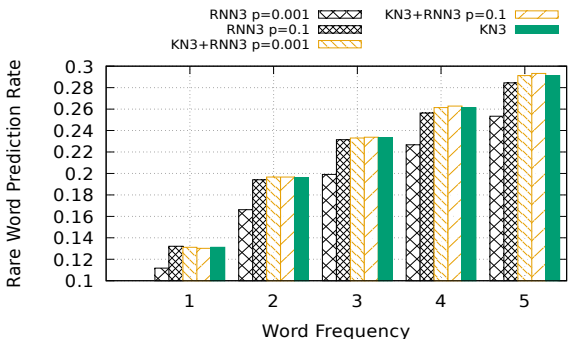

Figure 1: *The Figure 1 displays the first-pass decoding Rare Word Prediction Rate for words up to a frequency f in the training set, where $f = 1$ to 5. The language models compared here are built using different pruning threshold (p).*

An important aspect that is overlooked while approximating RNNLMs is how to model the information forgotten during the approximation process. Hence, we should find ways to model this residual information explicitly.

3.3. Subword Selection

Choice of the subword unit is an important factor while creating subword RNNLMs. In the context of speech recognition, selection of subword units for N-gram language models has also been well studied in earlier work [5]. Investigating whether similar trends will be followed by subword RNNLMs will be interesting. Previously, character-based language models (context size $\sim$ 100 characters) performed better on OOV detection than longer-subword models (context size $\sim$ 50 units) but, performed slightly worse overall. This effect may be due to the larger amount of non-words that single-character models must consider. The RNNLM is likely to be a good model for handling this kind of data sparseness, too.

The choice of the subword unit could affect the method used for approximating RNNLMs, and we plan to investigate these parameters in our next set of experiments.

3.4. Managing the size of Approximated RNNLMs

Decoding speed of a large approximated RNNLM can be a concern, and pruning might be required depending on an ASR task's requirements. Pruning might still affect the overall performance and we might have to combine the pruned language models with conventional N-gram LMs to mitigate any performance dips.

4. Concluding Remarks

Approximated RNNLMs slightly improve the overall performance against N-gram LMs but, as shown in our experiments, the approximated models can adversely affect the ASR performance on rare words. Though, all is not lost. Prior work has applied subword-based N-gram LMs to balance these two goals in conventional N-grams and RNNLMs. In the same vein, we plan to switch to training subword-based RNNLM approximation and outline the important factors for building these models.

5. References

[1] A. Deoras, T. Mikolov, S. Kombrink, M. Karafit, and S. Khudanpur, "Variational approximation of long-span language models for LVCSR," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5532–5535.

[2] H. Adel, K. Kirchhoff, N. T. Vu, D. Telaar, and T. Schultz, "Comparing approaches to convert recurrent neural networks into backoff language models for efficient decoding," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 651–655.

[3] G. Lecorvé and P. Motlicek, "Conversion of recurrent neural network language models to weighted finite state transducers for automatic speech recognition," in *Proceedings of Interspeech*, Sep. 2012, p. to appear.

[4] Z. Huang, G. Zweig, and B. Dumoulin, "Cache based recurrent neural network language model inference for first pass speech recognition," in *ICASSP*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), January 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/cache-based-recurrent-neural-network-language-model-inference-for-first-pass-speech-recognition/

[5] P. Smit, S. R. Gangireddy, S. Enarvi, S. Virpioja, and M. Kurimo, "Character-based units for unlimited vocabulary continuous speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2017, pp. 149–156.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[7] D. J. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kieling, "Speecon - speech databases for consumer devices: Database specification and validation." in *LREC*. European Language Resources Association, 2002.

[8] A. Rosti, A. Rämö, T. Saarelainen, and J. Yli-Hietanen, "Speechdat Finnish database for the fixed telephone network," Tampere University of Technology, Tech. Rep., 1998.

[9] A. Mansikkaniemi, P. Smit, and M. Kurimo, "Automatic construction of the Finnish parliament speech corpus," ser. Interspeech 2017, 2017-08, A4 Artikkeli konferenssijulkaisussa, pp. 3762–3766. [Online]. Available: http://urn.fi/URN:NBN:fi:aalto-201710157137

[10] "The Helsinki Korp version of the Finnish text collection," http://urn.fi/urn:nbn:fi:lb-2016050207, accessed: 2018-06-29.

[11] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 1995, pp. 181–184 vol.1.

[12] V. Siivola, T. Hirsimaki, and S. Virpioja, "On growing and pruning kneser-ney smoothed n-gram models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1617–1624, July 2007.

[13] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," 2007, pp. 45–50.

[14] I. Oparin, M. Sundermeyer, H. Ney, and J. L. Gauvain, "Performance analysis of neural networks in combination with n-gram language models," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 5005–5008.

[15] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016. [Online]. Available: https://arxiv.org/pdf/1602.02410.pdf

[16] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 2741–2749. [Online]. Available: http://dl.acm.org/citation.cfm?id=3016100.3016285