# LSTM-Based Language Models for Very Large Vocabulary Continuous Russian Speech Recognition System

Irina Kipyatkova(✉)

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia
kipyatkova@iias.spb.su

**Abstract.** This paper presents language models based on Long Short-Term Memory (LSTM) neural networks for very large vocabulary continuous Russian speech recognition. We created neural networks with various numbers of units in hidden and projection layers using different optimization methods. Obtained LSTM-based language models were used for N-best list rescoring. As well we tested a linear interpolation of LSTM language model with the baseline 3-gram language model and achieved 22% relative reduction of the word error rate with respect to the baseline 3-gram model.

**Keywords:** Speech recognition · Recurrent Neural Networks · Long Short-Term Memory · Language models · Russian speech

## 1 Introduction

A language model (LM) is one of the main parts of a speech recognition system. Nowadays, neural networks (NNs) are widely used for language modeling. As it was shown in many papers, NN-based LMs outperform standard n-gram models [1, 2]. For language modeling, the usage of recurrent NNs (RNNs) is preferable because this type of NN can store the whole context preceding the given word in contrast to feedforward NNs which store a context of restricted length.

A long short-term memory (LSTM) network is RNN, which contains special units called memory blocks. Each memory block is composed of a memory cell, which stores the temporal state of the network, and multiplicative units named gates (an input gate, an output gate, and a forget gate) controlling the information flow [3].

In our research we used a LSTM-based LM for N-best list rescoring for automatic speech recognition (ASR) system. The paper is organized as follows: in Sect. 2 we give a survey of application of LSTMs for language modeling, in Sect. 3 we give a description of our LSTM-based LMs, experimental results of N-best list rescoring using LSTM-based LMs are presented in Sect. 4.

A. A. Salah et al. (Eds.): SPECOM 2019, LNAI 11658, pp. 219–226, 2019.
https://doi.org/10.1007/978-3-030-26061-3_23

## 2 Related Works

LSTMs are widely used in speech recognition systems at N-best or lattice rescoring stage. In [4] comparison of LMs based on n-grams, feedforward, recurrent, and LSTM NNs in terms of perplexity and word error rate (WER) is presented. LMs were created for English and French. In the paper, it was shown that application of LSTM-based LMs for lattice rescoring outperforms other type of LMs. In addition, experimental analysis of relationship between perplexity of NN-based LMs and WER was performed. It showed that WER decreases with decreasing perplexity that is analogous to correlation between perplexity and WER for n-gram LMs.

In [5] LSTM-based LM was used for lattice rescoring for a YouTube speech recognition task. The proposed model decreased WER by 8% as compared with the result obtained with the n-gram model.

Automatic speech recognition for conversational Finnish and Estonian speech with LSTM LM is described in [6]. The authors tried subword-based and fullword-based language modeling and investigated the usage of classes for language modeling. LSTM LM was used for lattice rescoring. On both languages, the best results were obtained from class-based subword models.

Czech language modeling using LSTM is represented in [7]. As the baseline, 5-gram Knesser-Ney statistical model with 120 K vocabulary was used. The LSTM LMs were trained with limited vocabulary consisted of 10 K most frequent words. LSTM LM interpolated with the baseline model was used for rescoring of 1000-best list. Experiments were performed on the corpus of Czech spontaneous speech which was recorded from phone calls. Application of LSTM LM allowed increasing speech recognition accuracy by 3.7% in relative comparing to the result obtained with the baseline model.

A comparison of LMs based on LSTM and gated recurrent units (GRU) is presented in [8]. In experiments of lattice rescoring for English speech recognition task, LSTM-based LM outperformed GRU-based LM in terms of both perplexity and WER. Also experiments with Highway network based on GRU were performed that showed WER improvement, but similar investigation on the base of LSTM was not conducted.

In [9] a system which uses LSTM for both acoustic and language modeling is presented. The system uses CNN-BLSTM acoustic models and 4-gram LM for decoding and lattice rescoring. LSTM-based LM was applied for 500-best list rescoring. Relative WER reduction obtained after rescoring was about 20%.

Russian language modeling with the use of LSTM is described in [10]. The baseline 3-gram LM was trained on transcriptions of telephone conversations (390 h of speech) as well as on text corpus (about 200 M words) containes materials from Internet forum discussions, books etc. Vocabulary for the baseline model contains 214 K words. NN-based LMs were trained only with a part of the test corpus, and for this corpus the vocabulary of 45 K most frequent words was used. LSTM-based LM was used for rescoring of 100-best list. Relative WER reduction was equal to 8%.

In our previous researches on Russian language modeling [11, 12] we have experimented with LMs created on the base of RNN with one hidden layer using RNNLM toolkit [13]. We have obtained relative WER reduction of 14% as compared to the result obtained with our 3-gram model. The current research is aimed to investigation of another type of RNN for language modeling.

## 3 LSTM Language Models for Russian

For training of LSTM language models, we used TheanoLM toolkit [14]. We trained LMs on a text corpus composed with the use of on-line Russian newspapers [15]. The vocabulary size was 150 K word-forms. We created NN LMs consisting of a projection layer, which maps words to specified dimensional embeddings, one hidden LSTM layer, and a hierarchical softmax layer. Hierarchical softmax factors the output probabilities into the product of multiple softmax functions [16]. Thus, the output layer is factorized into two levels, both performing normalization over an equal number of choices [6], it allows using of very large vocabulary for language modeling. NN LM architecture is presented on Fig. 1, where $w_t$ is an input word at time $t$; $h_t$ is the hidden layer state, $c_t$ is LSTM cell state.

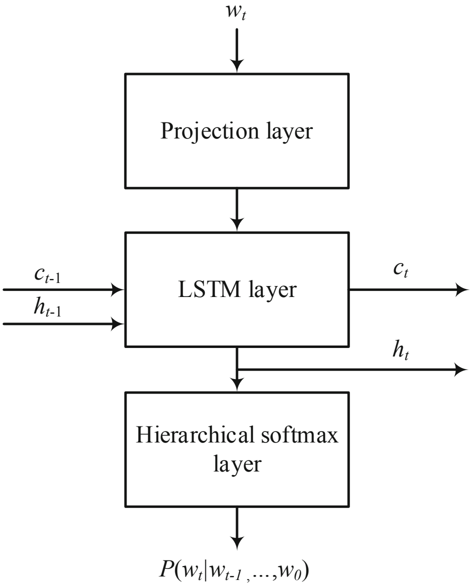

**Fig. 1.** LSTM-based LM architecture

We tried NNs with LSTM layer sizes equal to 256 and 512, and projection layer sizes equal to 100, 500, and 1000. LSTM-based LMs were trained using stochastic gradient descent (SGD) optimization method. The stopping criteria was "*no-improvement*" which means that learning rate is halved when validation set perplexity stops improving, and training is stopped when the perplexity does not improve at all with the current learning rate [14]. The maximum number of training epoch was 15. The initial learning rate was equal to 1.

As well, we made a linear interpolation of the LSTM-based LM and baseline LM. As a baseline, we used 3-gram LM with Kneser-Ney discounting trained on the same text corpus using the SRI Language Modeling Toolkit (SRILM) [17]. Perplexities of the obtained LMs computed on held-out text data are presented in Table 1. The interpolation coefficient of 1.0 means that only LSTM-based LM was used. The perplexity of the baseline model was 553.

**Table 1.** Perplexities of LSTM LMs

| Hidden layer size | Projection layer size | Interpolation coefficient | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| 256 | 100 | 339 | 336 | 336 | 343 | 359 | 431 |
| | 500 | 330 | 325 | 325 | 330 | 345 | 407 |
| | 1000 | 328 | 323 | 323 | 328 | 342 | 405 |
| 512 | 100 | 317 | 311 | 310 | 313 | 325 | 383 |
| | 500 | 308 | 301 | 299 | 302 | 311 | 363 |
| | 1000 | 306 | 300 | **297** | 300 | 309 | 361 |

The lowest perplexity was obtained with the NN with the projection layer size equal to 1000 and the hidden layer size equal to 512. Interpolation with the 3-gram model gave the additional improvement of perplexity. The interpolation coefficient equal to 0.7 provided the best result. Thus, relative reduction of perplexity was 46% as compared with the perplexity of the baseline model.

# 4 Experiments

## 4.1 Experimental Setup

For training the acoustic models and testing the speech recognition system, we used our own corpora of continuous Russian speech recorded at SPIIRAS. The total duration of the entire speech data is more than 30 h. The corpus is described in detail in [18].

We used hybrid DNN/HMMs acoustic models based on time-delay neural network with 5 hidden layers and time context [−8, 8]. Acoustic models were trained using the open-source Kaldi toolkit [19]. Mel-frequency cepstral coefficients (MFCCs) were used as input to the NNs. For speaker adaptation, 100-dimensional i-Vector [20] was appended to the 40-dimensional MFCC input. Detail description of our acoustic models is presented in [12]. We have obtained WER equal to 17.62% with our baseline 3-gram model, and WER equal to 15.13 was obtained after rescoring 500-best list with the help of RNN LM with one hidden layer interpolated with the 3-gram model.

LSTM-based LM was applied for rescoring of 500-best list of hypotheses and for selection of the best recognition hypothesis for the pronounced phrase. Interpolated LMs were used for rescoring as well. Obtained speech recognition results are presented in Table 2.

**Table 2.** WER after 500-best list rescoring (%)

| Hidden layer size | Projection layer size | Interpolation coefficient | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| 256 | 100 | 15.36 | 15.09 | 15.15 | 15.37 | 15.73 | 16.20 | 16.44 |
| | 500 | 15.54 | 15.39 | 15.41 | 16.62 | 15.86 | 16.09 | 16.52 |
| | 1000 | 15.21 | 15.06 | 14.94 | 14.79 | 14.94 | 15.22 | 15.67 |
| 512 | 100 | 15.17 | 14.83 | 14.59 | 14.74 | 14.85 | 15.02 | 15.39 |
| | 500 | 14.51 | 14.36 | 14.21 | **14.06** | 14.19 | 14.64 | 14.96 |
| | 1000 | 15.32 | 15.21 | 15.04 | 15.13 | 15.15 | 15.36 | 15.52 |

As one can see from the table, application of LSTM-based LMs allows to improve speech recognition results. Additional improvement was achieved with interpolated LSTM-based LM with baseline LM. The lowest WER (14.06%) was obtained using NN with projection layer size equal to 500 and hidden layer size equal to 512 interpolated with the baseline model with interpolation coefficient equal to 0.7, though this model was not the best in terms of perplexity. This may be connected with the fact that we used different texts material for estimation of perplexity and for speech corpora recordings.

Then we experimented with optimization method for NN training. We tried Nesterov Momentum [21], AdaGrad [22], and Adam [23] optimization methods, and compared them with SGD method in terms of perplexity and WER of the created models. We trained models with 512 units in the hidden layer and 512 units in the projection layer because LSTM with these parameters gave us the best results in terms of WER in our previous experiments with models with SGD optimization method. Initial learning rates were chosen according to recommendations of TheanoLM toolkit. Results of experiments on comparing optimization methods in term of perplexity and WER are presented in Tables 3 and 4 respectively.

**Table 3.** Results of experiments with LMs trained using different optimization methods in terms of perplexity

| Optimization method | Initial learning rate | Interpolation coefficient | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| SGD | 1.00 | 308 | 301 | 299 | 302 | 311 | 363 |
| Nesterov momentum | 1.00 | 299 | 292 | **289** | 291 | 300 | 346 |
| AdaGrad | 1.00 | 308 | 302 | 300 | 303 | 313 | 375 |
| Adam | 0.01 | 321 | 316 | 314 | 317 | 327 | 386 |

**Table 4.** Results of experiments with LMs trained using different optimization methods in terms of WER (%)

| Optimization method | Initial learning rate | Interpolation coefficient | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| SGD | 1.00 | 14.36 | 14.21 | 14.06 | 14.19 | 14.64 | 14.96 |
| Nesterov Momentum | 1.00 | 14.33 | 14.08 | **14.01** | 14.16 | 14.34 | 14.55 |
| AdaGrad | 1.00 | 15.00 | 14.93 | 14.91 | 14.81 | 14.89 | 15.36 |
| Adam | 0.01 | 14.78 | 14.63 | 14.53 | 14.48 | 14.68 | 14.78 |

Only Nesterov Momentum method slightly outperform SGD in terms of both perplexity and WER of the obtained models. Thus, the best results (perplexity equals 289; WER equals 14.01) were obtained after interpolation of LSTM LM trained using Nesterov Momentum optimization method interpolated with the baseline LM with interpolation coefficient equal to 0.7.

Then we trained NNs with 2 and 3 LSTM layers using the parameters of the best 1-layer LSTM. In these NNs we applied dropout at rate 0.3 between LSTM layers. Obtained results are presented in Table 4.

**Table 5.** Results of experiments with LMs with different number of LSTM layers

| Number of LSTM layers | Interpolation coefficient | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0.7 | | 0.8 | | 0.9 | | 1.0 | |
| | Perplexity | WER, % | Perplexity | WER, % | Perplexity | WER, % | Perplexity | WER, % |
| 1 | 289 | 14.01 | 291 | 14.16 | 300 | 14.34 | 346 | 14.55 |
| 2 | 286 | 13.88 | 279 | 13.80 | 292 | 13.90 | 323 | 13.93 |
| 3 | 294 | 14.05 | 301 | 14.23 | 327 | 14.35 | 357 | 14.62 |

Thus, the best result was obtained using NN LM with 2 LSTM layers interpolated with the baseline LM with interpolation coefficient of 0.8, in this case WER equaled 13.80%. Further increasing the number of the hidden layers led to increasing WER that may be caused by overtraining (Table 5).

## 5 Conclusions and Future Work

In the paper, we have investigated LSTM-based LMs for Russian speech recognition task. We have tried NNs with different hidden layer sized, projection layer sizes, optimization methods, and number of hidden layers. LSTM-based LMs were applied for N-best list rescoring. The lowest WER was achieved with the NN with 2 hidden layers, 512 units in hidden layer and projection layer of 500 trained with Nesterov Momentum optimization method. We have achieved 22% relative reduction of WER using LSTM LM with respect to the baseline 3-gram model. In further research, we are going to investigate other topologies of RNNs for language modeling.

**Acknowledgements.** This research is financially supported by the Ministry of Science and Higher Education of the Russian Federation, agreement No. 14.616.21.0095 (reference RFMEFI61618X0095).

## References

1. Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH'2010, Makuhari, Chiba, Japan, pp. 1045–1048 (2010)
2. Sundermeyer, M., Oparin, I., Gauvain, J.-L., Freiberg, B., Schluter, R., Ney, H.: Comparison of feedforward and recurrent neural network language models. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, B.C., Canada, pp. 8430–8434 (2013)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
4. Sundermeyer, M., Ney, H., Schlüter, R.: From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. Audio, Speech, Lang. Process. **23**(3), 517–529 (2015)
5. Kumar, S., Nirschl, M., Holtmann-Rice, D., Liao, H., Suresh, A.T., Yu, F.: Lattice rescoring strategies for long short term memory language models in speech recognition. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 165–172 (2017)
6. Enarvi, S., Smit, P., Virpioja, S., Kurimo, M.: Automatic speech recognition with very large conversational finnish and estonian vocabularies. IEEE Trans. Audio, Speech, Lang. Process. **25**(11), 2085–2097 (2017)
7. Soutner, D., Müller, L.: Application of LSTM neural networks in language modelling. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 105–112. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40585-3_14
8. Irie, K., Tüske, Z., Alkhouli, T., Schlüter, R., Ney, H.: LSTM, GRU, highway and a bit of attention: an empirical overview for language modeling in speech recognition. In: INTERSPEECH-2016, pp. 3519–3523 (2016)
9. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A.: The Microsoft 2017 Conversational Speech Recognition System. Preprint arXiv:1708.06073, https://arxiv.org/abs/1708.06073 (2017)
10. Medennikov, I., Bulusheva, A.: LSTM-based language models for spontaneous speech recognition. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS (LNAI), vol. 9811, pp. 469–475. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43958-7_56
11. Kipyatkova, I., Karpov, A.: Language models with RNNs for rescoring hypotheses of Russian ASR. In: Cheng, L., Liu, Q., Ronzhin, A. (eds.) ISNN 2016. LNCS, vol. 9719, pp. 418–425. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40663-3_48
12. Kipyatkova, I.: Improving Russian LVCSR using deep neural networks for acoustic and language modeling. In: Karpov, A., Jokisch, O., Potapova, R. (eds.) SPECOM 2018. LNCS (LNAI), vol. 11096, pp. 291–300. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99579-3_31

13. Mikolov, T., Kombrink, S., Deoras, A., Burget, L., Černocký, J.: RNNLM - recurrent neural network language modeling toolkit. In: ASRU 2011 Demo Session (2011)
14. Enarvi, S., Kurimo, M.: TheanoLM—an extensible toolkit for neural network language modeling. In: INTERSPEECH-2016, pp. 3052–3056 (2016)
15. Kipyatkova, I., Karpov, A.: Lexicon size and language model order optimization for Russian LVCSR. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS (LNAI), vol. 8113, pp. 219–226. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-01931-4_29
16. Morin, F., Bengio, Y.: Hierarchical probabilistic neural network language model. In: Cowell, R.G., Ghahramani, Z. (eds.) Proceedings of International Workshop on Artificial Intelligence and Statistics (AISTATS), New Jersey, USA, pp. 246–252. Society for Artificial Intelligence and Statistics (2005)
17. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: update and outlook. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop ASRU 2011, Waikoloa, Hawaii, USA (2011)
18. Kipyatkova, I.: Experimenting with hybrid TDNN/HMM acoustic models for Russian speech recognition. In: Karpov, A., Potapova, R., Mporas, I. (eds.) SPECOM 2017. LNCS (LNAI), vol. 10458, pp. 362–369. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66429-3_35
19. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding ASRU (2011)
20. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-vectors. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 55–59 (2013)
21. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, Heidelberg (2013)
22. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
23. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations, pp. 1–13 (2015)