

Improving neural network models for natural language processing in Russian with synonyms

Galinsky R., Alekseev A., Nikolenko S.

Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2016 FRUCT. Recent advances in deep learning for natural language processing achieve and improve over state of the art results in many natural language processing tasks. One problem with neural network models, however, is that they require large datasets, including large labeled datasets for the corresponding problems. In this work, we suggest a data augmentation method based on extending a given dataset with synonyms for the words appearing there. We apply this approach to the morphologically rich Russian language and show improvements for modern neural network NLP models on standard tasks such as sentiment analysis.

References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wat-tenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [2] N. Abramov. Dictionary of Russian Synonyms and Synonymous Phrases. Moscow: Russkie Sloviri, 1999.
- [3] R. Al-Rfou, B. Perozzi, and S. Skiena. Polyglot: Distributed word representations for multilingual nlp. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 183-192, Sofia, Bulgaria, 2013. Association for Computational Linguistics.
- [4] Z. E. Alexandrova. Dictionary of Russian Synonyms. Moscow: Russkii Yazyk, 2001.
- [5] Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3: 1137-1155, 2003.
- [6] Y. Bengio, H. Schwenk, J.-S. Senecal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning, pages 137-186. Springer, 2006.
- [7] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Des-jardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: A CPU and GPU math expression compiler. In Proceedings of the Python for Scientific Computing Conference (SciPy), 2010. Oral Presentation.
- [8] J. A. Botha and P. Blunsom. Compositional morphology for word representations and language modelling. In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 1899-1907, 2014.
- [9] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. Comput. Linguist., 18 (4): 467-479, 1992.
- [10] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, ACL '96, pages 310-318, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.
- [11] F. Chollet. Keras. <https://github.com/fchollet/keras>, 2015.

- [12] J. Chung, K. Cho, and Y. Bengio. A character-level decoder without explicit segmentation for neural machine translation. CoRR, abs/1603.06147, 2016.
- [13] R. Cotterell, H. Schiitze, and J. Eisner. Morphological smoothing and extrapolation of word embeddings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, 2016.
- [14] S. Dieleman, J. Schluter, C. Raffel, E. Olson, S. K. Sønderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, J. D. Fauw, M. Heilman, D. M. de Almeida, B. McFee, H. Weideman, G. Takacs, P. de Ri-vaz, J. Crall, G. Sanders, K. Rasul, C. Liu, G. French, and J. Degraeve. Lasagne: First release, 2015.
- [15] Y. Goldberg. A primer on neural network models for natural language processing. CoRR, abs/1510.00726, 2015.
- [16] J. T. Goodman. A bit of progress in language modeling. Comput. Speech Lang., 15 (4): 403-134, 2001.
- [17] A. Graves, S. Fernandez, and J. Schmidhuber. Bidirectional LSTM networks for improved phoneme classification and recognition. In Artificial Neural Networks: Formal Models and Their Applications-ICANN 2005, 15th International Conference, Warsaw, Poland, September 11-15, 2005, Proceedings, Part II, pages 799-804, 2005.
- [18] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18 (5-6): 602-610, 2005.
- [19] K. Kann and H. Schiitze. Single-model encoder-decoder with explicit morphological representation for reinflection. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers, 2016.
- [20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014.
- [21] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on, volume 1, pages 181-184 vol. 1, 1995.
- [22] M. Korobov. Morphological analyzer pymorphy. <http://pymorphy.readthedocs.io/>.
- [23] O. Kozłowa and A. Kutuzov. Improving distributional semantic models using anaphora resolution during linguistic preprocessing. In Proceedings of International Conference on Computational Linguistics "Dialogue 2016", 2016.
- [24] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In International Symposium on Circuits and Systems (ISCAS 2010), May 30-June 2, 2010, Paris, France, pages 253-256, 2010.
- [25] W. Ling, C. Dyer, A. W. Black, I. Trancoso, R. Fernandez, S. Amir, L. Marujo, and T. Luis. Finding function in form: Compositional character models for open vocabulary word representation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1520-1530, Lisbon, Portugal, 2015. Association for Computational Linguistics.
- [26] M.-T. Luong, R. Socher, and C. D. Manning. Better word representations with recursive neural networks for morphology. In CoNLL, Sofia, Bulgaria, 2013.
- [27] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013.
- [28] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. Recurrent neural network based language model. INTERSPEECH, 2: 3, 2010.
- [29] T. Mikolov, S. Kombrink, L. Burget, J. H. Cernocky, and S. Khudanpur. Extensions of recurrent neural network language model. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 5528-5531. IEEE, 2011.
- [30] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. CoRR, abs/1310.4546, 2013.
- [31] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In Advances in neural information processing systems, pages 1081-1088, 2009.
- [32] A. Panchenko, N. Loukachevitch, D. Ustalov, D. Paperno, C. M. Meyer, and N. Konstantinova. Russe: The first workshop on Russian semantic similarity. In Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue), pages 89-105, 2015.
- [33] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532-1543, Doha, Qatar, 2014. Association for Computational Linguistics.
- [34] M. Ranzato, G. E. Hinton, and Y. LeCun. Guest editorial: Deep learning. International Journal of Computer Vision, 113 (1): 1-2, 2015.
- [35] R. Soricut and F. Och. Unsupervised morphology induction using word embeddings. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1627-1637, Denver, Colorado, 2015. Association for Computational Linguistics.
- [36] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 649-657. Curran Associates, Inc., 2015.