Language Model Embeddings Improve Sentiment Analysis in Russian

Baymurzina D. R. (dilyara.rimovna@gmail.com)
Kuznetsov D. P. (kuznetsov.den.p@gmail.com)
Burtsev M. S. (burtsev.m@gmail.com)
Neural Networks and Deep Learning Lab
Moscow Institute of Physics and Technology

Goals

- To train language models on Russian corpora of different language styles (linguists emit five Russian language styles: scientific, *official*, *journalistic*, artistic and *colloquial*).
- To examine how language style affects the language model performance and quality of embeddings from language models.

Language modelling data

1. *Official style:* Wikipedia was spared from html-markup.
2. *Journalistic style:* WMT News is distributed preprocessed and cleaned.
3. *Colloquial style:* In Twitter all hashtags and user logins were replaced by special tokens.

Vocabulary size for each dataset is 1 million tokens.
Every dataset was split to train (98%) and validation (2%) sets.

Dataset	Number of words	Vocabulary size	Average number of words per sentence	File Size
Wiki	472 M	5.6 M	19.4	4.8 Gb
WMT News	1133 M	4.1 M	19.6	12.0 Gb
Twitter	887 M	11.3 M	8.7	7.9 Gb

Classification data

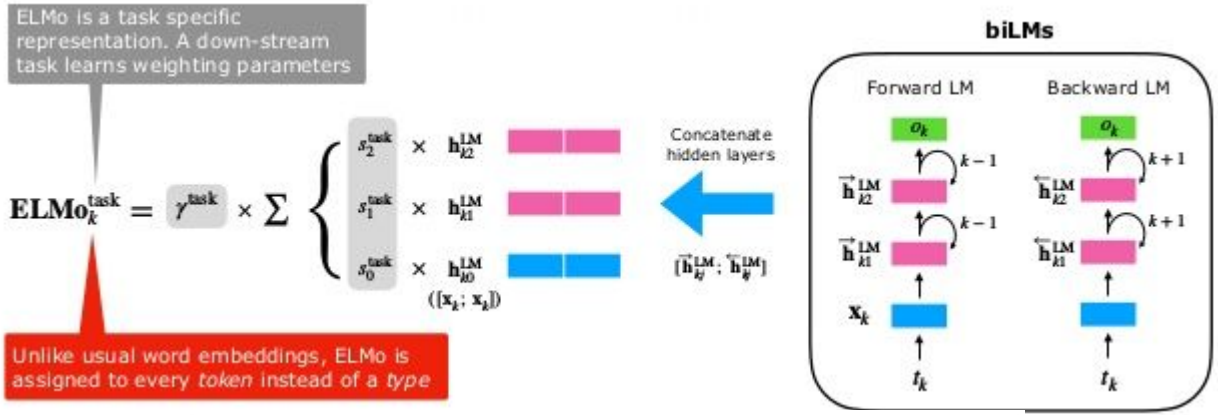
- RuSentiment[1] contains almost 30 thousands social media posts labelled with 5 classes: *positive, negative, neutral, speech and skip*.
- Subset “random posts” is divided to train and validation in proportion 9/1.
- RuSentiment relates to colloquial style.

[1]Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., and Gribov, A. (2018). Rusentiment: An enriched sentiment analysis dataset for social media in russian. In Proceedings of the 27th International Conference on Computational Linguistics, pages 755–763.

Experimental setup

- Train language models of official, journalistic and colloquial styles.
- Fine-tune language model on RuSentiment.
- Cross-compare bi-gram language models on different domains.
- Train classifiers with fastText embeddings of official/journalistic and colloquial styles.
- Train classifiers with ELMo of three different styles.

Training and fine-tuning of LMs



Data	Training time	Epochs	Perplexity on valid	Perplexity on RuSentiment
Wiki	6 days	10	43.692	17364.89
WMT News	14 days	10	49.876	360.97
Twitter	10 days	10	94.145	172.25
Fine-tuning of Twitter on RuSentiment	15 min	4	159.2	–

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
https://www.slideshare.net/shuntaroy/a-review-of-deep-contextualized-word-representations-peters-2018

Training and fine-tuning of LMs

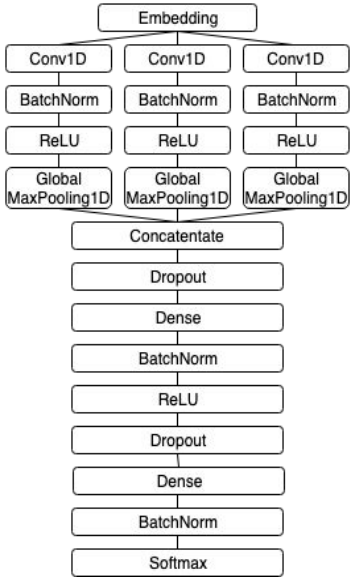
- Bi-gram KenLM[1] model is to predict conditional probability $\mathbb{P}(\omega_n|\omega_{n-1})$ of word ω_n given the preceding word ω_{n-1}.
- The resulting perplexities show how accurately a bi-gram model from one specific domain (rows) predicts words of test set from another specific domain (columns)

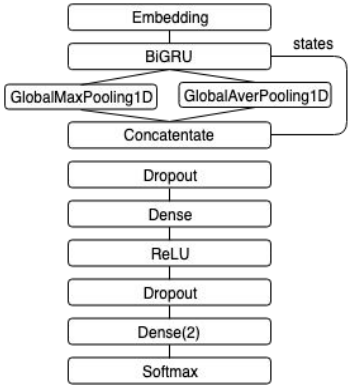
Bi-gram model\Data	RuSentiment	WMT News	Twitter	Wiki
RuSentiment	116.67	4847.68	**9094.83**	**7151.52**
WMT News	369864.24	640.55	434928.31	10381.87
Twitter	**46657.95**	1740.06	6762.07	8330.85
Wiki	189929.95	**1583.86**	197762.66	1586.13

[1]Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 690–696, Sofia, Bulgaria.

Training classifiers SWCNN[1] and BiGRU[2]

- fastText[3] embeddings trained on Russian Wiki and News corpora
- fastText[3] embeddings trained on Russian Twitter corpus
- ELMo trained on Russian WMT News dataset
- ELMo trained on Russian Wikipedia dataset
- ELMo trained on Russian Twitter dataset
- ELMo trained on Russian Twitter dataset and fine-tuned on RuSentiment





[1]Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
[2]Cho, K., Van Merri¨enboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoderdecoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
[2]Johnson, R. and Zhang, T. (2016). Supervised and semisupervised text categorization using lstm for region embeddings. arXiv preprint arXiv:1602.02373.
[3]Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Results

Model	Embeddings	Validation F1-weighted	Test F1-weighted
Rogers et al. [Rogers et al., 2018]	fastText VK	–	72.8
SWCNN	fastText Wiki+News	67.84	70.27
BiGRU	fastText Wiki+News	69.54	71.74
SWCNN	fastText Twitter	70.91	73.03
BiGRU	fastText Twitter	72.62	74.45
SWCNN	ELMo WMT News	70.27	72.42
BiGRU	ELMo WMT News	70.15	71.37
SWCNN	ELMo Wiki	68.11	71.28
BiGRU	ELMo Wiki	66.55	69.47
SWCNN	ELMo Twitter	75.40	**78.50**
BiGRU	ELMo Twitter	**75.89**	77.62
SWCNN	ELMo Fine-tuned	74.74	77.98
BiGRU	ELMo Fine-tuned	75.75	77.19

Examples

Text sample	True label	ELMo News	ELMo Wiki	ELMo Twitter
василий зе бест!	*positive*	skip	skip	*positive*
вкусняшка, омном-ном	*positive*	neutral	skip	*positive*
полнейший зашквар назначать некогда хорошего футболиста сразу главным тренером "реала"	*negative*	neutral	neutral	*negative*
я променяла вас на диплом! а ещё на министерское тестирование и гос экзамены!!я 0 числа уже с дипломом в зубах буду!!	*positive*	*positive*	skip	negative
все! завтра улетаю на евро- 0 в польшу болеть за сборную россии!	*positive*	*positive*	neutral	neutral
ну кто еще теперь задаст вопросы "зачем нами эта олимпиада?", "зачем нам спорт высоких достижений?". ведь можем же, когда захотим...	*neutral*	negative	*neutral*	negative

Conclusion

1. We have introduced pre-trained Russian language models of official, journalistic and conversational language styles.
2. We have trained two popular architectures on 6 different embeddings of different language styles.
3. Embeddings from language models are appropriate to be used in classification tasks if the domain of language model and target problem are close.

DeepPavlov

docs.deeppavlov.ai

Business solutions, support & partnerships

iPavlov.ai

This work was supported by National Technology Initiative and PAO Sberbank project ID 0000000007417F630002.