Sentiment Transfer for Russian Language

Mikhail Orzhenovskii

Dec 2020

Abstract

This document is the report for the Sentiment Transfer project.
`https://github.com/orzhan/russian-text-sentiment-transfer`

1 Introduction

Unsupervised sentiment transfer is changing the sentiment of the given text while preserving its original content without using pairs of parallel sentences. Sentiment transfer is a particular case of a wider problem - text style transfer.

One of the most popular approaches to sentiment transfer is removing original sentiment markers (words specific to a sentiment) from a sentence and then adding target sentiment markers.[Li et al., 2018] [Lee, 2020]

Proposed model is built based on this approach and makes use of a pre-trained language model [Brown et al., 2020] to generate fluent sentences. The model works with texts written in Russian.

2 Related Work

Many models use variations of delete and generate approach. [Li et al., 2018] additionally retrieve a sentence from corpus with similar meaning to source and target sentiment. [Xu et al., 2018] use reinforcement learning. [Sudhakar et al., 2019] use Transformer and pre-trained language model. [Lee, 2020] use classifier to delete original sentiment markers and use encoder-decoder to generate a new sentence. [Zhou et al., 2020] created attentional Seq2seq model and use word level style classifier.

There is also a method [Hu et al., 2018] of learning latent representations to separate style and content from sentences.

Style or sentiment transfer can also be seen as an unsupervised machine translation task. [Zhang et al., 2018] [Prabhumoye et al., 2018] Developing this approach, [Pant et al., 2020] incorporate sentiment based loss in the back-translation based style transfer.

One of possible extensions of the sentiment transfer task is fine-grained sentiment transfer. [Luo et al., 2019] [Xiao et al., 2020]

All of the works above were targeting text in other languages than Russian.

3 Model Description

Corpus of sentences $D = (x_1), ...(x_m)$, where x_i is a sentence, L - all possible sentences. $C(x_i)$ is content of sentence x_i, $S(x_i) \in V$ is true sentiment of sentence x_i (negative, positive and neutral). V={-1,1,0}.

We can't get true sentiment of a sentence except the ones from the corpus, but we can use a classifier to get some estimation:

$$\hat{S}(x) \in [-1, 1]$$

Also we can't extract real content of a sentence, but we can estimate the difference of content with a metric:

$$\hat{C}(x_1, x_2) \approx \|C(x_1) - C(x_2)\|$$

The sentiment transfer task now can be formalized as the following optimization task: for a given x and v find a sentence

$$x_v \in L : \begin{cases} \hat{C}(x, x_v) \to min \\ \hat{S}(x_v) \to v \end{cases}$$

As in many other implementations, the transfer is done in two stages.

During the first stage, text x is converted to neutral sentiment x_0 by deleting some of its tokens:

$$x_0 \in L : \begin{cases} \hat{C}(x, x_0) \to min \\ \hat{S}(x) \to 0 \end{cases}$$

On the second stage, language model is used to add the target sentiment to the neutral text x_0.

$$x_v in L : \begin{cases} \hat{C}(x_0, x_v) \to min \\ \hat{S}(x_v) \to v \end{cases}$$

Model structure is presented on Fig. 1.

3.1 Removing sentiment with FastText classifier

A FastText classifier[Joulin et al., 2016] is trained to determine the sentiment of a sentence. For each token in sentence we create a modified sentence without the token. Next, classification probability is calculated for each of those modified sentences. The modified sentence with the most neutral sentiment is chosen, so the most significant (in terms of sentiment) token is deleted or replaced with <unk>.

The process is repeated until sentence's classification probability drops below a certain level (parameter α). There is also a limit of how much of the sentence's tokens can be removed, this is regulated by parameter β. Tuning these parameters allows to choose between better sentiment removal and leaving more content unchanged.

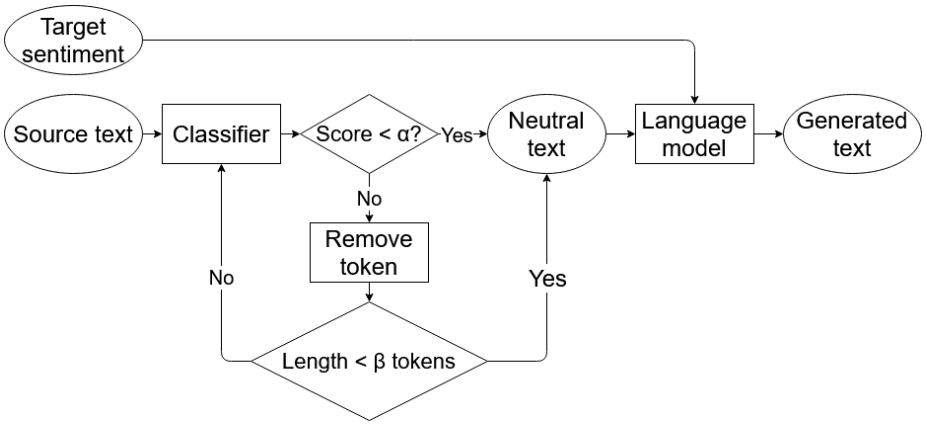

Figure 1: Model structure

3.2 Removing sentiment with BERT attention scores

BERT model[Devlin et al., 2019] is trained to determine the sentiment of a sentence. Then token attention scores are calculated as average attention value between all attention heads, from a token to [CLS] token, at the last layer. The token with the biggest attention score is deleted, until the sentence becomes neutral enough according to the classifier (parameter α) or maximum amount tokens is deleted (parameter β). The parameters play the same role as in FastText mode.

3.3 Generating sentences with target sentiment

To generate a text with specific sentiment from a neutral sentence we use GPT3 language model[Brown et al., 2020] (ruGPT3Large implementation by Sberbank AI). The model is fine-tuned with inputs in a special format:

$$x_0 \boxed{\text{SEP}} S(x) \boxed{\text{SEP}} x \boxed{\text{EOS}}$$

During inference, for source x and target sentiment v the following prompt is passed:

$$x_0 \boxed{\text{SEP}} v \boxed{\text{SEP}}$$

The output of the language model is expected to be x_v - the modification of original sentence with target sentiment.

4 Datasets

4.1 RuReviews

RuReviews: An Automatically Annotated Sentiment Analysis Dataset for Product Reviews in Russian presented in [Smetanin and Komarov, 2019] and can be obtained on GitHub[1]. This dataset contains 90000 reviews divided into three classes: with positive, neutral and negative sentiment.

For this dataset we were solving a task of converting negative sentences to positive and in opposite direction. Neutral sentences were not used. Also 2877 reviews written in English were detected and removed. See Table 1.

	Train	Validation	Test
Articles	57123	2000	2000
Positive	26540	1000	999
Negative	26583	1000	1101

Table 1: Statistics of the RuReviews dataset

4.2 Toxic Russian Comments

The dataset was introduced in OK ML Cup competition and is available on Kaggle[2]. It contains 248290 comments with 3 multiclass labels: insult, obscenity and threat. 203685 of the comments have none of those labels. See Table 2.

We used sentences with any of these labels as examples of Toxic class and other sentences as examples of Normal class. Normal examples were downsampled to match the count of the toxic examples.

	Train	Validation	Test
Articles	83210	2000	2000
Normal	41605	1000	1000
Toxic	41605	1000	1000

Table 2: Statistics of the Toxic Russian Comments dataset

5 Experiments

5.1 Metrics

Sentiment transfer task implies that original text content is preserved and only sentiment is changed. The best metric to use would be human judgement.

[1] https://github.com/sismetanin/rureviews

[2] https://www.kaggle.com/alexandersemiletov/toxic-russian-comments

However it is expensive and slow. Because of that we used automatic evaluation with the following metrics(having original sequence of texts $\{x_i\}$, transformed sequence $\{\hat{x}_i\}$ and target sentiment v):

- $BLEU(\{x_i\}, \{\hat{x}_i\})$ [Papineni et al., 2002] as content preservation metric
- Accuracy - fraction of target sentences having the target sentiment (calculated with BERT classifier)

$$ACC(\{\hat{x}_i\}, v) = \frac{|x \in \{\hat{x}_i\} : \hat{S}(x) = v|}{|\{\hat{x}_i\}|}$$

- Score computed using BLEU and accuracy - single metric to compare different models.

$$G(\{\hat{x}_i\}, \{x_i\}, v) = BLEU(\{x_i\}, \{\hat{x}_i\}) * ACC(\{\hat{x}_i\}, v)$$

5.2 Experiment Setup

The data was split into training, validation and test sets for the both datasets.

Training data was used to train FastText classifier and fine-tune BERT model and GPT model (separate models were trained for each dataset). Hyper-parameters a and β and hyper-parameters of the language model were picked based on validation performance. Results were obtained on the test set.

5.3 Baselines

A simple model was used as a baseline: Logistic Regression classifier was trained on TF/IDF of words and bigrams to detect sentiment of a text. For RuReviews dataset most significant 50 positive and top 50 negative strings were replaced with the manually edited opposing values (if applicable). For Toxic Russian dataset negative strings were deleted, and positive strings were replaced with random negative strings.

6 Results

Model	Accuracy	BLEU	Score
Baseline	0.06	0.99	0.06
FastText+LM	0.46	0.74	**0.34**
Attention+LM	0.44	0.57	0.26

Table 3: Results on RuReviews - Positive to Negative

The samples for RuReviews dataset could be found in Tab. 9. The samples for Russian Toxic dataset could be found in Tab. 10.

Model	Accuracy	BLEU	Score
Baseline	0.04	0.99	0.04
FastText+LM	0.34	0.70	**0.24**
Attention+LM	0.26	0.51	0.13

Table 4: Results on RuReviews - Negative to Positive

Model	Accuracy	BLEU	Score
Baseline	0.06	0.99	0.06
FastText+LM	0.40	0.72	**0.29**
Attention+LM	0.35	0.54	0.19

Table 5: Results on RuReviews - Overall

Model	Accuracy	BLEU	Score
Baseline	0.09	0.98	0.09
FastText+LM	0.71	0.52	**0.37**
Attention+LM	0.37	0.34	0.12

Table 6: Results on Russian Toxic - Normal to Toxic

Model	Accuracy	BLEU	Score
Baseline	0.02	0.97	0.02
FastText+LM	0.50	0.60	**0.30**
Attention+LM	0.61	0.33	0.20

Table 7: Results on Russian Toxic - Toxic to Normal

Model	Accuracy	BLEU	Score
Baseline	0.06	0.98	0.06
FastText+LM	0.61	0.56	**0.34**
Attention+LM	0.49	0.33	0.16

Table 8: Results on Russian Toxic - Overall

7 Conclusion

Two datasets were prepared for the task. Own model was implemented and displayed better results compared to the simple baseline. Model with BERT based token removal performed worse than Fasttext based removal. The model got higher score transforming from positive to negative sentiment than in the opposite direction. Metric values do not differ significantly between the datasets.

Positive to negative
Кофта приятная, цвет супер →
Кофта приятная, пришла с дырой
товар соответствует описанию в Ростов пришёл за 3 недели →
Товар не соответствует описанию в Ростов пришёл не 3 недели
Negative to positive
Размер вообще не совпадает, на куклу одежда, даже на S размер не тянет →
Все как в описании, все очень хорошо совпадает, на русский s одежда
кофточка пришла не по размеру заказывала xl а как будто s →
кофта пришла по размеру подошла xl а как просила s

Table 9: Output samples on RuReviews dataset

Normal to toxic
праведно зделоли пусть сидит мы его встретем →
праведно зделоли пусть сидит сволочь его встретем
наверное позора не вынесут и покинут россию навсегда! →
сраные суки позора не вынесут и покинут россию навсегда!
Toxic to normal
сталина на них нет. расстрелять и все →
да, да! сталина на них нет. как и все
дебилом не рождаются, а ими стоновятся. потому что в башке пусто!!!!! →
взрослые не рождаются, а ими стоновятся. потому что в голове пусто!!!!!

Table 10: Output samples on Russian Toxic dataset

References

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Hu et al., 2018] Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. (2018). Toward controlled generation of text.

[Joulin et al., 2016] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[Lee, 2020] Lee, J. (2020). Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 195–204, Dublin, Ireland. Association for Computational Linguistics.

[Li et al., 2018] Li, J., Jia, R., He, H., and Liang, P. (2018). Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

[Luo et al., 2019] Luo, F., Li, P., Yang, P., Zhou, J., Tan, Y., Chang, B., Sui, Z., and Sun, X. (2019). Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2022, Florence, Italy. Association for Computational Linguistics.

[Pant et al., 2020] Pant, K., Verma, Y., and Mamidi, R. (2020). Sentiinc: Incorporating sentiment information into sentiment transfer without parallel data. In Jose, J. M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M. J., and Martins, F., editors, *Advances in Information Retrieval*, pages 312–319, Cham. Springer International Publishing.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[Prabhumoye et al., 2018] Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. (2018). Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

[Smetanin and Komarov, 2019] Smetanin, S. and Komarov, M. (2019). Sentiment analysis of product reviews in russian using convolutional neural networks. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, pages 482–486.

[Sudhakar et al., 2019] Sudhakar, A., Upadhyay, B., and Maheswaran, A. (2019). Transforming delete, retrieve, generate approach for controlled text style transfer.

[Xiao et al., 2020] Xiao, L., Qu, X., Li, R., Wang, J., Zhou, P., and Li, Y. (2020). Fine-grained text sentiment transfer via dependency parsing. In Giacomo, G. D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., and Lang, J., editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2228–2235. IOS Press.

[Xu et al., 2018] Xu, J., Sun, X., Zeng, Q., Zhang, X., Ren, X., Wang, H., and Li, W. (2018). Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.

[Zhang et al., 2018] Zhang, Z., Ren, S., Liu, S., Wang, J., Chen, P., Li, M., Zhou, M., and Chen, E. (2018). Style transfer as unsupervised machine translation. *CoRR*, abs/1808.07894.

[Zhou et al., 2020] Zhou, C., Chen, L., Liu, J., Xiao, X., Su, J., Guo, S., and Wu, H. (2020). Exploring contextual word-level style relevance for unsupervised style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7135–7144, Online. Association for Computational Linguistics.