Large Dataset and Language Model Fun-Tuning for Humor Recognition

Vladislav Blinov[1,3], Valeriia Bolotova-Baranova[1,3], and Pavel Braslavski[1,2]

[1]Ural Federal University, Yekaterinburg, Russia
[2]Higher School of Economics, Saint Petersburg, Russia
[3]Tinkoff.ru
{vladislav.blinov,pavel.braslavsky}@urfu.ru, lurunchik@gmail.com

Abstract

The task of humor recognition has attracted a lot of attention recently due to the urge to process large amounts of user-generated texts and rise of conversational agents. We collected a dataset of jokes and funny dialogues in Russian from various online resources and complemented them carefully with unfunny texts with similar lexical properties. The dataset comprises of more than 300,000 short texts, which is significantly larger than any previous humor-related corpus. Manual annotation of about 2,000 items proved the reliability of the corpus construction approach. Further, we applied language model fine-tuning for text classification and obtained an F1 score of 0.91 on test set, which constitutes a considerable gain over baseline methods. The dataset is freely available for research community.

1 Introduction

Humor is an important element of everyday human communication (Martin, 2007). With a rapid development of conversational systems and NLP applications for social media content, the task of automatic recognition of humor and other types of figurative language has gained a lot of attention (Nijholt et al., 2017). Standard publicly available datasets significantly contribute to steady and measurable progress in solving NLP tasks. To the date, there are several humor-related datasets, but the majority of them contain English texts only, are relatively small, and focus predominantly on puns, thus don't reflect a wide variety of humorous content.

In this work we describe the creation of a large dataset of funny short texts in Russian. We started with an existing dataset and more than tripled it in size. The texts were automatically collected from various online sources to ensure their diversity and representativeness. A separate task was the compilation of a contrasting corpus of unfunny texts in such a way that their distinguishing characteristic was absence of humor, and not their lexical properties and style. The dataset comprises of more than 300,000 short texts in total, about half of them being funny. Manual annotation of 1,877 examples confirmed the validity of the automatic approach and formed a golden test set.

We implemented a humor detection method based on the universal language model fine-tuning. Unlike most approaches to humor recognition described in the literature, this method neither draws upon an existing theory of humor, nor makes explicit assumptions about the structure and 'mechanics' of jokes; it needs no feature engineering and is purely data-driven. This approach is justified in the case of a large heterogeneous collection. Evaluation of the trained model on several test collections of Russian jokes shows that it has not been overfitted and generalizes well.

The compiled dataset publicly available[1]. We hope that the resource will intensify research on multilingual computational humor.

2 Related Work

Humor recognition is usually formulated as a classification problem with a wide variety of features – syntactic parsing, alliteration and rhyme, antonymy and other WordNet relations, dictionaries of slang and sexually explicit words, polarity and subjectivity lexicons, distances between words in terms of *word2vec* representations, word association measures, etc. (Taylor and Mazlack, 2004; Mihalcea and Strapparava, 2005; Kiddon and Brun, 2011; Yang et al., 2015; Zhang and Liu, 2014; Liu et al., 2018; Cattle and Ma, 2018; Ermilov et al., 2018). A cognate task is humor ranking (Shahaf et al., 2015; Potash et al., 2017). Features engineered for classification/ranking are of-

[1] https://github.com/computational-humor/humor-recognition/tree/master/data

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4027–4032
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

ten inspired by linguistic theories of humor, see a survey in (Attardo, 1994). Most recent studies (Yang et al., 2015; Liu et al., 2018; Cattle and Ma, 2018) employ Random Forest classifiers for humor recognition and word embeddings as feature vectors. At the moment, there are a few studies that use neural architectures to directly address humor recognition: Ortega-Bueno et al. (2018) and Hasan et al. (2019) exploit LSTM, while Chen and Soo (2018) use CNN architecture.

The dataset collected by Mihalcea and Strapparava (2005) became a *de facto* standard for humor recognition. It contains 16,000 one-liners and 16,000 non-humorous sentences from news titles, proverbs, British National Corpus, and Open Mind Common Sense collection. Another dataset used in several studies (Yang et al., 2015; Cattle and Ma, 2018) comprises of 2,400 puns and an equal number of negative samples from the news, Yahoo!Answers, and proverbs. In both cases authors tried to ensure lexical and structural similarity between the humorous and 'serious' classes. Two datasets were prepared within SemEval 2017 shared tasks: *#HashtagWars* (Potash et al., 2017) and *English Puns* (Miller et al., 2017). The former dataset comprises of 12,000 tweets corresponding to about 100 episodes of a TV show, each annotated with a 3-point funniness score. The latter one contains about 4,000 contexts, 71% of which are puns, annotated with WordNet senses. Most of humor recognition research deals with English; exceptions are studies working with Italian (Reyes et al., 2009), Russian (Ermilov et al., 2018), and Spanish (Castro et al., 2018).

3 Data

STIERLITZ and PUNS. We started with a dataset of Russian one-liners and non-humorous texts used previously in (Ermilov et al., 2018). The balanced dataset was assembled by complementing a collection of jokes from social media (Bolotova et al., 2017) with non-humorous proverbs, news headlines, and sentences from fiction books. Following the authors, we refer to the dataset as STIERLITZ.[2] We also use a small collection of Russian puns from (Ermilov et al., 2018) for evaluation. Puns as a special type of humor seem to be less articulated in the Russian culture compared to British/US tradition. The authors were able to spot only few online collections of puns.

Dataset	*Jokes*	*Non-jokes*	*Total*
STIERLITZ	46,608	46,608	**93,216**
train	37,447	37,447	65,530
validation	4,682	4,682	9,364
test	9,361	9,361	18,722
PUNS	213	0	**213**
FUN	156,605	156,605	**313,210**
train	125,708	125,708	251,416
test	30,897	30,897	61,794
GOLD	899	978	**1,877**

Table 1: Datasets for humor recognition.

FUN: dataset expansion. Our goal was to significantly expand STIERLITZ and to ensure that funny/serious counterparts are more similar in terms of vocabulary, style, and structure than in the original collection.

First, we collected more than 1M jokes from multiple humorous public pages from the largest Russian social network *VK.com* through its API (556K) and from the website *anekdot.ru* (477K), the oldest online resource of humorous content on the Russian Web.

Then, we filtered out less popular jokes based on user ratings, duplicates, and jokes already presented in STIERLITZ and PUNS collections. The newly obtained collection is quite diverse: it contains one-liners, multi-turn jokes, and short sketches.

Second, we downloaded 10M posts from a large online forum of the city portal of Yekaterinburg *E1.ru*[3]. We opted for online forums as a source of negative examples, since social media and human conversations are immediate application domains of humor recognition. We indexed the forum data with Elastic[4] and returned a BM25-ranked list of matching forum posts for each joke. To filter out potential occurrences of jokes in the forum data, we removed all forum snippets with Jaccard similarity higher than 0.4 to the query joke. This threshold was inferred empirically from the data. After that, we added the highest-ranked post for each joke to the collection. Here is an example of such a joke/non-joke pair (hereafter, we cite English translations of original texts in Russian):

[2] Stierlitz is a protagonist of a popular TV series, a Soviet spy working undercover in Nazi Germany. He is also a popular character of jokes in post-Soviet countries.

[3] https://www.e1.ru/talk/forum/

[4] https://www.elastic.co/

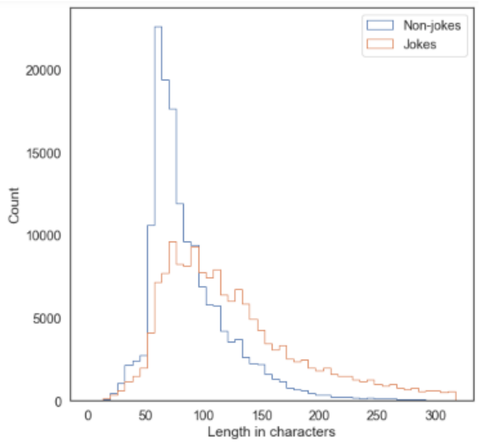

Figure 1: Item length distributions for jokes/non-jokes in the FUN dataset.

FUN*: Russian Mars rover will hand out Russian passports to the Martians.*
FORUM*: They say in the Crimea, too, they are handing out Russian passports or have already handed them out.*

To assess lexical diversity of the resulted dataset, we calculated KL-divergence of add-one smoothed word frequency distributions of non-jokes with respect to jokes for both STIERLITZ and FUN. The resulted in 0.18 for FUN and 0.50 for STIERLITZ, which demonstrates that joke/non-joke classes in the new dataset are lexically more similar than in the STIERLITZ dataset. We also examined words with most frequency disproportions in funny/non-funny parts of the dataset. Local toponyms, abbreviations, and non-standard spellings typical for online communications appeared to be the most salient words in the 'unfunny' part of the dataset that stems from an online forum, while names of some jokes characters and *mat* (Russian profane language) are most typical for the funny part.

When compiling the dataset we introduced high/low cut-off thresholds for text lengths, but didn't try to balance out lengths distributions of jokes/non-jokes subsets. Figure 1 shows that the jokes' length distribution is skewed towards longer texts compared to non-jokes.

We also removed URLs and user names, retaining only unique entries. Finally, we partitioned the dataset into train/test sets (80:20) ensuring the original STIERLITZ train and validation subsets belong to FUN train and test subsets, respectively.

GOLD: dataset validation. To verify that our automatically created collection contains valid jokes and non-jokes, we conducted an evaluation using an online interface, where 1,000 random jokes and 1,000 random non-jokes were assessed on a 3-point scale: 'not a joke', 'an unfunny joke' and 'a joke'. We were able to recruit more than 100 volunteers through online social networks; evaluation resulted in 1,877 examples being labeled by at least three assessors. In case of 238 items (12.7%) we could observe opposite assessments, i.e. 'not a joke' and 'a joke', which is an acceptable agreement for a crowdsourcing setting. Majority voting resulted in 94% of non-jokes marked as 'not a joke' and 95% of jokes marked as either 'an unfunny joke' or 'a joke', which demonstrates a good performance of the automatic procedure.[5] The errors in the non-jokes are mostly humorous responses from the forum users, for example:

Invite a girl, cook a dinner for two... but do not ask "how to get rid of a girlfriend?" a week later.

Texts from humorous sources marked as 'not a joke' are examples of dry humor or context-dependent jokes, e.g.:

Ten to the power of thirty of electrons is almost a kilogram.

Table 1 summarizes statistics of the datasets used in the study.

4 Classification Methods

Recently, various neural network architectures have achieved state-of-the-art results in many areas of natural language processing. Given that we have a large enough corpus, we opted for universal language model fine-tuning method (ULMFiT) for text classification (Howard and Ruder, 2018) that has demonstrated good performance and generalization capabilities.

In case of humor recognition, it is desirable to model deeper word and context dependencies, as humorous effect is usually enabled by combinations of words rather than individual words themselves. Language models (LMs) have been used as baselines in several humor recognition studies (Shahaf et al., 2015; Yang et al., 2015; Cattle and Ma, 2018). In contrast to most previous humor recognition studies, we didn't engineer any

[5] For example, manual verification of the dataset in (Mihalcea and Strapparava, 2005) revealed 9% of noise.

linguistic features. However, LM-based approach can be seen as indirect reflection of some common humor features such as incongruity, unexpectedness, or nonsense.

Our humor corpora are relatively small compared to the corpora that are used to train language models. To overcome this limitation, we first trained a language model on 10M online forum texts for 15 epochs. Texts were tokenized using unigram subword tokenization method implemented in SentencePiece library (Kudo and Richardson, 2018) with the vocabulary size of 100,000. Architecture and parameters were directly transferred from (Howard and Ruder, 2018). Further, we used either Stierlitz or Fun dataset to fine-tune the model for five epochs. Finally, we replaced the target task with humor classifier by augmenting the model with linear blocks and trained the model with gradual unfreezing followed by 14 consecutive epochs. We further refer to this model as ULMFun.

As a baseline classification method, we chose an SVM classifier on top of *tf.idf* features, which is usually a good starting point in text classification tasks. In addition, the authors of (Ermilov et al., 2018) kindly agreed to run their best learned model on our new dataset.

5 Results and Discussion

The goals of the experiment were to estimate the impact of the increased dataset size and its construction methods, to introduce a strong baseline based on deep neural network approach, to compare it with a baseline and published work, as well as to evaluate generalization abilities of the model.

In the first series of experiments we trained a linear SVM baseline on *tf.idf* features and ULMFun model on Stierlitz train set. We tested the obtained models on held-out test sets of Stierlitz and Fun, as well as on smaller manually annotated Gold and Puns collections. In addition, we were able to apply the best model from (Ermilov et al., 2018) to the test data. Table 2 summarizes performance of the models. What stands out from the results is that baseline SVM outperforms a previous feature-rich approach (Ermilov et al., 2018). Due to high lexical diversity between positive and negative classes in Stierlitz, it seems to be trivial to distinguish between jokes and non-jokes with lexical features only. Even a simple linear model achieves F1 score of 0.91. Unsurprisingly, ULMFun outperforms both Stierlitz SVM and the baseline. Since Fun was constructed in quite a different way compared to Stierlitz and represents a much harder task, classification quality of all three methods decreases on Fun test set. For instance, Stierlitz SVM achieves only 23% recall on Fun non-jokes. It is interesting to note that more versatile Stierlitz SVM features demonstrate better transferability and help the method to beat the baseline on ‘unfamiliar’ Fun. Performance of the baseline is stable on Gold, while the other two classifiers’ scores decrease more significantly than one can expect based on manual verification results, i.e. by 5-6%. Variance of recall scores of the three methods on Puns is much higher, though the results must be treated with caution due to small size of the collection.

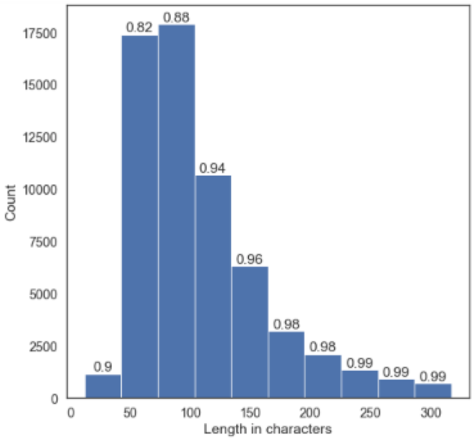

Figure 2: F1-scores for humor class depending on the text length in FUN test set.

Table 3 represents results of baseline SVM model and ULMFun trained on Fun training set. As expected, more data significantly improve classification quality on Fun test set in case of both methods. However, performance on presumably ‘simpler’ Stierlitz test set drops since Fun dataset is a lot more diverse in terms of joke types and topics. Performance of both methods on Gold decreases less than by 5% of noise expected in the data.

Figure 2 shows that the lowest humor detection quality is observed for texts in the range from 50 to 100 characters, which can be explained by the imbalance of the dataset in regard of length. Moreover, longer jokes are easier to detect due to a richer context. Manual inspection suggests that misclassified jokes can be divided into three categories. The most common

Model	STIERLITZ *Test F1*	FUN *Test F1*	GOLD *F1*	PUNS *Recall*
Baseline SVM	0.910	0.677	0.643	0.725
Stierlitz SVM (Ermilov et al., 2018)	0.884	0.735	0.638	0.695
ULMFun	0.965	0.768	0.662	0.920

Table 2: Humor detection quality – models trained on STIERLITZ train.

Model	STIERLITZ *Test F1*	FUN *Test F1*	GOLD *F1*	PUNS *Recall*
Baseline SVM	0.787	0.798	0.803	0.436
ULMFun	0.921	0.907	0.890	0.892

Table 3: Humor detection quality – models trained on FUN train.

one is jokes whose comprehension requires external world knowledge, for example:

The absolute record in worldwide compact disk sales was set by a little-known band called CD-R with its new single 700MB.

The following examples demonstrate two other error types – hard to get jokes, e.g.

No GMO, no artificial dyes, no plans for the future, no meaning in life, and no preservatives.

and noisy non-jokes from the positive class:

Would you like to celebrate your birthday in Las Vegas?

Similarly, misclassified examples from negative class are occasionally present noisy jokes:

No doctor is as worried about the patient's high heart beat rate as a pathologist.

ULMFun also triggers on context changes that are typical for many jokes, for example:

Model of an ideal person – and an out-of-class fire-breathing dragon!

6 Conclusion and Future Work

In this paper, we introduced a publicly available dataset for humor recognition in Russian that exceeds in size all previous public datasets. We compared the performance of a baseline SVM method and a more sophisticated ULMFiT method on this dataset, with the latter yielding favorable results. In the future, we aim to analyze how changes in the training procedure and hyperparameters of ULMFiT affect resulting model performance. On top of that, we hope to improve model generalization by augmenting negative examples with a split of jokes into setups and punchlines, as they should not be funny by themselves. We also plan to reproduce the experiment on English data.

Acknowledgments

We thank volunteers who took part in data annotation. We are also thanksful to authors of (Ermilov et al., 2018), who ran their code on our data.

The authors acknowledge support from the Ministry of Education and Science of the Russian Federation (project no. 1.3253.2017) and the Competitiveness Program of the Ural Federal University.

References

Salvatore Attardo. 1994. *Linguistic Theories of Humor.* Walter de Gruyter.

Valeria Bolotova, Vladislav Blinov, Kirill Mishchenko, and Pavel Braslavski. 2017. Which IR model has a better sense of humor? Search over a large collection of jokes. In *Proceedings of the Dialogue Conference*, pages 29–42.

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018. A crowd-annotated Spanish corpus for humor analysis. In *Proceedings of SocialNLP Workshop*.

Andrew Cattle and Xiaojuan Ma. 2018. Recognizing humour using word associations and humour anchor extraction. In *COLING*, pages 1849–1858.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *NAACL-HLT*, pages 113–117.

Anton Ermilov, Natasha Murashkina, Valeria Goryacheva, and Pavel Braslavski. 2018. Stierlitz Meets SVM: Humor Detection in Russian. In *Artificial Intelligence and Natural Language*, pages 178–184.

Md. Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md. Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed E. Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *CoRR*, abs/1904.06618.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL*, pages 328–339.

Chloe Kiddon and Yuriy Brun. 2011. That's what she said: double entendre identification. In *ACL-HLT*, pages 89–94.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *EMNLP: System Demonstrations*, pages 66–71.

Lizhen Liu, Donghai Zhang, and Wei Song. 2018. Exploiting syntactic structures for humor recognition. In *COLING*, pages 1875–1883.

Rod A. Martin. 2007. *The Psychology of Humor: An Integrative Approach*. Elsevier.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *HLT-EMNLP*, pages 531–538.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 Task 7: Detection and Interpretation of English Puns. In *Proceedings of the SemEval Workshop*, pages 58–68.

Anton Nijholt, Andreea Niculescu, Alessandro Valitutti, and Rafael E. Banchs. 2017. Humor in human-computer interaction: A short survey. In *Adjunct Conference Proceedings INTERACT*, pages 192–214.

Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. UO UPV: Deep Linguistic Humor Detection in Spanish Social Media. In *Proceedings of the IberEval Workshop)*.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 Task 6: #HashtagWars: Learning a Sense of Humor. In *Proceedings of the SemEval Workshop*, pages 49–57.

Antonio Reyes, Davide Buscaldi, and Paolo Rosso. 2009. An analysis of the impact of ambiguity on automatic humour recognition. In *Text, Speech and Dialogue*, pages 162–169.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *PKDD*, pages 1065–1074.

Julia M Taylor and Lawrence J Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 1315–1320.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *EMNLP*, pages 2367–2376.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on Twitter. In *CIKM*, pages 889–898.