Short and long-range comedy generation and understanding using Large Language Models

Edison Marrese-Taylor[1, 2], Machel Reid[3], Alfredo Solano[2]
National Institute of Advanced Industrial Science and Technology[1]
Graduate School of Engineering, The University of Tokyo[2]
Google DeepMind[3]
edison.marrese@aist.go.jp, machelreid@google.com
asolano@weblab.t.u-tokyo.ac.jp

Abstract

We study the automatic detection and generation of humorous and ironic text, both in short and long range scenarios. For the former, we propose a style-transfer approach, which we utilize to generate humorous news headlines by exploiting a combination of classification and generative models based on medium-sized language models. For the latter, we introduce a new dataset of full stand-up comedy special scripts, which we use as an arena to generate and classify humorous content using LLMs.

1 Introduction

Humor is admittedly an important part of communication. From teasing, yet harmless jokes made between friends to the sprinkling of humor in a long presentation to maintain the audience' s attention, humor can be seen ubiquitously in a large variety of settings, times, and cultures. Advances in machine learning and natural language processing have contributed to the ability of machines to detect humor, but this remains an open and challenging problem.

The linguistic study of humor started as early as the classical times [1]. Several theories have been proposed to describe and explain humor, with work on this subject tracing back to Aristotle, and extending to various disciplines such as semantics, psychology, and linguistics [2, 1, 3].

Furthermore, humor is also related to irony and sarcasm, two closely related linguistic phenomena that encompass the concept of meaning the opposite of what is literally expressed. There is no consensus in academic research on the formal definition, both terms being non-static, depending on different factors such as context, domain and even region in some cases. Humor is often seen as an umbrella term for many such phenomena [4], with the difficulties in understanding them lying in the ability of models in capturing linguistic nuances, context-dependencies and latent meaning, due to the richness of dynamic variants and figurative use of language [5].

The rise of Large Language Models (LLMs) has lately led to significant advancements in text generation. The development of conversational AI based on such models has made computational humor methods highly relevant and in demand for practical applications [4]. In light of this, this paper studies the automatic detection and generation of humorous and ironic text for both short and long-term forms.

For the former, we propose a method based on style-transfer utilizing models with millions of parameters, while for the latter, we propose a technique which generates stand-up comedy transcripts given a prompt, based on language models with billions of parameters. Our results show that although it is possible for smaller language models to generate short humorous text in the form of news headlines, we observe inherent limitations which make applicability limited. In turn, our long-range humor generation experiments suggest that larger language models offer a valid mechanism to generate and classify long humorous content, although, again, with clear limitations.

2 Related Work

Sarcasm Detection Work on automatically detecting sarcasm and irony is extensive and goes back to rule-based systems [6]. Statistical methods and classic machine learning algorithms such as Support Vector Machines [5, 7], Naive Bayes and Decision Trees [8] were also utilized for this task. Also, [9] proposed a model which uses an intra-attentional component in addition to an RNN. A comprehensive survey on automatic sarcasm detection was done by [10], while computational irony detection was reviewed by [11].

Humor Detection Work on humor detection first relates to the task of joke identification, where we find a large set of approaches have been employed over the years, including classical ones like regression trees, as well as more recent deep learning models like CNNs. The approach of [12] recently showed that by training a ROBERTa-based [13] classifier on jokes gathered from multiple sources, each covering a different type of humor, one can obtain a robust classifier. Some of these claims were recently discussed by [4], who showed that the out-of-domain performance of such models is unstable, suggesting that they may learn non-specific humor characteristics, while instead showing that LLMs demonstrate competitive results and a more stable behavior in this regard

3 Proposed Approach

Short-range humor understanding Previous work on humor generation and humor classification has mainly focused on short-range text. Table 1 shows a summary of well-known datasets in the area. From these datasets, in this paper we focus specifically on HUMICROEDITS [16], a dataset that consists of regular English news headlines paired with versions of the same headlines that contain simple replacement edits designed to make them funny. Our choice is explained by the fact that HUMICROEDITS allows us to directly study the role of humor in text generation, as other confounding variablesn such as the context in which the humorous text is to be produced, are controlled for in this scenario.

To tackle this task, we propose an approach based on text style transfer. In this setting, a model changes the style of a source text into a target style, while otherwise changing as little as possible about the input. While so far this approach has mainly been used to tackle domains such as sentiment and politeness, in this work we follow [18] and apply their ideas to humor and irony generation, regarding plain text as belonging to the "source" style, and humorous text as the "target" style.

Concretely, we adapt LEWIS [19] a state-of-the-art edit-based generation model for style transfer. This approach works by generating synthetic data via a domain-specific pretrained conditional language model, to have a parallel corpus that can be used to learn how to transform sentences from one style into the other. Critically, the generation of pseudo-parallel data is derived from an attentive-style (humor) classifier, from which style-agnostic templates can be extracted based on the values of the attention. This means that having a robust humor classifier is of high importance for generation, which further motivates our adoption of LEWIS as a means to study this phenomenon. Our choice of model contrasts with [18], who directly relied on training a Transformer-based sequence-to-sequence model on the data but unfortunately did not offer a performance study via automatic evaluation, thus providing little insight on how the generation is performed.

For this part, we follow [12] and train a RoBERTa-based classifier, an inherently attention-based model, for humor on WELLER (PUNS + SHORT JOKES + REDDIT HUMOR FULL), denoted as M_W, and another equivalent model solely on HUMICROEDITS (M_H). We also consider a sarcasm model, training on SARC 2.0 (M_S). We choose to train on this dataset as its contents most adequately match our evaluation regime— news headlines. As seen in Table 1 other irony/sarcasm dataset contains data that differs significantly in nature from our domain.

Table 2 summarizes the results of our humor and sarcasm classification training efforts. We see that our humor model trained on WELLER obtains a classification accuracy of 98%, which compares favorably against the value of 93% reported by [12]. On the other hand, our sarcasm detection model report a 74.1% accuracy on SARC 2.0, which is competitive with the baseline model from [20], which reports a value of 75.8%, and with the model by [21] who obtained an accuracy of 77.3%. We point out that while these models offer superior performance compared to our RoBERTa model, we are unfortunately unable to use them for our experiments with LEWIS as they do not rely on attention.

We also see that classifiers trained on the concatenated data, following [12] do not perform well across datasets, even across the training portions. One key point here is that M_W obtains a performance of 50.5% on HUMICROEDITS, which contrasts to an accuracy of 87.4% by M_H. This

Table 1 Summary datasets on Humor and Irony/Sarcasm detection relevant to our work. In the table, "Human★" means that this data was provided or annotated directly by humans, "Self" means that the annotations are automatically derived from the context (e.g. presence of a hashtag on Twitter, or extracted from a source such as a subreddit), "MTurk" indicates that the data was annotated by means of crowdsourcing via Amazon Mechanical Turk.

Task	Dataset	Size	Mean Length	Source	Annotations
Humor	Puns [14]	4,827	14 ± 5	punoftheday.com	Human★
	Short Jokes [15]	405,400	22 ± 12	Kaggle/WMT	Human★
	Reddit Humor [12]	20,046	72 ± 122	Reddit	Self
	Weller [12]	429,668	25 ± 31	Various	Various
	Humicroedits [16]	40,638	15 ± 5	Reddit	MTurk
	Scraps from the loft (ours)	416	11,654 ± 4,441	Blog	Self
Irony	SARC 2.0 [17]	321,748	30 ± 18	Reddit	Self

Table 2 Out-of-domain performance of our trained humor/sarcasm classifiers based on RoBERTa. Results are reported on the test portions of each dataset.

Dataset	Accuracy		
	M_W	M_H	M_S
Puns	94.7	68.2	50.6
Short Jokes	98.8	60.6	56.2
Reddit Humor	61.8	51.0	44.1
Weller	98.0	60.5	56.0
Humicroedits	50.5	87.4	49.7
SARC 2.0	49.6	49.0	74.1

Table 3 Summary of our results on comedy generation based on our style-transfer using LEWIS. In the table, LEWIS_H and LEWIS_S denote the models trained on humor and irony data, respectively and LEWIS is the original model [19]. Hum., Yelp and Polite denote the dataset by [16], [25] and [26], respectively. These last two values are taken directly from [19].

Model	Data	BLEU	S-BLEU	BScore	S-BScore
LEWIS_H	Hum.	21.4	27.7	0.431	0.360
LEWIS_S	Hum.	47.8	60.6	0.649	0.770
LEWIS	Yelp	24.0	58.5	50.0	72.2
LEWIS	Polite	-	75.3	-	81.4

shows that while RoBERTa is able to adequately model regularities in this dataset by directly finetuning on it, the model remains unable to perform the task when finetuning on Weller. This provides further evidence supporting claims by [4], that these techniques lead to limited generalization capabilities.

Once the style-agnostic templates have been extracted for each dataset, they are fed to the denoising models for each domain to generate synthetic parallel data. As denoising models we utilize the pretrained BART [22], and a version finetuned on our data — Weller and SARC 2.0 for Humor and Sarcasm, respectively. Finally, we train a sequence-to-sequence model on the synthetic parallel data, learning to map original sentences to their in-filled templates. For this final step, we again rely on pre-trained BART, which we finetune in our data.

For evaluation of the humor generation model, we follow previous work [19] and utilize BLEU [23] and BERTScore [24] measured against the reference "target" sentence to evaluate lexical overlap with human annotation. In addition to this, we measure Self-BLEU and Self-BERTScore, meaning we compare our generated sentence against the source, to measure content preservation.

Table 3 summarizes the results we obtained on the humorous news headline generation using our style-transfer approach. Our final LEWIS models are trained on a synthetic parallel corpus of 1,731,589 and 314,352 examples for humor and sarcasm respectively. As shown on the table, we see that models are able to obtain reasonable performance, with BLEU and BERTScore comparable to those obtained by this model on style transfer for sentiment and politeness, showing the effectiveness of our approach. We also specifically see that our sarcasm-based model is able to outperform the humor-based model, with a higher lexical overlap with the source headline. Overall, this suggests that headlines on our dataset are more aligned with the specific phenomena of sarcasm.

Long-range humor understanding We find that one major shortcoming of the approach presented above is its innate inability to handle long contexts when generating humorous text. [5] argues that indeed the difficulties in understanding and generating humor are due to context-dependencies and latent meaning, due to dynamic variants and figurative use of language. Following this line of argument, we believe that understanding and generating such short text may be of limited interest.

In light of this issue, we turn our interest to a different setting: stand-up comedy. Stand-up is a comedic performance where a person addresses a live audience directly from the stage. Our interest in this style of performance spawns from

Table 4 Performance on our long-range comedy generation and classification on Scraps from the loft via our LLM-based approach. In the Table, P, R and F1 are short for Precision, Recall and F1-Score.

Model	BLEU	BERTScore	P	R	F1
GPT2 (1.5B)	-	-	-	-	-
+ finetuning	15.9	0.815	0.209	0.468	0.273
LLama 2 (7B)	2.8	0.793	-	-	-
+ finetuning	14.1	0.809	0.117	0.357	0.165
Llama 2 (13B)	6.1	0.796	-	-	-
+ finetuning	13/8	0.807	0.219	0.546	0.301

the fact that stand-up usually consists of a wide variety of humor variations, including but not limited to one-liners, stories and general observations. To collect stand-up comedy transcripts, we rely on `scrapsfromtheloft.com`, an international magazine that focuses on entertainment and pop culture, offering reviews and essays as well as stand-up transcripts. From this website, we obtain full transcripts of 416 stand-up comedy specials. For each transcript, we collect its *title* and the name of the *comedian(s)*. We utilize 340 scripts for training, and 76 for testing.

To tackle long-range humor generation, we adopt an LLM-based approach in a chat-based scenario. Concretely, we propose to generate comedic content by instructing/prompting such models to do so. To create suitable prompts that can direct a given model to generate comedic content, we first propose to automatically obtain summarized versions of each stand-up transcript, as well as an explanation of the intent of the performer. The idea is to create instructions that contain the critical elements that need to be accounted for when generating. Once we have obtained these summaries, we use an LLM to obtain a prompt that can can successfully capture the details of the required output.

For the above steps, in this paper we rely on ChatGPT (*gpt3.5-turbo-16k*) and proceed as follows. First we ask the model to, given a full comedy transcript, generate a summary and the intent of the comedian with the prompt: *"$\{transcript\}$ Summarize the above and tell me what the intent of the comedian is within three lines."*. Once we have obtained the summarized content, and in the same chat session, we ask the model to generate suitable instructions using the prompt *"Now rewrite the above as an instruction you would give a comedian to reproduce the routine"*.

For the empirical study, we consider the recently-released LLama 2 models [27], specifically the instruction-tuned versions. We utilize the 7B and 13B models, which we quantize to 4-bits and finetune using QLoRA [28] in order to fit our GPU memory. Each model is finetuned directly on the training portion of our data, where the input to the model is the prompt, and the output is the full stand-up special. As baselines, we consider a GPT2-xl finetuned in our data, as well as base versions of the larger models.

In addition to performing long-range comedy generation in this fashion, we also study how well the above LLMs can classify long-range humor into genres. To this end, we rely on Wikipedia comedy genres, which we map to our example using the names of the comedian. We design a different prompt where we ask the model to classify the stand-up directly, as follows: *"### User: Could you please provide me with the comedy categories that best describe the comedy script below? $\{transcript\}$ ### Agent: Class 1 ||| Class 2 ..."*. Models are then trained in the same fashion as above.

Table 4 summarizes our results on comedy generation and classification, respectively. We see that GPT-2 offers competitive performance on both tasks, showing that medium-sized LLMs are able, to some extent, to generate long humorous content. While these results are encouraging, they also suggest that little progress has been made in this sense, as results on our task do not improve with model scale, from hundreds of millions to billions of parameters, an indication of the complexity in understanding humor computationally.

4 Conclusions

We study the automatic detection and generation of humorous text. We show that while current models are able to generate short-range humorous content, there are limitations with this approach. We then propose a long-range scenario for comedy generation based on stand-up specials. We show that LLMs are able to tackle both tasks. While the results we have obtained are encouraging, they also clearly suggests the limitations of current models and datasets.

References

[1] Salvatore Attardo. Humor in language. In **Oxford Research Encyclopedia of Linguistics**. 2017.

[2] Victor Raskin. Linguistic heuristics of humor: a script-based semantic approach. **International journal of the sociology of language**, Vol. 1987, No. 65, pp. 11–26, 1987.

[3] Jon E Roeckelein. **The psychology of humor: A ref-**

erence guide and annotated bibliography. Greenwood Press/Greenwood Publishing Group, 2002.

[4] Alexander Baranov, Vladimir Kniazhevsky, and Pavel Braslavski. You told me that joke twice: A systematic investigation of transferability and robustness of humor detection models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 13701–13715, Singapore, December 2023. Association for Computational Linguistics.

[5] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. Harnessing Context Incongruity for Sarcasm Detection. **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)**, Vol. 51, No. 4, pp. 757–762, 2015.

[6] Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. Parsing-based Sarcasm Sentiment Recognition in Twitter Data. **Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15**, pp. 1373–1380, 2015.

[7] Piyoros Tungthamthiti, Kiyoaki Shirai, and Masnizah Mohd. Recognition of Sarcasm in Microblogging Based on Sentiment Analysis and Coherence Identification. **Journal of Natural Language Processing**, Vol. 23, No. 5, pp. 383–405, 2010.

[8] Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in Twitter. **Language Resources and Evaluation**, Vol. 47, No. 1, pp. 239–268, 2013.

[9] Yi Tay, Anh Tuan Luu, , Siu Cheung Hui, and Jian Su. Reasoning with sarcasm by reading in-between. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1010–1020. Association for Computational Linguistics, 2018.

[10] Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. Automatic Sarcasm Detection: A Survey. Vol. 50, No. 5, 2016.

[11] Byron C. Wallace. Computational irony: A survey and new perspectives. **Artificial Intelligence Review**, Vol. 43, No. 4, pp. 467–483, 2015.

[12] Orion Weller and Kevin Seppi. Humor Detection: A Transformer Gets the Last Laugh. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3621–3625, Hong Kong, China, November 2019. Association for Computational Linguistics.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv:1907.11692 [cs]**, July 2019.

[14] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor recognition and humor anchor extraction. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 2367–2376, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[15] Peng-Yu Chen and Von-Wun Soo. Humor Recognition Using Deep Learning. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 113–117, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[16] Nabil Hossain, John Krumm, and Michael Gamon. "President Vows to Cut \textlessTaxes\textgreater Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 133–142, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[17] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A Large Self-Annotated Corpus for Sarcasm. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[18] Orion Weller, Nancy Fulda, and Kevin Seppi. Can Humor Prediction Datasets be used for Humor Generation? Humorous Headline Generation via Style Transfer. In **Proceedings of the Second Workshop on Figurative Language Processing**, pp. 186–191, Online, July 2020. Association for Computational Linguistics.

[19] Machel Reid and Victor Zhong. LEWIS: Levenshtein editing for unsupervised text style transfer. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 3932–3944, Online, August 2021. Association for Computational Linguistics.

[20] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A Large Self-Annotated Corpus for Sarcasm. 2017.

[21] Suzana Ilić, Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo. Deep contextualized word representations for detecting sarcasm and irony. In **Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis**, pp. 2–7, Brussels, Belgium, 2018. Association for Computational Linguistics.

[22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[24] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **International Conference on Learning Representations**, September 2019.

[25] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[26] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. Politeness transfer: A tag and generate approach. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1869–1881, Online, July 2020. Association for Computational Linguistics.

[27] Hugo Touvron and LLama 2 Team. Llama 2: Open foundation and fine-tuned chat models, 2023.

[28] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 10088–10115. Curran Associates, Inc., 2023.