Can Humor Prediction Datasets be used for Humor Generation? Humorous Headline Generation via Style Transfer

Orion Weller
Brigham Young University
orionw@byu.edu

Nancy Fulda
Brigham Young University
nfulda@byu.edu

Kevin Seppi
Brigham Young University
kseppi@byu.edu

Abstract

Understanding and identifying humor has been increasingly popular, as seen by the number of datasets created to study humor. However, one area of humor research, humor generation, has remained a difficult task, with machine generated jokes failing to match human-created humor. As many humor prediction datasets claim to aid in generative tasks, we examine whether these claims are true. We focus our experiments on the most popular dataset, included in the 2020 SemEval's Task 7, and teach our model to take normal text and "translate" it into humorous text. We evaluate our model compared to humorous human generated headlines, finding that our model is preferred equally in A/B testing with the human edited versions, a strong success for humor generation, and is preferred over an intelligent random baseline 72% of the time. We also show that our model is assumed to be human written comparable with that of the human edited headlines and is significantly better than random, indicating that this dataset does indeed provide potential for future humor generation systems.

1 Introduction

Understanding and identifying humor has long been a goal for natural language understanding systems (Taylor and Mazlack, 2004; Hempelmann, 2008; Purandare and Litman, 2006; Mihalcea and Strapparava, 2005), with many attempts seeking to identify whether a sentence is a joke. These systems have seen impressive gains in recent years (Yang et al., 2015; Chen and Soo, 2018), with systems achieving scores in the mid 90's. As such, other areas of humor research have grown in popularity, including distinguishing between jokes (Weller and Seppi, 2019) and generating humorous text (He et al., 2019; Luo et al., 2019).

This rise in popularity has even translated to a SemEval task of predicting the level of humor in text (Hossain et al., 2019). In their work, as well as others (Mihalcea and Strapparava, 2005; Weller and Seppi, 2019) that seek to understand various aspects of humor, the authors note that their work may be influential in helping create systems that can automatically generate humor. To the best of our knowledge, however, no work has attempted to explore whether these humor prediction datasets encode information that can be used by a generative system. Instead current systems rely on retrieve-and-edit models (He et al., 2019) or models based on word senses (Luo et al., 2019).

The recent works of Hossain et al. (2019, 2020b) have created pairs of minimal changes that turn a regular news sentence into a humorous news sentence, by only changing one phrase. Because of its popularity and impact, as well as the clear insight that can be gained from minimal pair datasets (Kaushik et al., 2019; Gardner et al., 2020), we choose to examine the former as an initial exploration of what can be done. Our contributions include:

- Proposing the first model for humor style transfer, building a transformer model that "translates" from regular to humorous English[1]
- Examining whether the format of popular humor prediction datasets can be used to successfully generate humorous text. We explore this through a crowdsourced human evaluation, showing that our system performs equally to human edits (a difficult challenge for abstractive generative humor systems) as well as showing that our model provides more than random effects

[1] We publicly release our code and models at https://github.com/orionw/humorTranslate

Proceedings of the Second Workshop on Figurative Language Processing, pages 186–191
July 9, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

Original Headline	Humorous Edit
Meet the wealthy **donors** pouring millions into the 2018 elections	Meet the wealthy **sadists** pouring millions into the 2018 elections
Trump has the upper hand in North Korea **talks**	Trump has the upper hand in North Korea **handshakes**
Manhattan DA reportedly dropped felony fraud case against Trump's kids after donation from Trump's **lawyer**	Manhattan DA reportedly dropped felony fraud case against Trump's kids after donation from Trump's **doppleganger**

Table 1: Example instances of the Humicroedit dataset, containing the original headline and a humorous edited version. Edited phrase is in bold. Note that the edited headlines are designed to be humorous in light of the original.

2 Related Work

Many humor datasets have been created in order to explore humor in different circumstances. These datasets include diverse domains such as puns (Yang et al., 2015), TV shows (Purandare and Litman, 2006), Ted Talks (Chen and Soo, 2018), and online forums (Weller and Seppi, 2019, 2020). Humor prediction has even been included in this year's SemEval Task 7 (Hossain et al., 2020a) with humorous data created by online crowdsourcers who modified news headlines. Concurrent to our work, Hossain et al. (2020b) generate additional crowdsourced data through interactive online games.

In the humor generation area, previous approaches have relied strongly on templated approaches for specific types of puns or jokes (Ritchie, 2005; Binsted et al., 1997; Binsted, 1996), such as ”I like my coffee like I like my [insert phrase here].” Others have used templates based on word similarity or uncommonness to generate humorous samples (Petrović and Matthews, 2013; Valitutti et al., 2016). Recent work has started to break off from the template trend, creating abstractive models such as an RNN that creates puns (Yu et al., 2018), a retrieve-and-edit model that adds surprise to create jokes (He et al., 2019), and a Pun GAN with a word-sense disambiguater as a discriminator (Luo et al., 2019). However, none of these models employ a humorous corpus to generate their jokes, leading to the question: are they useful for models attempting to generate humor?

Our work also utilizes methods from stylized text generation (Fu et al., 2018), where work has shown success with parallel data (Zhang et al., 2018; Dai et al., 2019) as well as dealing with the lack of such data (Prabhumoye et al., 2018; Shen et al., 2017). These methods, recently employing transformer models proposed by Vaswani et al. (2017), have also been applied to formality of language (Etinger and Black, 2019) and sarcasm (Mishra et al., 2019). However, to the best of our knowledge we are the first to use style transfer for humor generation.

3 Experimental Setup

Dataset. In order to explore the utility of recently published humor corpora, we use the Humicroedit dataset created by Hossain et al. and used in the 2020 SemEval Task 7, containing more than 15,000 humorous headlines. These headlines were generated by taking a dataset of normal headlines and asking crowdsourced workers to make humorous edits. They specifically limited the workers to a single edit, where an edit was defined as the insertion of a single-word noun or verb to replace an existing entity or single-word noun or verb. Although our system will not enforce such strict edits, we use this data as a training set because of its popularity and its parallel corpus of minimal edits.

This dataset further assumes that the reader is aware of the original headline (from already popular news, for example), so that the additional word play in the edits will be humorous in light of the original headlines and topic (example instances are shown in Table 1). We note that the original Humicroedit dataset contains humor ratings for each crowdsourced edited headline: however, due to the scarcity of training data we include all headlines regardless of the humor level rating, as each edited headline was a human generated attempt at humor. In order to provide a fair training/test split, we remove all instances from the dataset which contain the same original headline, in order to prevent data leakage. We then divide the data into an 80/20 train test split, with 9000 and 2300 instances.

Model[2]. We build our model similarly to the trans-

[2] We do not report results from pre-trained models (i.e. BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019)) as we want to explore the effects of the HumicroEdit dataset apart from the effects of pre-training.

Original	President Trumps Golden Age of Trolling
Edited	President Trumps **Infinite** of Trolling
Random	President **Big Spenders Accentuation** Age of Trolling
Translated	President Trumps Golden Age of **Sassy** Trolling
Original	How CBS News reported the last national military parade in 1991
Edited	How CBS News reported the last national military **buffet** in 1991
Random	How **Ides Hemophiliac** reported the last national military parade in 1991
Translated	How CBS News **choreographed** the last national military parade in 1991
Original	Trump lawyers scramble to prepare for new stage of Russia probe
Edited	Trump lawyers scramble to prepare for new stage of **dog** probe
Random	Trump lawyers scramble to prepare for new **bailey** of Russia probe
Translated	Trump lawyers scramble to prepare for new **eggs** of Russia probe

Table 2: Example instances of all three systems: the human edited headlines, random edits, and our translated edits. Edited replacements of the original headline are in bold.

A/B Test	Ours	Other
Translated vs Edited	24	26
Translated vs Random	36*	14

Table 3: Results from A/B testing. * indicates statistical significance from a one sample test of proportions

former initially described in Vaswani et al. (2017), using an encoder-decoder architecture with eight attention heads and two layers in both encoder and decoder. We trained on the training set for 200 epochs and manually inspected checkpoint samples from the training data along the way. We chose the best performing model from the checkpoints to generate the samples for our evaluation.

Baseline. To show the effectiveness of our model on the data, we use a baseline that would generate similar surprisal effects in headline edits. We recognize the capacity of the human mind to make connections when there are none, thus, we want a baseline that randomly replaces words in a sentence and relies on that connective ability. However, a purely random model would be too naive, creating headlines that are ungrammatical and unintelligible. Thus, we create an intelligent random model that probabilistically replaces specific parts of speech with other words in that same part of speech, capitalizing the replacement phrase if the original was capitalized. We randomly replace nouns, noun-phrases, verbs, and adjectives, in order to replicate human edits in the original dataset.

4 Experiments

We use the test set described in Section 3, consisting of over 2k instances. In order to show a comprehensive view of our model, we compare the human and random edits with our translated samples. For convenience in writing, we term these models *edited*, *random*, and *translated*. We label the original non-humorous headlines, *original*. Example instances of each humor system are displayed in Table 2. We attempt to provide instances showing both positive and negative aspects of the human edited and random models. Samples examining limitations of our translation model are shown in Table 4.

We perform three experiments: a rating task, and two A/B tests between our model and the other systems. For each experiment we randomly sample 50 instances from the test set and employ users from Amazon's Mechanical Turk for feedback. We randomize the order of appearance for each trial of the A/B tests, so that order preference is controlled. We present the instances to the user and ask them "Which one of the following changes to this headline is more humorous," giving them the original headline for comparison. We limit each respondent to 5 annotations in order to avoid annotator burnout. The rating tasks are given by displaying an instance to the user and asking them to rate the headline on a 1-5 scale for fluency of language and level of humor. We then ask the user whether or not they think the headline is human generated. Our final score for both A/B tests and rating tasks consist of the average (or mode) of three annotators.

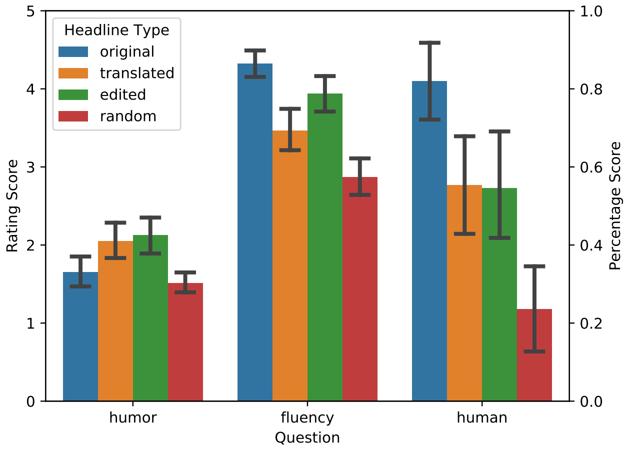

Figure 1: Results from human evaluation of headline humor, fluency, and proportion of whether or not they think it is human generated. Results are gathered from Mechanical Turk. Error bars indicate two standard errors from the mean.

5 Results

We see the results of the experiments in Table 3 and Figure 1. On the A/B tests, the human edited version of the original headline was preferred to the translated version 26 to 24 times, or 52% of the time. However, our system was preferred to the intelligent random baseline 36 times to 14, or 72% of the time. To determine significance, we conduct one-sample hypothesis tests ($\alpha = 0.05$), finding that our model is statistically significant compared to the random model, but not statistically different than the human edited version. Although this may not seem like a positive result at first glance, matching human performance on humor tasks is difficult to accomplish.

When we examine each headline in isolation, we find that the unedited human headlines are ranked significantly higher than all three systems in sentence fluency and proportion of users that thought the headline was human generated. This is to be expected, as humor does not always follow standard grammar rules. In the humor category, we see that the original headline ranked below the human edited and translated versions, also as expected. The random model performed significantly below all others in almost every task, indicating that the human edited and translated headlines contained elements that were more than random associations. We further see that the translated system performed statistically similar to the human edited version on all questions.

Example 1:	Repeats
Original:	Couple who rented condo to Pruitt pays fine to D.C.
Edit:	Couple who wore wizard to Pruitt pays to D.C. D.C. D.C.
Example 2:	Not Humorous
Original:	China says to ban some petroleum exports to North Korea
Edit:	China says to ban some 2005 to North Korea
Example 3:	No Change
Original:	California to sue Trump administration for repeal of fracking rules
Edit:	California to sue Trump administration for repeal of fracking rules

Table 4: Examples of poor performing samples from the humor translation model

We see instances in Table 4 where the translation system failed to generate a humorous edit. We observed the following categories of failure in the model: failure to change the sentence, repeated words, and non-humorous edits. From a manual inspection of a random sample of 100 instances, these errors occurred less than 5% of the time.

We see that despite the above mentioned limitations, our humor translation model matches the performance of human generated edits. As humor is a a linguistic phenomenon that depends upon the human receiving it to appreciate the humor, it is difficult to generate humor that is better than human generated content. However, our results indicate that the HumicroEdit dataset of pairs combined with our translation model is able to provide creative ways of reformulating headlines. As this work is exploratory and non-exhaustive, this gives a positive signal that the communities' efforts in humor collection have strong potential for further advances in humor generation.

6 Conclusion

In this work, we explored whether humor prediction data, such as the HumicroEdit dataset from SemEval Task 7, could be used to generate humor, examining whether this humor provides more than random surprisal effects. We use these human edited headlines as training for a machine translation system that automatically "translates" normal headlines into humor. We then build a intelligent random system as a baseline, showing that our generative headlines are significantly better than random effects, illustrating that our results are due to more than spurious correlations. We further find that our system's humorous headlines are preferred equally with those of the human generated edits, with equal proportions of crowdsourcers thinking these headlines are human generated and humorous. As this initial positive result shows that a humor prediction dataset can be used successfully for generating humor, we hope that future generative systems for humor will consider utilizing and improving from such resources.

References

Kim Binsted. 1996. Machine humour: An implemented model of puns.

Kim Binsted, Helen Pain, and Graeme D Ritchie. 1997. Children's evaluation of computer-generated punning riddles. *Pragmatics & Cognition*, 5(2):305–354.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*.

Isak Czeresnia Etinger and Alan W Black. 2019. Formality style transfer for noisy, user-generated conversations: Extracting labeled, parallel data from unlabeled corpora. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 11–16.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Quan Zhang, and Ben Zhou. 2020. Evaluating nlp models via contrast sets. *ArXiv*, abs/2004.02709.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *North American Association for Computational Linguistics (NAACL)*.

Christian F Hempelmann. 2008. Computational humor: Beyond the pun? *The Primer of Humor Research. Humor Research*, 8:333–360.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. " president vows to cut¡ taxes¿ hair": Dataset and analysis of creative text editing for humorous headlines. *arXiv preprint arXiv:1906.00274*.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020a. Semeval-2020 Task 7: Assessing humor in edited news headlines. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020b. Stimulating creativity with funlines: A case study of humor generation in headlines. *ArXiv*, abs/2002.02031.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Fuli Luo, Shunyao Li, Pengcheng Yang, Baobao Chang, Zhifang Sui, Xu Sun, et al. 2019. Pun-gan: Generative adversarial network for pun generation. *arXiv preprint arXiv:1910.10950*.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 531–538, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. 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)*, pages 6146–6155.

Saša Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for f*r*i*e*n*d*s*. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Julia M. Taylor and Lawrence J. Mazlack. 2004. Computationally recognizing wordplay in jokes. *In Proceedings of CogSci 2004*.

Alessandro Valitutti, Antoine Doucet, Jukka M Toivanen, and Hannu Toivonen. 2016. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering*, 22(5):727–749.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Orion Weller and Kevin Seppi. 2019. 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)*, pages 3612–3616.

Orion Weller and Kevin Seppi. 2020. The rjokes dataset: a large scale humor collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France. European Language Resources Association.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.