Survival of the Wittiest: Evolving Satire with Language Models

Thomas Winters and **Pieter Delobelle**
Dept. of Computer Science; Leuven.AI
KU Leuven, Belgium
{firstname}.{lastname}@kuleuven.be

Abstract

Large pre-trained transformer-based language models have revolutionized the field of natural language processing in recent years. While BERT-like models perform exceptionally well for analytical tasks such as classification and regression, their text generation capabilities are usually limited to predicting tokens within a given context. In this paper, we introduce GALMET, a model that generates text by using genetic algorithms with BERT-like language models for evolving text. We use GALMET with the RoBERTa language model to automatically evolve real headlines into more satirical headlines. This is achieved by adapting the masked language head to the headlines domain for the mutation operator and finetuning a regression head to distinguish headlines from satire for the fitness function. We evaluated our system by comparing generated satirical headlines against human-edited headlines and just the fine-tuned masked language head. We found that while humans generally outperform the model, generations by GALMET are also often preferred over human-edited headlines. However, we also found that only using the fine-tuned masked language model gives slightly preferred satire due to generating more readable sentences. GALMET is thus a first step towards a new way of creating text generators using masked language models by transforming text guided by scores from another language model.

Introduction

Large pre-trained language models such as BERT and GPT-2/GPT-3 have recently revolutionized the field of natural language processing (Vaswani et al. 2017; Devlin et al. 2019; Radford et al. 2019; Brown et al. 2020). These models are usually fine-tuned to achieve state-of-the-art performance on a wide variety of language tasks. While sequential models like GPT-2 are popular for text generation, bidirectional models such as BERT and RoBERTa also allow for generating a small number of tokens through its masked language model head. This head is trained to predict the probability of a token in a particular masked place of a sentence. For example, given a sentence *"Computational creativity is a discipline with roots in <mask> science."*, BERT predicts *"computer"* with a probability of 67.6% and *"cognitive"* with a probability of 27.6%. These probabilities can be used to generate small token modifications to a sentence.

In this research, we aim to investigate if we can improve the generative capabilities of BERT-like models by combining them with genetic algorithms. To achieve this, we introduce GALMET *(Genetic Algorithm using Language Models for Evolving Text)*. To enable these textual transformations, this framework uses two different strengths of BERT-like models, namely their text classification capabilities and their masked language model. This study thus pilots a possible combination of BERT models' masked language model head (which enables small textual modifications), with genetic algorithms' mutation operators (which requires a function that slightly modifies an individual). It also simultaneously studies BERT models' power for textual regression (thus labeling sentences with real-valued numbers) with genetic algorithms' fitness functions (which require a function to label individuals, preferably with real-valued numbers). We then evaluate if the framework can be applied for evolving headlines into more satirical-sounding texts. While the results for humor were not incredibly satisfactory, the proposed mechanism could still enable a new way for creative language generation in several other domains as well (e.g. poetry or adversarial text generation).

Background

Language Models

BERT (Devlin et al. 2019) is a language model that uses the encoder stack of transformer models (Vaswani et al. 2017), which consists of multiple *attention heads* that correlate co-occurrences of words or tokens. For an analytical detail of the attention mechanism, see the introductory paper (Vaswani et al. 2017). It is highly suited for classification and regression tasks on an input sequence, ranging from named entity recognition to high-level sentiment analysis. The BERT model later got robustly evaluated and optimized in the RoBERTa model (Liu et al. 2019). These encoder-only language models are initially trained on the masked language modeling (MLM) task that is based on the Cloze task, where the training objective is to predict a masked word or token T_i at a certain position based on the context. Interestingly, this can be interpreted as a probabilistic model, with the model generating a conditional distribution for each masked token T_i following

$$Pr\left(T_i \mid T_1 \dots T_{i-1}, T_{i+1} \dots T_n\right).$$

These tokens are in most cases words, based on the frequency of each word appearing in a dataset. A common word can usually be expressed by a single token, while less frequent words are usually expressed as multiple tokens. The MLM task allows the model to learn linguistic knowledge in a self-supervised manner from unlabeled text sequences and is usually only used for pre-training. With transfer learning, a new head can be fine-tuned in a supervised manner to perform another type of language task.

Genetic Algorithms

The genetic algorithm is a prominent type of evolutionary algorithm that uses techniques inspired by natural selection to discover high-quality solutions for a search problem where solutions can be evaluated (Holland and others 1992). The algorithm generates an initial population of μ individuals, evaluates its fitness and selects the best few for the next generation. These selected individuals are often crossed-over, where new individuals have elements from multiple parents, and are often slightly mutated. This continues until the stopping criteria are reached, such as the desired fitness value.

Satire Detection

Recently, several researchers released datasets for performing satire detection. Most notably, a dataset called *"Humicroedit"* contains headlines and the edits humans did to create funnier headlines, which were then rated by other humans (Hossain, Krumm, and Gamon 2019). In a competition, 48 different teams created models to achieve the best performance for estimating the perceived funniness of the edited headlines (Hossain et al. 2020a). As expected, most teams used pre-trained language models such as BERT and RoBERTa, with the winner using an ensemble of six different pre-trained language model architectures (Hossain et al. 2020a).

Satire Generation

There have been several research projects aiming to automate satire. One approach uses a genetic algorithm, that substitutes words from movie titles with words related to the satirical target to create satirical movie titles. An apprentice then learns to replicate from this algorithm and humans on Twitter creating the same type of movie title variations with this context using a neural sequence-to-sequence model (Alnajjar and Hämäläinen 2018). Other researchers also used the earlier mentioned Humicroedit dataset to train a transformer model to generate satirical edits to real headlines (Weller, Fulda, and Seppi 2020). Another recent approach used BERT summarization models that map true headlines, leading paragraphs and Wikipedia contexts to satirical headlines, achieving 9.4% funny headlines (Horvitz, Do, and Littman 2020).

Data

We combined datasets containing real and satirical headlines used previously in research and competitions, and added a funniness label. In this work, we use "real headlines" to refer to headlines that were created by actual news websites, and "satirical headlines" to refer to headlines from satirical websites. Some datasets also contain edited real headlines and edited satirical headlines, to be respectively more or less funny than the original. If a funniness rating was present in a dataset, it was normalized and used as a label, otherwise real headlines received label 0 and satirical headlines label 1. If a headline was already in the combined dataset, it was not included again. We first added datasets that contain rated funniness, namely 15K edited headlines from Humicroedit (Hossain, Krumm, and Gamon 2019), 8K from Funlines (Hossain et al. 2020b) and 2.7K from Unfun (West and Horvitz 2019). The first two edited real headlines by changing one word, while the third made satirical headlines less funny by changing as few words as possible. We also included unrated real and satirical headlines, namely 20K from Unfun, 26K from the Sarcasm Detection dataset (Misra and Arora 2019) and 22K from The Onion or Not dataset[1].

After splitting into training, validation and test sets, the dataset was augmented by converting it to different casings (lowercase, uppercase, title case). Not only because this could give a hint to the original dataset (where for example lowercase was more prevalent), or because the edited headlines sometimes had a different case for the edit, but also because these sentences map to completely different tokens, and thus to different input sequences for a RoBERTa model. This resulted in a training dataset of 236k training, 30k validation and 34k test instances in their respective datasets. The distribution of the dataset is displayed in Figure 1, showing that it mostly contains real headlines and real satire.

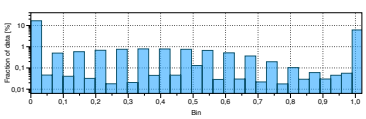


Figure 1: Histogram of the labels of the complete regression dataset. Note that the extreme bins have more sequences due to binary datasets, and more data every other bin due to Humicroedit and Funlines having 16 possible score bins.

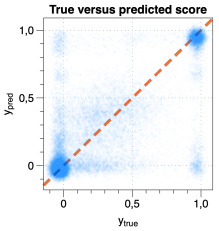


Figure 2: Predicted regression scores in function of the true scores.

RoBERTa Models

We fine-tuned two RoBERTa heads on the dataset, one regression model predicting the funniness rating and one masked language model on only the text sequences (referred to as *Satire MLM* in the remainder of the paper). The regression model achieves $MSE = 0.0447$ and $R^2 = 0.548$ on the held-out test set (see Figure 2 for the distribution). While the RMSE of 0.2113 seems surprisingly low compared to the Humicroedit competition winners (with RSME of 0.5016), this is due to our score normalization, and is also made incomparable due to our large dataset augmentation.

[1] https://www.kaggle.com/chrisfilo/onion-or-not

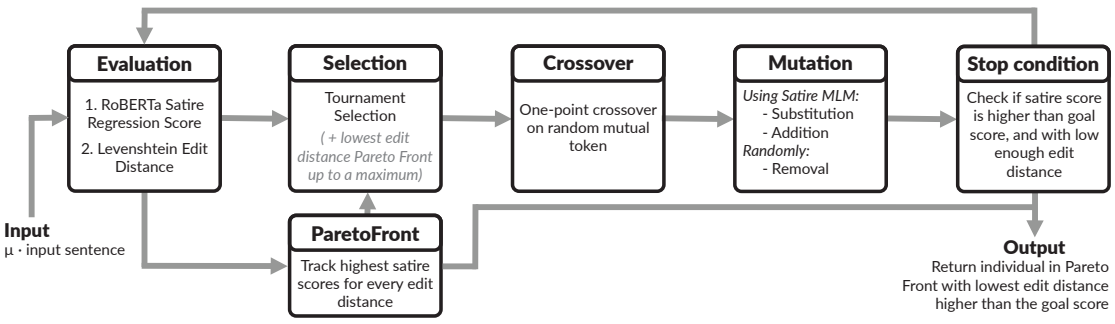


Figure 3: Overview on the components used in GALMET and the population flow.

Hyperparameter	Value
adam_epsilon	10^{-8}
adam_betas	$\beta_1 = 0.9, \beta_2 = 0.999$
fp16	False
gradient_accumulation_steps	$i \in \{1, 2, 3, 4\}$
learning_rate	$[10^{-6}, 10^{-4}]$
max_grad_norm	1.0
max_steps	-1
num_train_epochs	3
per_device_eval_batch_size	8
per_device_train_batch_size	8
seed	1
warmup_steps	100
weight_decay	$[0, 0.1]$

Table 1: The hyperparameter space used for finetuning all language models.

GALMET

We introduce the GALMET model, a Genetic Algorithm using Language Models for Evolving Text. We use the fine-tuned RoBERTa models to implement the genetic algorithm operators for a GALMET model that aims to transform headlines into more satirical counterparts.

Flow

GALMET starts by receiving an input sentence, here an initial headline. This sentence is tokenized and duplicated μ times as the initial population. GALMET then repeats several classic genetic algorithm steps, namely evaluating and selecting the best individuals, crossing and mutating them until a stop condition is achieved (Figure 3). In this case, evaluation happens by predicting the funniness of the sentences using the fine-tuned regression model as a fitness function, and calculating the Levenshtein distance function to prefer fewer edits. The best sequence per edit distance is saved in a Pareto front and added into the next generation. Further selection happens using a tournament selection with $\kappa = 3$ individuals. The crossover operator uses a variant on one-point cross-over, that crosses sentences on a random word that both individuals contain. This operator thus potentially combines successful modifications to the left of that word, with modifications to the right of that word in another individual. For the mutation operator, we created three different operators. The first is token substitution, which substitutes a token with a mask and uses the probability distribution from the Satire MLM to sample a replacement token. For example, for a sentence *"The lion roars."*, the token for *"lion"* could be selected to be replaced by a mask to create *"The <mask> roars."*, which is then filled in using the Satire MLM to create the sentence *"The alarm roars."*. The second mutation operator is token addition, which randomly adds a mask in the sequence and fills it in similarly. The third removes a random token from the sequence. The algorithm stops when an individual receives a score from the regression model above a certain threshold, e.g. 0.99.

Parameter	Value
μ (population size)	50
p_c (crossover probability)	0.2
p_m (total mutation probability)	0.8
$p_m^{substitution}$	0.7
$p_m^{addition}$	0.05
$p_m^{removal}$	0.05
max # generations	30
goal fitness	0.99
max edit distance	7
max # elites	6
elite duplicates	3

Table 2: The parameters used for transforming headlines into more satirical headlines using GALMET

An example execution of GALMET is shown in Table 3, where the sentence *"Most Americans Want Congress To Investigate Michael Flynn"* is transformed into *"224 Americans Asked To Investigate Michael Jordan"*. An example that ran for more iterations is given in Table 4, for which the evolution of the fitness functions is summarized in Figure 4. As expected, the edit distance rises, but is limited thanks to the Pareto front. We can see that the desired regression score is achieved at generation 14, but that the edit distance was 8, and thus not sufficient for the stop condition demanding an edit distance of 7 or less. We can also see that the regression score stays near 0, and then quickly jumps to the 1 regions, a clear bias originating from the binary datasets with little values between 0 and 1 (as illustrated in Figure 2).

d_{edit}	Score	Phenotype
0	0.010341	Most Americans Want Congress To Investigate Michael Flynn
1	0.019262	Why Americans Want Congress To Investigate Michael Flynn
2	0.070894	Most Americans Want To Investigate Michael Jordan
3	0.458300	224 Americans Asked To Investigate Michael Flynn
4	1.001400	224 Americans Asked To Investigate Michael Jordan

Table 3: Example of a Pareto front after 5 iterations, containing the highest scoring mutation for each edit distance to the original. Here, the sentence from $d_{edit} = 0$ is thus transformed into the sentence for $d_{edit} = 4$.

d_{edit}	Score	Phenotype
0	0.008290	Amazon removes Indian flag doormat after minister threatens visa ban
1	0.011139	Amazon removes Indian flag doormat after minister visa ban
2	0.018676	Amazon removes Indian flag doormat after violating visa ban
3	0.235192	NFL removes Indian flag doormat after violating visa ban
4	0.716432	NFL removes Indian flag doormat after violating OT ban
5	0.941911	NFL welcomes Indian flag doormat after violating OT ban
6	0.949949	NASA adds rainbow flag doormat after massive OT ban
7	1.004551	NASA releases rainbow leg doormat after violating OT ban

Table 4: An example that ran for 16 iterations (see Figure 4).

Code

The code, fine-tuned RoBERTa models and further implementation, training and parameter details are available on `https://github.com/twinters/galmet`.

Evaluation

We evaluated GALMET by transforming specific headlines into more satirical versions and checking which ones are funnier than human-created satirical versions. We also evaluated if the genetic algorithm is better than its mutation operator component by making it compete with our Satire MLM. Given a headline, the MLM head applies the substitution mutation operator an equal number of times as GALMET's maximum allowed edit distance, which we set to seven tokens (note that one word might be composed of several tokens). We randomly sampled elements from the Humicroedit test set 408 times, which contain a real headline and human-edited version. Then, both GALMET and the Satire MLM generate a new headline that is at most seven token edits away from the original, real headline. This resulted in 816 possible matchups. In the experiment, participants were shown two headlines in a shuffled order and asked to select the funnier headline. Thus, unlike previous work using these types of datasets (Hossain, Krumm, and Gamon 2019), we did not ask participants to add a score to each element, but instead just asked to pick their favorite from two given texts, as this is a more natural task to perform. In total, 18 participants labeled 1498 pairs, 756 against human-edited and 742 against the Satire MLM, resulting in the results in Figure 5.

While human-edited headlines are usually preferred over GALMET-evolved headlines, it is still able to perform a funnier edit than humans in 26% of the times[2], which is quite an achievement given the intrinsic difficulty of humor for machines. However, the lines generated by its component, the Satire MLM, are preferred in 53% of the cases over the GALMET-generated lines. At first glance, it would seem that for the current parameters, just using the MLM head is enough. However, we see two likely issues that could have caused this near-uniform preference, disadvantaging GALMET. First, several evaluators pointed out the difficulty of evaluating broken headlines. Given that GALMET additionally adds and removes random sub-word tokens in random locations, this likely often results in broken sentences (e.g. *"In long-feared twistlamp leak rattling American people"*), or in nonsense words (e.g. inventing words like *Vomoted* by adding a token in the middle of *Vote*, and *Haveared* by removing a token from *Have Cleared*). Second, since the training data only contained sensible words, these sequences containing the aforementioned non-existent (and thus out-of-domain) words likely received near-random scores, reducing the usefulness of genetic algorithm components like the regression model.

[2]Note that Humicroedit participants only changed one word, whereas GALMET can update several (subword) tokens, as to not turn the genetic algorithm into a simple search problem of depth 1.

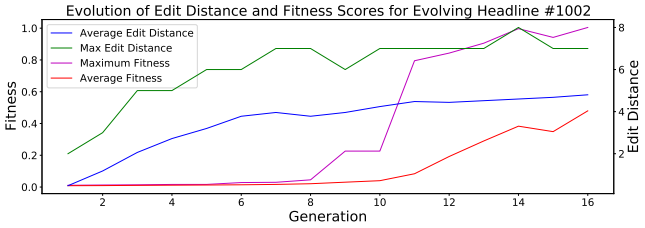
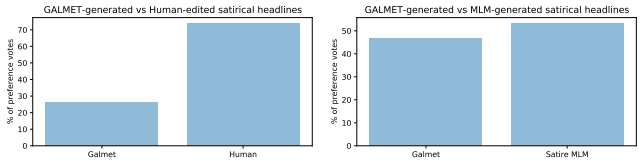


Figure 4: Evolution of the fitness value and edit distance corresponding to generate the example of Table 4.



(a) GALMET versus human edits (b) GALMET versus Satire MLM

Figure 5: Preferences of participants in our human evaluation for the source of the transformed headlines.

Future Work

While GALMET did not display outstanding results in this particular study, we still believe it can be the basis for a powerful mechanism for transforming textual sequences. There are several improvements for the satirical headlines domain we envision. First, it would be beneficial to improve the dataset balance by removing some binary class datasets, as they bias the regression model strongly towards either close to 0 or to 1. Second, only allowing real words to appear in sentences, by only replacing full words, and allowing to insert a random number of neighboring masks instead of single masks when using the mutation operators. Third, adding a detector to steer away from broken sentences, which is achievable for a RoBERTa-based model in a humor detection setting (Winters and Delobelle 2020). Fourth, it would

be beneficial to either improve the cross-over operator to create better individuals than our current cross-over operator or just leave the cross-over out and use the components in a search setting instead of genetic algorithms.

We expect the GALMET framework to be interesting for other text generation domains too, such as poetry generation or generating adversarial examples to textual classifiers. The framework could also have a use in co-creative applications by suggesting improvements to given text sequences.

Conclusion

We investigated how to improve the generative capabilities of analytical language models by combining them with genetic algorithms. For this, we created a novel text generation method for evolving text into text from a different domain, in our case, transforming headlines into satirical headlines. To achieve this, we introduced several new genetic operators based on pre-trained language models. On evaluation, we found that it performs similar to one of its components, and identified several causes and potential solutions. We believe this framework could open the way for novel co-creative applications where users can evolve their texts towards particular goal text domains, even if that domain might usually be hard for computers to grasp, such as poetry or humor.

Acknowledgments

We would like to thank the volunteers for judging the headlines in the human evaluation. Thomas Winters is a fellow of the Research Foundation-Flanders (FWO-Vlaanderen, 11C7720N). Pieter Delobelle was supported by the Research Foundation - Flanders (FWO) under EOS No. 30992574 (VeriLearn) and also received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

Alnajjar, K., and Hämäläinen, M. 2018. A master-apprentice approach to automatic creation of culturally satirical movie titles. In *Proceedings of the 11th International Conference on Natural Language Generation*, 274–283.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Holland, J. H., et al. 1992. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Horvitz, Z.; Do, N.; and Littman, M. L. 2020. Context-driven satirical news generation. In *Proceedings of the Second Workshop on Figurative Language Processing*, 40–50. Online: Association for Computational Linguistics.

Hossain, N.; Krumm, J.; Gamon, M.; and Kautz, H. 2020a. Semeval-2020 task 7: Assessing humor in edited news headlines. *arXiv preprint arXiv:2008.00304*.

Hossain, N.; Krumm, J.; Sajed, T.; and Kautz, H. 2020b. Stimulating creativity with funlines: A case study of humor generation in headlines. *arXiv preprint arXiv:2002.02031*.

Hossain, N.; Krumm, J.; and Gamon, M. 2019. "president vows to cut taxes hair": Dataset and analysis of creative text editing for humorous headlines. *arXiv preprint arXiv:1906.00274*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Misra, R., and Arora, P. 2019. Sarcasm detection using hybrid neural network. *arXiv preprint arXiv:1908.07414*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 5998–6008.

Weller, O.; Fulda, N.; and Seppi, K. 2020. 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*, 186–191.

West, R., and Horvitz, E. 2019. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7265–7272.

Winters, T., and Delobelle, P. 2020. Dutch humor detection by generating negative examples. In *Proceedings of the 32st Benelux Conference on Artificial Intelligence (BNAIC 2020) and the 29th Belgian Dutch Conference on Machine Learning (Benelearn 2020)*.