Cognitively Controlled Sentiment-based Review Generation

Abstract

In recent years, computational, neurocognitive, and psychological approaches to fiction and poetry have become more prevalent. This includes the use of statistical modeling in analyzing the features of literature, studying the impact of reading literary fiction on reader empathy (e.g., for real or fictional demographics), and considering the neurological bases of reading poetry. At the same time, computational psycholinguistics and cognitive natural language processing (NLP) have used neural networks to model cognitive data (e.g., eye-tracking and neuroimaging) and applied these to influence NLP tasks or enrich understanding of how humans process written language.

Natural Language Generation (NLG) is an area that has seen multiple advances with the advent of large neural language models. Controllable NLG enables generation to be steered so that the produced texts possess target attributes, such as positive or negative sentiment. As NLG becomes a large-scale, widely deployed aspect of our lives due to advances in powerful generative language models, the urgency to understand the impact of generated texts on readers increases—such as manipulation of reader affect. One approach to assessing this impact would be to compare the target sentiment of generated text to judgments of sentiment made by human readers of the text. We hypothesize that modeling human cognitive responses as part of this process may both amplify the potential impact—by increasing transferability between artificial and human judgments—and improve interpretability of the mechanisms of this impact by engendering more human-readable patterns in models.

In this work, we ask whether models can be successfully biased to produce movie reviews of targeted sentiment in a cognitive fashion—by inducing more human-like biases with synchronized measurements of eye movements and brain activity—and whether these biases augment generated reviews such that human readers agree more often with their putative sentiment than readers do those of non-cognitive models' reviews. If cognitively controlled reviews are more congruent with human judgments than baseline control without cognitive influence on models, we can clarify the relationship between the modeled cognitive data in the context of machine generation and human reading: such as by examining which features are enhanced or attenuated and how they are manifested in reviews with high correlation with human judgments. This may enable the refinement of computational models of cognitive language processing while gleaning information which can be used to improve NLG controllability in alignment with human goals.

Introduction and Overview

The advent of large pretrained generative language models such as GPT-2 (Radford et al., 2018) in Natural Language Processing (NLP) has increased research and deployment of Natural Language Generation (NLG) tools in academia and society. An area of study in this subfield is controllable NLG, whereby various techniques are implemented to influence the characteristics of generated text. The type of texts generated may vary, such as stories or movie reviews. One aspect of texts that can be influenced is sentiment, such as a positive or negative movie review.

With Plug and Play Language Models (PPLM; Dathathri et al., 2020), a controllable NLG method, a text can be generated with a preferred sentiment, by influencing a generative model's output with a sentiment classifier, such as BERT (Devlin et al., 2017). One issue with classifiers is ensuring their outputs correlate with human judgments. Typically, classifiers are trained by supervision with self-reported human labels, so that the implicit goal is to induce a human-like bias. However, self-reports are known to be unreliable, and can not provide information inaccessible to annotators, such as eye movements or brain activity.

Thus, to induce a more robust human-like bias, cognitive data may be used to complement human labels in the training of classifiers, potentially yielding more unique decisions. This is done by training models on the same input sequences that were read by human subjects whose cognitive activity was recorded. In order to verify the validity of unique, cognitively biased decisions about classified inputs, such as movie reviews, human judgments may be employed: if humans frequently agree with cognitively informed BERT models when they disagree with baseline BERT models regarding the original input data, we can claim that they stand on equal footing on the ground truth.

The size of datasets used to elicit recorded human data is relatively limited, consequently limiting the generalizability and robustness of training and evaluating machine learning models using these data. However, a cognitively supervised sentiment classifier may be used to influence the text generated by a model such as GPT-2, possibly transferring that human-like bias. Generating samples would remove the low number of samples, potentially improving robustness of statistical investigation of agreement between human judges and model decisions.

The goal of this research is to determine whether cognitive data collected from eye-tracking and electroencephalography (EEG) brain recordings of human subjects reading reviews from a dataset can be successfully leveraged to induce biases in neural network-based models used to generate movie reviews, whether these biases more effectively elicit target ratings of sentiment from human readers of these reviews, and, if

so, what cognitively informed properties might manifest in the generated reviews which distinguish these effects.

Our hypotheses are that the qualitative differences which appear in classification with a BERT model informed by cognitive data can transfer to reviews generated by GPT-2 under the PPLM method, and that human readers may agree more often with the ratings of cognitively informed reviews. These can be tested by training classifiers, generating reviews, and enlisting human judges to annotate the reviews, subsequently investigating their differences with statistical methods.

Research on cognitive NLP is still nascent, and the methods of applying cognitive data to train models is not well-developed, nor the effects well-understood. Extending the research to include controllable NLG might illuminate the generalizability and impact of human-like biases. If the aim of NLP is to explicitly or implicitly emulate human language processing, research which produces such illumination is useful. At the same time, the goal of value-alignment may be served by investigating ways which enable explicit control over models to align them more with certain cognitive processes, or to attenuate that alignment in the case of human prejudices.

The ability to create models which perform equally well according to certain objectives (e.g., classification accuracy) while offering qualitatively different outputs can create the opportunity to shape models to be more interpretable, by aligning outputs with the sort of attention or biases that human auditors would expect. For a human-in-the-loop procedure such as co-authoring texts with a generative model, this can be additionally beneficial by giving the human co-author a more intuitive interface with the machine co-author.

Related Work

Related work can be divided among the major methodological components of our research agenda: To create a cognitively informed classifier, to use it to influence model generation in a controllable NLG system, and to assess the valence of generated texts. In terms of analyzing human judgments of the positive or negative valence of generated text, previously, Sheng et al. (2019) created a *regard* metric—akin to sentiment but designed to more directly measure bias—to assess biases in GPT-2 continuations using BERT classifiers, exploring whether continuations generated with particular prompts showed prejudice against certain demographics, such as women.

In terms of controllable NLG, Sheng et al. (2020) followed up this work with the use of adversarial triggers—certain phrases in inputs which influence outputs—to analyze and mitigate biases toward demographics when generating text in a controlled fashion. As noted, Dathathri et al. (2020) have used Plug and Play Language Models (PPLM) to control attributes of generated text without retraining models or modifying model architecture. The gradients of a classifier are used to modify the hidden states of a generator. This will be the method used in our research, as it allows us to leverage cognitively informed BERT classifiers to influence GPT-2 models.

The method we use to inform classifiers with cognitive data is inspired by Barrett et al. (2018), which is effectively a multi-task learning (MTL) approach where the cognitive data are treated as proxies for attention and used to compute an auxiliary attention loss. This attention supervision is meant to influence which words are attended to by making model attention weights more similar to cognitive data through a process of blame assignment by backpropagating loss. Hollenstein et al. (2019) have combined eye-tracking (ET) and EEG data for a suite of tasks by concatenating cognitive features to word embeddings as well as predicting features as an auxiliary task, such as predicting fixation duration, and their work provides the data we will use in our own. Similar to Barrett et al., Muttenthaler et al. (2020) implement an attention loss based on EEG data, while Malmaud et al. (2020) predict eye-tracking data in order to augment question answering tasks.

Research Design and Methodology

In order to cognitively control the generation of reviews and produce outputs that humans can assess, we will use the popular generative neural language model GPT-2 in the PPLM framework. We will use a cognitively informed BERT sentiment classifier to influence GPT-2's outputs, with modification of PPLM. As noted, PPLM is an *ex post facto* technique for controlling a pretrained generator such as GPT-2 by using another pretrained model as a discriminator (e.g., a sentiment classifier labeling reviews as positive or negative), providing feedback on GPT-2's outputs, without additional training or alteration of models. In our study, we use BERT as a sentiment classifier because BERT possesses a similar neural architecture to GPT-2, but is designed for discriminative tasks. GPT-2 and BERT are characterized by the use of a self-attention mechanism to selectively emphasize words in sentences and create contextualized representations of each word.

We will use GPT-2 steered by base BERT sentiment predictions as well as those of BERT modified to approximate human responses. The latter version is created by attention supervision with the Zurich Cognitive Language Processing Corpus (ZuCo; Hollenstein et al., 2018), which contains synchronized eye-tracking and brain activity (EEG) data: cognitive responses to reading movie review excerpts from the Stanford Sentiment Treebank (SST). Words in these samples have corresponding gaze and brain data, which can act as proxies for human attention, and patterns in the data can be leveraged to inform the words to which BERT most attends. With an attention-focused approach, we are able to evaluate whether classifiers learn human-like biases by comparing the similarity of attention distributions in conjunction with task performance metrics such as accuracy.

Unlike typical recurrent networks, BERT uses self-attention with multiple attention heads (Vaswani et al., 2017) to create its contextualized representations, where the features of a given word representation reflect the features of similar words in the sentence, as determined by a scaled dot product between the tokens. This dot product is passed through softmax to create a distribution of attention weights over the tokens in the sentences, so that each token in a sequence of length N has N attention weights, creating an attention weight matrix for each head, which is matrix multiplied with token representations to rescale features. A special classification token, [CLS], taken from the final layer is used as the pooled sentence representation used to classify sentences. Thus, influencing the attention weights which correspond to this token in turn influences the sentence representation used for classification.

For the ZuCo data, each token in a sentence has a set of eye fixation (a focused paus on a particular area of interest) and a set of electroencephalography (EEG) electrode activity values, measured in milliseconds and microvolts, respectively. These vectors of

values for each token can be reduced to scalars and passed through softmax so that each sentence has a distribution of weights reflecting the magnitude of activity or attention human readers allotted to words when reading sentences.

In order to induce human-like biases in model attentions with BERT and its more sophisticated attention mechanism, we can take the final layer's [CLS] attention weights and average them over attention heads, computing the Kullback-Leibler divergence between cognitive data and adding this as an auxiliary loss to the main sentiment classification loss.

Previous research (McGuire & Tomuro, 2021) has found that BERT attention can be successfully influenced to be similar to ZuCo data, and that, while for such models classification accuracy remains the same as baseline BERT models, the cognitive BERT models more often misclassify different samples from baseline, and these unique errors tend to be false negatives. This extended the related work which was limited to less sophisticated attention mechanisms for recurrent models, while quantifying differences and similarities in errors and attention. Base BERT can be seen as biased toward subjective, self-reported human labels as ground truth, while cognitive attention supervision adds an objective, physiological ground truth. The validity of these standards can be substantiated with human judgments of samples on which respective BERT models disagree.

To combine this assessment with open-ended generation so that we can evaluate transferability to generated texts and potential differences in human annotation, given partial SST samples as prompts, GPT-2 will be influenced to generate continuations based on both baseline and cognitive BERT predictions, producing two sets of reviews. Previously we have found that a ratio based on the symmetric difference of errors between baseline and cognitive BERT models was higher than the mismatches between baseline and randomly supervised models. That is, non-random attention supervision with ZuCo data resulted in more disagreements with baseline BERT, using samples taken from the SST dataset. We can similarly assess mismatches and related properties such as attentional differences with reviews generated by GPT-2 with the two classifiers at the helm. Each generated review will be classified by both BERT models to discover mismatches—where the BERT models disagree.

To add ground truth validity to one decision or the other, we will use a popular crowdsourcing platform, Amazon Mechanical Turk (MTurk), to hire a trio of judges to annotate the sentiment of all samples that engender mismatches, so that each sample has the originating BERT label, its counterpart BERT label, and three human labels. Agreement can be determined with Fleiss' kappa and Spearman's correlation. While crowdsourcing can complicate such calculations by allowing large numbers of different workers to form different trios and annotate different samples, a fixed trio of annotators working on the same samples can be enforced through more constrained task design.

We will assess which models human judges tended to agree with, gleaning insight into the efficacy of using cognitive information in a pipeline to generate texts that elicit positive or negative responses in human readers.

By creating a sentiment classifier which exhibits unique human-like biases due to the application of cognitive data and steering the generation of reviews, we can explore differences between baseline BERT and cognitive BERT models by an extension of their discriminative biases through generative processes, which may uncover pathologies unseen in less open-ended classification tasks. In some cases, it may be that a classifier learns to attend more or less to certain words important for sentiment, and that GPT-2 in turn learns to sample these words more or less often during generation. Such processes can in part be discovered by examining the top-*k* most attended words overlapping between models and human data.

Additionally, ascertaining whether such biases can be transferred across models and tasks expands the scope of cognitive NLP studies, suggesting a myriad of experimental designs for investigating the use of cognitive data by adding generative tasks rather than standard classification tasks. Confinement to the latter may constrain patterns or suppress them in service to straightforward objectives such as accuracy or F1 scores, while the former allows for a wide variety of deployments.

For NLG, the ability to successfully employ cognitive data would allow exploration of human biases extant in generative models, such as racism or sexism supported by objective physiological measurements of reading behavior, rather than solely relying on patterns models learn from self-reported judgments of training data. As there is a rich history of exploring human cognitive biases in psychology and cognitive science, this may enable a greater bridge between the fields in developing experimental understanding.

This will be an early step in exploring the impact of emerging technologies which allow indefinitely large amounts of fluent text to be generated and read from a relatively small system: a large pre-trained language model.

Plan of Work and Outcomes

Activity	Purpose	Timeline	Outcome
Train BERT sentiment classifier with ZuCo data	This will create models with putative human-like biases we can measure and attempt to transfer to generated texts.	2-7 days - . We expect this to proceed quickly, based on prior work which procured and processed the cognitive data, established the code base for attention supervision to induce biases in model attentions, and the general methodological approach for training and measuring models.	Trained BERT sentiment classifier with quantifiable attention biases induced by ZuCo data.
Use GPT-2 to generate reviews with PPLM	Generating reviews with GPT-2 with target sentiments steered by cognitively biased BERT sentiment classifiers gives the opportunity to measure whether biases could be successfully transferred, and produce generated review datasets which can be classified by baseline and cognitive BERT models as well as human judges.	Two weeks to modify PPLM code to accommodate transformer architectures and evaluate mismatches and related differences between the two datasets to be sent to human judges.	Generated reviews.
Employ human judges to annotate reviews	To determine whether humans agree as much or more with the target sentiments of cognitively biased vs. baseline samples, we hire a trio of judges with the crowdsourcing platform AMT, collecting and analyzing results.	Annotation moves quickly, but data processing and analyses may take an estimated two weeks.	Measurements of human and machine agreement on generated samples.

Conclusions and Future Work

If human judges agree more with the sentiment analyses by cognitive BERT models of generated reviews steered, this would suggest that such models may offer judgments that correlate better with human judgments. Additionally, in the case of reviews generated by GPT-2 models influenced by baseline or cognitive BERT models, we can examine what features were learned which may have contributed to these correlations. Further experimentation could review how to amplify or reduce these features, perhaps, allowing the generation of texts which increase, for example, the sort of negativity a creative writer might want in a horror story, or to reduce feelings of prejudice against certain demographics.

References

Barrett, Maria, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. "Sequence classification with human attention." In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 302-312. 2018.

Dathathri, Sumanth, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. "Plug and Play language models: A simple approach to controlled text generation." *arXiv preprint arXiv:1912.02164* (2019).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Gabriel, Shira, and Ariana F. Young. "Becoming a vampire without being bitten: The narrative collective-assimilation hypothesis." *Psychological Science* 22, no. 8 (2011): 990-994.

Hollenstein, Nora, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. "ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading." *Scientific data* 5, no. 1 (2018): 1-13.

Hollenstein, Nora, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. "Advancing NLP with cognitive language processing signals." *arXiv preprint arXiv:1904.02682* (2019).

Keen, Suzanne. *Empathy and the Novel*. Oxford University Press on Demand, 2007.

Malmaud, Jonathan, Roger Levy, and Yevgeni Berzak. "Bridging Information-Seeking Human Gaze and Machine Reading Comprehension." *arXiv preprint arXiv:2009.14780* (2020).

McGuire, Erik, and Noriko Tomuro. "Relation Classification with Cognitive Attention Supervision." In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 222-232. 2021.

Muttenthaler, Lukas, Nora Hollenstein, and Maria Barrett. "Human brain activity for machine attention." *arXiv preprint arXiv:2006.05113* (2020).

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

Sheng, Emily, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. "The woman worked as a babysitter: On biases in language generation." *arXiv preprint arXiv:1909.01326* (2019).

Sheng, Emily, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. "Towards controllable biases in language generation." *arXiv preprint arXiv:2005.00268* (2020).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Appendix: Budget

Amazon Mechanical Turk[1]	$25-100
• Masters Qualification	5% of worker reward

We will programmatically ensure workers are paid rewards in what amounts to a fair[2] hourly wage for analyzing 1-2k generated reviews, which should take less than one hour. Additional expenses are occurred with premium[3] prices set per assignment by Amazon for setting more exacting requirements for annotators to meet, such as age ($0.50) fluency ($1.00) in the language of the samples, educational level ($0.65 for US graduate degree), and worker reputation or experience (to increase quality of annotations)—Masters Qualification.

[1] https://www.mturk.com/pricing
[2] https://fairwork.stanford.edu/
[3] https://requester.mturk.com/pricing