Context-Driven Satirical Headline Generation

Zachary Horvitz Nam Do Michael L. Littman
Department of Computer Science
Brown University
Providence, RI 02912
{zachary_horvitz,nam_do,michael_littman}@brown.edu

Abstract

While mysterious, humor likely hinges on an interplay of entities, their relationships, and cultural connotations. Motivated by the importance of context in humor, we consider methods for constructing and leveraging contextual representations in generating humorous text. Specifically, we study the capacity of transformer-based architectures to generate funny satirical headlines, and show that both language models and summarization models can be fine-tuned to regularly generate headlines that people find funny. Furthermore, we find that summarization models uniquely support satire-generation by enabling the generation of topical humorous text. Outside of our formal study, we note that headlines generated by our model were accepted via a competitive process into a satirical newspaper, and one headline was ranked as high or better than 73% of human submissions. As part of our work, we contribute a dataset of over 15K satirical headlines paired with ranked contextual information from news articles and Wikipedia.

1 Introduction

Despite long-term interest in the foundations of humor, work to date in the NLP community on humorous text has largely relied on surface-level features (e.g., puns). We study employing richer contextual representations to generate satirical news headlines, which necessitate a balance between funniness and topicality. While our particular focus is humor, our methods are broadly applicable to tasks that require reasoning over textual knowledge.

Consider the following satirical headline from The Onion (TheOnion.com):

TC Energy Says Keystone Pipeline Failed Due To Protestors Making It Lose Confidence In Itself

In addition to knowing the connection between failure and self-confidence, processing the humor of this headline presupposes knowing (or inferring):

1. TC Energy Oversaw the Keystone XL Pipeline.
2. The Keystone XL pipeline failed amid protests.

Thus, satirical news requires an appreciation of a real-world, non-funny context.

Existing literature on the psychology of humor emphasizes the role of complex representations and relationships (Morreall, 2016; Martin, 2010; Attardo, 2001; Raskin, 1985; Attardo, 2014). Psychologists have offered multiple theories of humor. According to "Superiority Theory," jokes hinge on the power-relations between entities, while "Relief Theory" ventures that humor releases conflict between desires and inhibitions. "Incongruity Theory" sees humor as emerging from low-probability juxtapositions between objects and events.

Regardless of the theoretical framework, moving from surface-level features to a deeper analysis of humor requires an implicit calculus of entities, their relationships, and even cultural connotations. Recent NLP and NLG research has sought to apply psychological hypotheses to understand and generate humorous text. J.T. Kao (2016) applied the incongruity framework to analyze and predict the funniness of puns, and found that puns rated funnier tended to be more ambiguous. Building on the aforementioned work, Yu et al. (2018) applied neural models to pun generation, and He et al. (2019) found that puns could be procedurally created by inserting low-probability (as determined by a language model) homophones into non-funny sentences.

Other related work has established style-transfer and translation approaches for sarcasm generation. For example, Mishra et al. (2019) introduced a pipeline for converting an input sentence to a sarcastic form by neutralizing its sentiment, translating it into strong positive sentiment, and then combining it with a negative event. This pairing creates

Proceedings of the Second Workshop on Figurative Language Processing, pages 40–50

https://doi.org/10.18653/v1/P17

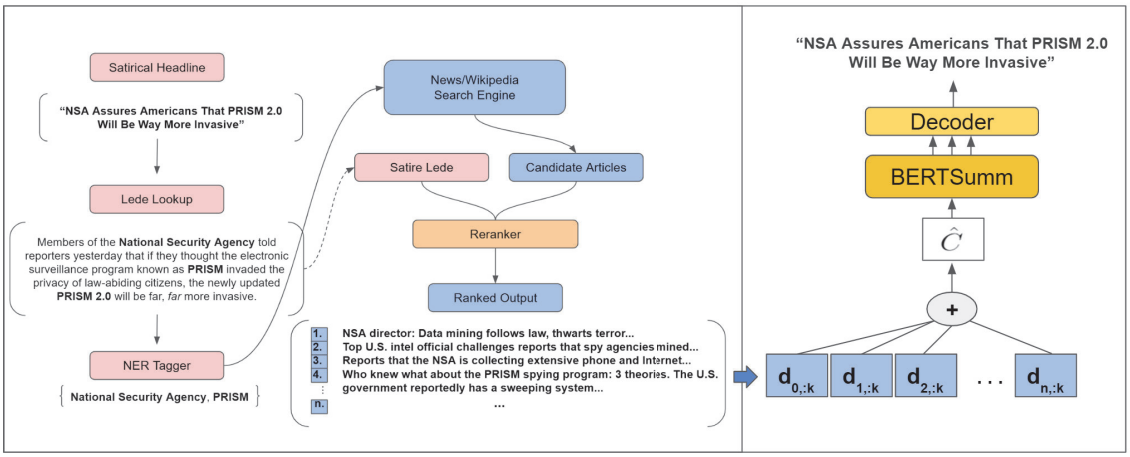

Figure 1: Pipeline for retrieving real-world textual context for a satirical headline. The extracted context is combined into a synthetic document, which is used as the input to a pretrained abstractive summarization model. The pipeline extracts named entities from the lede of the satirical article. These named entities are queried on Wikipedia and CNN. The results are then ranked by comparing their similarity to the original article across several metrics. We task the model with decoding the original satirical headline.

an incongruity between the underling event and sentiment expressed in the sentence. In contrast to pun wordplay or sarcasm, satirical headlines require a significantly richer context. Therefore, we explore satirical headlines as a testbed for humor generation that leverages richer contextual features.

Recent work in satire has explored the curation of corpora mapping from non-satirical to satirical forms. West and Horvitz (2019) built a corpus of unfunny headlines via a game that asks crowdworkers to make minimal edits that render satirical headlines unfunny and then analyzed structural differences between matched pairs of serious and satirical headlines. Taking an alternative approach, Hossain et al. (2019) introduced a corpus of news headlines with one-word edits While both of the aforementioned research efforts make inroads into understanding the rules underlying satire, both of the collected datasets are relatively small and curated. More importantly, both datasets do not consider the broader context that forms the basis of the joke.

Beyond puns and sarcasm, there has been little research on the generation of humorous text. Notable exceptions include Alnajjar and Hämäläinen (2018), who apply evolutionary algorithms and a master-apprentice approach to transform English movie titles into "creative" titles related to Saudi Arabia, and Winters et al. (2019) who automatically extract schemas from ranked lists of one-liner jokes. Rather than generation, the emphasis thus far has been on humor classification and ranking. Work by Shahaf et al. (2015) built classifiers to rank the funniness of submissions to the New Yorker Magazine caption contest, and Hossain et al. (2019) have introduced a headline-editing evaluation task.

Raskin (2012) notes that both humor detection and generation research have been hindered by "the difficulty of accessing a context sensitive, computationally based world model," but that "such difficulties are eliminated when the humor analysis is done with a system capable of capturing the semantics of text." Our work follows the second vein: we build on recent advances in contextual embeddings and summarizing architectures to extract meaningful features from text, and leverage them for conditional headline generation. We treat the satire generation task as one of taking a real-world text input, and generating a satirical headline output.

We propose a novel approach (as detailed in Figure 1) wherein we first construct a dataset of real-world context–satirical headline pairs in which the context is constructed by procedurally retrieving and ranking real-world stories, events and information related to the entities that appear in the original satirical headline. Then, we fine-tune BertSum, a state-of-the-art abstractive summarization architecture pretrained on news corpora, to encode the real-word context and generate the original satirical headline.

Our contributions are as follows: (1) we intro-

duce a novel approach for modeling satirical news headlines as conditioned on a real-world context, and an information retrieval pipeline for constructing the real-world context for a given real satirical headline; (2) we generate a dataset of more than 15K real-world context–satirical headline pairs for conditional humor generation; (3) we formulate satirical headline generation as an abstractive summarization task, mapping from a real-world text document to a humorous headline, and (4) we show that both the language and summarization models can be fine-tuned to generate headlines that people find funny. We find that summarization models best support satire generation by enabling humorous text that is both coherent and topical.

The context-based model appears to capture aspects of a "humor transformation" that include "edgy"/taboo topics and the satirical news register. Additionally, our model appears to learn how to mimic known principles of humor, including false-analogy, and to use incongruous relationships between entities and ideas. We compare the context-based approach to a context-free language modeling baseline. While the context-free approach can produce funny results, we find that people rate the context-based approach as generating funnier headlines. The context-based approach is also able to generalize, and generate headlines for unseen news contexts. Together, the results demonstrate that summarization models, which provide rich textual feature extraction, may offer important tools for future work in computational humor.

In machine generation of humor, it is important to control for the possibility that the humor is emerging from amusing generation failures. Our comparisons with non-satirical baselines evince that, fortunately, annotators are laughing *with* our model, not *at* it.

2 Our Approach

We now provide background on our methods.

2.1 Headline Representation

We model a satirical headline S_i as a function of an underlying latent joke J_i, which is, in turn, dependent on real-word context C_i,

$$S_i = \textsc{Headline}(J_i), J_i = \textsc{Humor}(C_i).$$

The goal of satirical news generation is then to map from context C_i to a satirical headline S_i.

2.2 Retrieving Real World Context

Hossain et al. (2019) attribute the lack of progress in computational humor research to "the scarcity of public datasets." In previous work, humans have been an essential in the labeling of these corpora. However, in the present work, we introduce an automatic, scalable pipeline for recovering background information for a satirical joke. We reconfigure the problem of matching headlines to a context as an unsupervised information retrieval task. The flow of our pipeline is displayed in Figure 1. We leverage the paradigm introduced by West and Horvitz (2019) of translating from funny text (a known satirical headline) to an unfunny related document. However, we expand this mapping to include a larger textual context.

In satirical headlines, the full "joke" may never be explicitly stated. However, the first line in a satirical article, referred to as the *lede*, contextualizes the headline by providing a grammatical, extended description and introducing named entities.

1. For a given satirical headline, we look up its lede, the first sentence in the body of the original satirical article.

2. We run the SpaCy Named Entity Recognition Tagger to extract named entities from the lede (Honnibal and Montani, 2017).

3. We then query these named entities on the news site CNN.com to retrieve contemporaneous news content published the week that the satirical article was written, along with all paragraphs from those entities' Wikipedia entries.

The output of our pipeline (Figure 1) is a dictionary that maps satirical headlines to a ranked list of Wikipedia paragraphs and CNN news articles. We then combine these results into an aggregate text document to serve as fodder for training and evaluating.

2.3 Building a Synthetic Document

To build our aggregate context document, we take the first k sentences from the top n most relevant ranked documents $\{d_0, ..., d_{n-1}\}$. This synthetic document of retrieved entity text serves as the approximation of the real world context:

$$\hat{C}_i = [d_0; ...; d_{n-1}] \approx C_i.$$

Once we have mapped every satirical headline to a corresponding context, we then train our model to approximate:

$$\hat{C}_i \mapsto S_i.$$

In other words, we train our summarization model to encode the contextual representation, augment it with pretrained embeddings, and then decode the original satirical headline.

2.4 Datasets

We retrieve documents for 15199 satirical Onion headlines.

To build our training and testing datasets, we include the first four sentences from the top two CNN articles and from the top three remaining documents by rank. (This design biases our contextual document towards news content, when it is available). We then trim these aggregate contexts, which we refer to as synthetic documents, down to approximately 512 tokens. We experimented with several document-creation schemes, including building a larger corpus by stochastically sampling from the different text sources.

The resulting dataset comprises over document–headline pairs. We employed human annotators to confirm that our retrieved documents are regularly relevant to the original satirical article. These results are included in the Appendix (See A.1).

For our news-based context-free baseline model, we used the roughly 10K real news headlines from the Unfun.me corpus. As an additional baseline, we trained the model on the 2758 unfunned-satirical pairs provided in the corpus. Each satirical headline had multiple "unfunned" candidates, so we ensured that no such duplicates appeared in both the train and test corpora.

We used an 85-15 train-test split for all satire models.

2.5 Models

We leverage recent breakthroughs in document summarization by employing the abstractive summarization model introduced by Liu and Lapata (2019)(Nallapati et al., 2016). The architecture is state-of-the-art on the CNN-DailyMail Test.[1] Their architecture, BERTSum, augments BERT (Devlin et al., 2018) (Bidirectional Encoder Representations from Transformers) with sentence embeddings to build document-level encodings. These encoder embeddings are then fed to a Transformer decoder (Vaswani et al., 2017). The BERTSum encoder is then secondarily pretrained on an extractive summarization task before finally being fine-tuned on an abstractive summarization task. For our work, we fine-tuned a BERTSum model pretrained for abstractive and extractive summarization on 286K CNN and Daily Mail articles.

We settled on three main fine-tuning schemes, which yielded three distinct context-based models. For the **Encoder-Weighted-Context** (E-Context) model, we trained the encoder and decoder with learning rates of 0.002 and 0.02, respectively. For the **Abstractive-Context** (A-Context) model, we trained the network on contexts that had been preprocessed by the pretrained abstractive summarizer. For **Decoder-Weighted-Context** (D-Context), we trained the decoder with a learning rate of 0.02, and an encoder with learning rate 0.00002. For all models, we used batches of size 200, a warmup of 500, and decayed the learning rate using the function implemented by Liu and Lapata (2019).

We applied these varied schemes as a means of exploring the relationship between learning a new encoder representation and fine-tuning a new 'satirical' decoder atop the pretrained summarization encoder module. Additionally, we include the abstractive approach to test the value of a more concise document formulation.

For the context-free baselines, we fine-tuned Hugging Face's large GPT-2 model on the satirical headlines in our corpus (Radford et al., 2019; Wolf et al., 2019). We also fine-tuned the GPT-2 model on a corpus of 10K real news headlines from the Unfun.me corpus.

3 Experimental Design

We tested our models by sampling headline generations and evaluating their quality via crowdsourcing.

3.1 Satire Generation

We began by greedily generating headlines from our baseline models: the GPT-2 context-free satire model and the GPT-2 context-free news model. Since language models only condition on the previous tokens in a sequence, generating diverse outputs requires random sampling. However, we found that common approaches (such as Top-k and Top-p sampling) rapidly degraded headline quality. Thus, from our validation set of 1955 satirical head-

[1] https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail

lines collected from The Onion, we extracted the first two words from each headline, and used these two words as prompts for greedy generation. For the context-based modes, we generated headlines by feeding in the synthetic documents from our test set. In contrast to our language-model baselines, our context-based model never sees any segment of the original satirical headline.

3.2 Satire Generation Evaluation

We employed human annotators to evaluate the performance of different models on the satire-generation task. Workers on Amazon Mechanical Turk answered three questions for every generation: (1) Is the headline coherent? (2) Does the headline sound like The Onion? and (3) Is the headline funny? To control for funniness induced by incoherence, we instructed annotators to mark all 'incoherent' headlines as not funny.

For each generated headline, we received three unique annotations. To qualify, annotators must have come from the United States, have had more than 50 annotations approved in previous studies, and have had an annotation approval rating higher than 95%.

We had 750 headlines annotated for each model. Labels were determined by majority agreement $(2/3+)$.

4 Results

This section describe the results of our evaluation.[2]

Table 1: Model Comparison

Model	Coherence	Onion	Funny	F | C
Onion (Gold)	**99.5%**	**86.6%**	**38.2%**	**38.4%**
Satire GPT-2	86.5%	57.7%	6.9%	7.9%
News GPT-2	**89.2%**	36.9%	2.4%	2.7%
D-Context	88.4%	**58.8%**	**9.4%**	10.4%
E-Context	80.2%	57.8%	8.7%	**10.8%**
A-Context	85.3%	54.9%	8.8 %	10.3%

[2]We excluded the model trained on the Unfun.me corpus, as only 2.5% of its generations were rated as funny, and 64% of generations were simply duplicates of the original input context. The redundant generations likely resulted from the small corpus size, and the significant word-overlap between the 'unfunned' input and labels.

4.1 Quantitative Results

Table 2 contrasts the performance of the different headline-generation techniques as rated by the annotators. Coherence, Onion and Funny columns describe the majority vote among the three annotators for the category. Column $F|C$ contains the probability of a headline being rated funny, given that it is rated coherent. Because all funny annotations were also by default rated coherent, we computed $F|C$ by dividing the number of Funny headlines by Coherent headlines.

We also collected annotations for original Onion headlines, which we compared to the results for each of our models. As expected, expert-generated satirical headlines from The Onion perform best on the Coherence, Onion and Funny metrics, as well as $F|C$. In contrast, the news-based model was judged as Coherent, but not rated well on the humor-related metrics.

Importantly, the D-Context model achieved the highest Funny rating among all models, followed by the E-Context model. (The former had a Funny score $\sim 4\times$ that of the News GPT-2 baseline). Additionally, the context-based models received higher Funny scores than the Satire GPT-2 language model (a 2% increase, approximately). This delta is especially impressive given that the context-free satirical language model was prompted with the first two words of a true satirical headline.

These results support the claim that context-based models more regularly produce funny generations than the context-free approaches. Additionally, all satire-trained models substantially outperformed the News GPT-2 baseline, providing critical evidence that the humor judgments are not simply due to awkward machine-generated language, but are a consequence of the fact that the models are learning to generate coherent, humorous text.

While the D-Context was rated over 8% more coherent than the E-Context model, a smaller fraction of the coherent generations are rated Funny. Our examinations of these generations reveal that primarily fine-tuning the decoder on satire may lead to coherent, but more standardized generations that are less conditioned on context. However, we measured the quantitative similarity of generations for all context-based models to their input document (See A.2), and found that all models produced generations that were as similar to their input document as the original satirical headline.

We will now examine the patterns that character-

ize these generations.

4.2 Qualitative Analysis

In our initial analyses of the characteristic behaviors of the models, we have observed a transformation from events referenced in the context into a "newsy" register, the introduction of expressions of uncertainty, sweeping generalizations, and incongruous juxtapositions (see Figure 2).

The adoption of a newsy tone is readily apparent; the model invents "studies" and "reports" even when none are mentioned in the original context. Additionally, common forms include "X announces/unveils Y," where X and Y are extracted from the context, or are particularly polarizing topics from The Onion corpus, like "abortion" or "sex."

The model also refers to general entities often referenced by Onion writers. These include common satirical terms for everyday people, like 'area man.' When the model employs these characters, it tends to decode out more observational headlines, like *area man just wants to know what he 's doing*, that are less related to the given context. Our Decoder-Weighted-Model exhibited this behavior more often.

The context-based generations also introduce apparent "incongruities" in a variety of ways. For example, the models catch the satirical news trick of juxtaposing a 'study' with an unscientific remark. For example: *study finds americans should be more obese by now*. Another, most obvious example of incongruity is the mention of absurd, yet contextually relevant events (e.g. *study finds majority of americans still in oil spill*).

However, the most fascinating cases are when the reality articulated in the input context is inverted. For example, *god admits he's not the creator* when the context very much states that He is. Similarly, in Figure 2, we see *scientists discover that oil spills caused by natural causes*, when the context argues quite the opposite. This juxtaposition works as a humorous construction and suggests that the model has latched onto something like a general principle.

We submitted these two generated headlines, along with others, to the Brown Noser[1], Brown University's campus satirical newspaper:

- *God Unveils New Line of Sex*
- *U.S. Asks Pugs If they Can Do Anything*

Staff writers rated the latter as high as or better than 73% of writer submissions. Both were accepted for publication and express several aspects of observed humor transformation captured by our context-based models. The first juxtaposes newsy language (for example, Unveils, New line of, U.S.) with incongruous entities like 'God' and 'Sex.' The second relates pugs to U.S. governmental affairs. The latter article was published with an accompanying article written by a human satirical writer (See A.3).

4.3 Sensitivity Analysis

The latent space of Transformer-based architectures is fundamentally difficult to analyze. However, our summarization approach gives us the ability to probe the relationship between context and output: We can perturb the input context and examine the resulting change to the decoding headline. Thus far, we have observed that our model is less sensitive to changes to the context in the form of adjectives or negations than it is to changes in the entities. Additionally, key terms in the context can activate certain headlines. For example, mentions of the Royal family tend to prompt the same original headline: *Royal baby born in captivity*. However, in other instances, the entire tone of the resulting headline can be changed by single verb substitution. For example:

> "harriet hall of science based medicine reviewed the film in an article entitled ' does the movie fed up make sense ? ' . the film [**makes/disputes**] the claim that drinking one soda a day will increase a child's chance of becoming obese by 60 %"
>
> 1. **makes**: study finds americans should be more obese by now
> 2. **disputes**: study finds average american has no idea how to get overweight

In both cases, the model introduced a fictional study. However, the latter appears to capture the uncertainty around the disputed claim that one can become obese by drinking a soda. Future work is necessary to explore the relationship between context and output.

[1] http://thenoser.com/

Input: a creator deity or creator god [often called the creator] is a deity or god responsible for the creation of the earth , world , and universe in human religion and mythology . in monotheism , the single god is often also the creator . a number of monolatristic traditions separate a secondary creator from a primary transcendent being , identified as a primary creator...
E-Context: god 's name a big hit / god admits he 's not the creator
D-Context: god 's god calls for greater understanding of all the things
A-Context: god admits he 's not a good person
Onion: Biologists Confirm God Evolved From Chimpanzee Deity
GPT-2 Satire: biologists confirm
GPT-2 News: biologists confirm human ancestor

Input: the jet propulsion laboratory is a federally funded research and development center and nasa field center...on 26 november 2011 , nasa's mars science laboratory mission was successfully launched for mars ... the rover is currently helping to determine whether mars could ever have supported life , and search for evidence of past or present life on mars ...
E-Context: nasa announces plan to put down mars / nasa announces plan to hunt mars
D-Context: nasa launches new mission to find out what life is doing
A-Context: mars scientists successfully successfully successfully successfully
Onion: Coke-Sponsored Rover Finds Evidence Of Dasani On Mars
GPT-2 Satire: coke - a little too much
GPT-2 News: coke - the new 'dancing with the stars'

Input: the boston globe called for a nationwide refutation of trump's 'dirty war' against the news media, with the hashtag enemy of none. more than 300 news outlets joined the campaign. the new york times called trump's attacks 'dangerous to the lifeblood of democracy...
E-Context: trump vows to destroy all his words / trump: ' i 'm not the best guy in the world '
D-Context: trump vows to destroy all the things he 's doing
A-Context: trump : ' we 're not going to let people know what it is '
Onion: Trump's Attacks On The Press
GPT-2 Satire: trump'sick and tired of hearing' trump say
GPT-2 News: trump'sick of being in the middle of a fight'

Input: a 2014 study of the effects of the oil spill on bluefin tuna funded by national oceanic and atmospheric administration...found that tuna and amberjack that were exposed to oil from the spill developed deformities of the heart and other organs that would be expected to be fatal or at least life-shortening . the scientists said that their findings would most likely apply to other large predator fish and even to humans.. bp was guilty of gross negligence and willful misconduct . he described bp's actions as 'reckless ...
E-Context: study finds majority of americans still in oil spill
/ study finds majority of tuna spills now in danger of human suffering
D-Context: scientists discover that oil spills caused by natural causes
A-Context: report : bluefin fish may have been killed by fish
Onion: Shrimp Boat Captain Worn Out From Long Day Of Putting Human Face On Crisis
GPT-2 Satire: shrimp boat to be built in new york
GPT-2 News: shrimp boat sinks in gulf

Figure 2: A sample of documents (abbreviated) and resulting generations. These generations incorporate entities from the context while maintaining Onion-like language. This includes irreverent, observational tone, and the addition of frequently Onion corpus terms like "study" and "announces." We also observed that generations could invert facts expressed within the context (e.g. *God admitting he is **not** the creator*, or *oil spills result from natural causes*). We observe the decoder-weighted model resorting to more casual, repetitive language (e.g. *"all the things..."*).

4.4 Topical Generation for News Stories

We can apply our model to novel news stories. While all of our training headlines were collected before COVID-19 was declared a pandemic in March 2020, our model shows an ability to generalize to pandemic-related news stories and generate topical headlines (Figure 3). We preprocessed the beginning of CNN articles from April 2020 with the pretrained BERTsum model to generate concise abstracts, and then input these summaries into our networks. Our models appear to condition on these new contexts and generate related satirical headlines.

Input: president donald trump doubled down on an unproven therapy for the novel coronavirus . without citing evidence , he said it's a "great " and "powerful" anti-malaria drug " . trump said it ' s

- trump doubles down on prescription drug that can cause coronavirus

Input: questions over whether downing street was fully transparent about the prime minister's health . but important issues risk overshadowing the true picture of the uk's struggle against coronavirus . the uk is on a similar, grim trajectory as the uk is

- nation 's love of coronavirus now a little more complex

Input: president donald trump announced tuesday he is halting funding to the world health organization while a review is conducted . trump said the review would cover the " role in severely mismanaging and covering up the spread of coronavirus " . trump has sought to assign blame elsewhere , including at the who and in the news media

- world health organization unveils new " ' plan to cover up coronavirus

Figure 3: We preprocessed CNN articles from April using the pretrained abstractive summarization model provided by Liu and Lapata (2019). Our approach appears to generalize to these novel news contexts.

The resulting generations incorporate named entities from the context, and embed them in a humorous generation.

5 Future Work

We intend to further investigate the latent and embedding spaces of our model, and hope to better elucidate the neural-logic that transposes everyday events into humorous language.

Additionally, our context-driven approach allows us to examine the relationship between real-world, input events, and the resulting satirical output. We plan to continue probing this relationship, and to refine our understanding of the processes underlying our generations, and how their relationship to real-world events can be interpreted within proposed theories of humor, including the Script-Based Semantic Theory of Humor (See A.4).

While we treat document retrieval as a prepossessing step, we can also explore including retrieval in end-to-end training, as employed by Guu et al. (2020).

Lastly, we are fascinated by the potential for leveraging other text-based contextual representations, ranging from alternative types of document and longer-form text, to graph-encoded representations of events and entities. These approaches can provide alternative blends of depth, concision, and structure.

6 Conclusion

We introduced a methodology and formalization for modeling satirical news headlines as conditioned on a real-world context, and presented an information retrieval pipeline for constructing that context. We found that pretrained abstractive summarization models provide powerful feature extractors atop these rich contextual representations. Conditioning on such contextual documents enabled the generation of topical satire that people determined to be funnier than headlines generated via context-free methods. Additionally, we discovered that the approach generalizes to new topics and circumstances.

Moving beyond the focus on generating satirical headlines, we believe that variations of our approach are broadly applicable to tasks ranging from information retrieval and dialogue to reading comprehension. Our work provides further evidence that neural architectures, augmented with task-relevant contextual information, have the potential to reason about sub-textual concepts, including the subtleties of humor.

Acknowledgments

Special thanks to Ellie Pavlick for providing valuable feedback through the course of this research. Additionally, we are grateful to Jacob Lockwood, an editor of the Brown Noser Satirical Newspaper, for offering the necessary human talent to parlay our model's output into a full satirical article (Figure 4).

This work is supported in part by the ONR PERISCOPE MURI award N00014-17-1-2699.

References

Khalid Alnajjar and Mika Hämäläinen. 2018. A master-apprentice approach to automatic creation of culturally satirical movie titles. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 274–283, Tilburg University, The Netherlands. Association for Computational Linguistics.

Salvatore Attardo. 2001. *Humorous Texts: A Semantic and Pragmatic Analysis*. Mouton de Gruyter.

Salvatore Attardo. 2014. Humor in language. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *HUMOR: International Journal of Humor Research*, 4(3/4):293–348.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *ArXiv*, abs/2002.08909.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. 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)*, pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> 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)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

N. D. Goodman J.T. Kao, R. Levy. 2016. A computational model of linguistic humor in puns. *Cognitive Science*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP/IJCNLP*.

Rod A Martin. 2010. *The Psychology of Humor: An Integrative Approach*. Elsevier.

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 6144–6154, Hong Kong, China. Association for Computational Linguistics.

John Morreall. 2016. Philosophy of humor. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, winter 2016 edition. Metaphysics Research Lab, Stanford University.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GulÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

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

Victor Raskin. 1985. *Semantic Mechanisms of Humor*. Reidel.

Victor Raskin. 2012. A little metatheory: Thought on what a theory of computational humor should look like. In *AAAI Fall Symposium: Artificial Intelligence of Humor*.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Robert West and Eric Horvitz. 2019. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". *CoRR*, abs/1901.03253.

Thomas Winters, Vincent Nys, and Danny De Schreye. 2019. Towards a general framework for humor generation from rated examples. Proceedings of the 10th International Conference on Computational Creativit, pages 274–281. Association for Computational Creativity; University of North Carolina at Charlotte, USA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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, Melbourne, Australia. Association for Computational Linguistics.

A Appendices

A.1 Dataset Evaluation

We employed 62 annotators via Amazon Mechanical Turk to evaluate the relevance of a sample of our retrieved documents to the original satirical articles. Annotators must (1) have come from the United States, (2) have had more than 50 annotations approved in previous studies, and (3) have an annotation approval rating higher than 95%. We evaluated the retrieved documents for 1500 of the satirical headlines. For each article, we present three different annotators with the satirical headline, lede and the top five retrieved documents from our ranker, each trimmed to contain only the title and the first three sentences of content. We asked annotators to evaluate how "relevant" each context document is relative to the satirical article, with relevant being defined as either (1) the context document covering the event discussed by the satirical article, or (2) the context document provides useful context to understand the satirical article. Table 2

Table 2: Top 5 Retrieved Document Relevance

Consensus	Top 1	Top 2	Top 3	Top 4	Top 5
≥ 1	86.7%	96.3%	97.7%	98.5%	99.0%
Majority	48.3%	61.5%	69.1%	74.5%	80.7%

describes the result of the evaluation. The first row indicates the percentage of all headlines where at least one annotator considered one of the top n retrieved documents to be relevant. The second gives the fraction of headlines where $2/3+$ annotators agreed that at least one of the top n retrieved documents was relevant to the original satirical article.

For approximately 80% of the headlines, a majority of annotators deemed at least one of the top five retrieved documents to be relevant. For 99% of headlines, at least one annotator believed that one of the top five retrieved documents were relevant. These results indicate our automatic pipeline systematically retrieves relevant articles. We believe that we could further improve retrieval accuracy via (1) filtering out satirical headlines less likely to have contextual information (e.g. "Nervous New Driver Going To Stick To Sidewalks Until He's More Confident"), (2) human labeling of the corpus, and/or (3) training a reranker with human labels on a subset of the corpus.

A.2 Generation Relevance to Context

To evaluate the extent to which our generations incorporate the input context, we performed quantitative evaluations of similarity. For each contextual model, we computed the Jaccard index between each headline in the test set and its corresponding input context. We then normalized these scores by the *average* Jaccard index between that headline and every context. This metric captures similarity to the specific context, relative to an arbitrary context.

Table 3: Generation-to-Context Similarity

Model	Normalized Jaccard (Avg.)
Onion	5.0
Summarizer	9.9
D-Context	6.2
E-Context	6.7
A-Context	6.0

Table 3 contains the generation-to-context similarity score for each model. These values were computed by tokenizing the headline and input, stemming tokens, and filtering out stopwords.

Critically, we see that all models receive scores significantly greater than 1. This result indicates that the headlines are substantially more similar to their input context than to an arbitrary one. As expected, the pretrained summarization model (with no satirical fine-tuning) receives the highest generation-to-content similarity score. Additionally, D-Context, which has an attenuated encoder learning rate, is less contextually relevant than E-Context.

However, most notably, all models produce generations that are more relevant to the input context than to the original satirical headline. These quantitative evaluations reveal that our architecture's generations are not only funny, but apt.

A.3 Satirical Newspaper Publication

After our headlines were accepted into the Brown Noser's April 2020 issue, Jacob Lockwood, an undergraduate satirical writer, volunteered to write an accompanying article (See Figure 4). We are enthusiastic about the potential for future AI–Human satirical collaborations.

U.S. Asks Pugs If They Can Do Anything

If you don't like this, take revenge on your nearest computer. This is the world's first AI-generated headline.

By Jacob Lockwood
Published Friday, April 24th, 2020

In a statement released Tuesday, citizens of the United States asked pugs if they can do anything, and if so, what? "Is there anything at all that you can do?" the amassed nation collectively wondered aloud while staring at the country's population of pugs, who were all simultaneously panting with their tongues out. "What exactly is your purpose? Do you have skills? Or are you not useful for anything? Please, pugs, tell us what you do. We're all very unsure." Responding to the inquiry of all Americans, the gathered pugs began to drool.

Figure 4: A satirical headline generated by our model was published in the Brown Noser Satirical Newspaper, and accompanied by an article written by a human satirical writer. The article can be found here.

.

A.4 The Script-Based Semantic Theory of Humor

The Script-Based Semantic Theory of Humor (SSTH) (Raskin, 1985) provides a framework for interpreting our model's output. According to SSTH, for a text to be "funny" it must satisfy the following conditions:

1. The text is compatible, fully or in part, with two different scripts.

2. The two scripts with which the text is compatible are opposite in a special sense (Raskin, 1985).

Many of our generations exhibit these properties. For example, consider the generated headline from Figure 2:

God Admits He's Not The Creator

Within this generation, there is at least one possible script opposition:

1. God as the divine creator (as described in the context).

which opposes the script:

2. A person making an admission to the media.

These opposing scripts are related via the logical mechanism of false-analogy: God is a famous entity, and thus likely to appear in the news, but God is also a deity, *not* a person, and is infallible (West and Horvitz, 2019; Attardo and Raskin, 1991).

Consider another example generation:

Royal Baby Born in Captivity

With opposing scripts:

1. The royal baby is a human.

2. The baby is, like an animal, born into captivity.

These two scripts are again related through the mechanism of false-analogy: The royal baby is a baby, like an animal born in captivity. However, the baby is human, making it unlikely to be born in captivity.

It remains unclear whether our architecture is explicitly modeling "opposing scripts" in its latent space, or rather translating entities from the context into headlines with Onion-style language. However, in either case, our approach is incorporating contextual entities, using contextual information, and generating text that imitates the properties of humor.