

High Quality Real-Time Structured Debate Generation

Eric Bolton
Columbia University
edb2129@columbia.edu

Alex Calderwood
Columbia University
adc2181@columbia.edu

Niles Christensen
Columbia University
nnc2117@columbia.edu

Jerome Kafrouni
Columbia University
jk4100@columbia.edu

Iddo Drori
Columbia University
idrori@cs.columbia.edu

Abstract

Automatically generating debates is a challenging task that requires an understanding of arguments and how to negate or support them. In this work we define debate trees and paths for generating debates while enforcing a high level structure and grammar. We leverage a large corpus of tree-structured debates that have metadata associated with each argument. We develop a framework for generating plausible debates which is agnostic to the sentence embedding model. Our results demonstrate the ability to generate debates in real-time on complex topics at a quality that is close to humans, as evaluated by the style, content, and strategy metrics used for judging competitive human debates. In the spirit of reproducible research we make our data, models, and code publicly available.

1 Introduction

Automatic generation of human-level debates has important applications: providing coherent deconstructions of given topics, assisting in decision-making, and surfacing previously overlooked viewpoints as demonstrated in Table 1. Debate generation would enable policymakers and writers to improve their rhetorical abilities by practicing against an automated agent in real-time.

The field of debate is closely related to linguistic negation, which has studied logical rules for forming various types of negation (Malpas and Davidson, 2012b), at the low level of words and sentence. In this work we use a large dataset and pre-trained neural language model to synthesize long debates with a high-level linguistic structure. Our work focuses on informal logic, the study of arguments in human discourse, rather than in rigorously specified formal languages (Malpas and Davidson, 2012a).

Source:	Income inequality makes people unhappy. Inequality is bad for everyone.
Response:	The economic outcome of an increase in income is not a good factor in determining the effects of starvation and poverty.
Source:	Many religious teachings promote violence.
Response:	Religious protesters have a strong history of engaging in violence against violence. Many of the people who oppose the violence are often on a case-by-case basis.

Table 1: Sample results: novel responses generated to be contradictory.

Machine debate generation must meet several requirements: the need to respect long-distance dependencies while maintaining semantic coherence with the prompt, the need for consistent themes and tone over long durations, and finally, the need for adherence to grammar rules.

These requirements entail compelling, thoughtful responses to a given writing prompt. Models that excel at this task must understand the low-level grammar of the language, but also need to incorporate an understanding of the relationships between sentences. Thus, existing story generation models constitute a natural tool for generating arguments with an understanding of what it means to refute or support an argument’s point.

1.1 Kialo

We collected a data set from Kialo (Kialo, 2019), an online debate platform, and trained a text gen-

eration model (Fan et al., 2018) to generate arguments in favor or against given debate prompts. Users submit responses (pros and cons), which are then approved by a moderator. Other users can then respond to those responses as though they themselves were prompts, forming a debate tree.

We leveraged the tree structure of the Kialo dataset along with its existing pro/con annotations to develop a novel self-supervised corpus creation strategy, using the insight that the metadata constraints we used to define an argument (our generative unit) effectively enforced a high-order grammar on the output. We experiment with this type of tree traversal strategy to model multi-turn text generation. This strategy is applicable to other seq2seq models that model long-form text generation.

1.2 Contributions

In this work, we describe a system which learns to participate in traditional debates by modeling different sequences of argument nodes through tree-structured debates to generate prompts and responses *corresponding to sources and targets in the translation task, such*

Our main contributions are:

- **Defining debate tree and paths for generating text:** Enforcing a high level structure and grammar, defining complex arguments, multi-turn debates, as described in Section 2.1 and 2.2.
- **Multi-turn debate generation:** An extension of the prompt to response generation paradigm to enable multi-turn generation, as described in Section 2.3.
- **Measure of debate quality:** as evaluated with criteria used in debate competitions: content, style, and strategy, repeated over multiple rounds.

Based on these key contributions, our results demonstrate near-human quality debates as shown in Section 3.

1.3 Related Work

Our goal is to create debates that perform at a high-level according to the same evaluation metrics as human debates. Quantitatively assessing the quality of debates and arguments is a challenging task. Recent work evaluates dialogue (Kulikov et al.,

2018) and stories (Purdy et al., 2018) by humans using Amazon’s Mechanical Turk (AMT). We also use AMT for human evaluation of debates, based on the criteria commonly used in debate competitions (Speech and Competition, 2018).

Our focus is debate generation, which is structured in multiple rounds. The state of the art in structured text generation focuses on stories rather than debates, and is based on generating sequences of predicates and arguments, replacing entities with names and references (Fan et al., 2019). Other structuring strategies include using outlines (Drissi et al., 2018), plans (Yao et al., 2018), or skeletons (Jingjing et al., 2018). All of these works build upon significant advances on the problem of text generation using hierarchical models (Fan et al., 2018). These methods can be improved and extended by using very large language models which have shown excellent results in language generation and reading comprehension (Devlin et al., 2018; Peters et al., 2018; Radford et al., 2018, 2019).

The performance of these neural language models has relied on curating large datasets for different tasks, such as question answering (Rajpurkar et al., 2016) and reading comprehension (Yu et al., 2018; Clark and Gardner, 2018). In contrast, our work creates a large dataset specifically for the task of generating debates. IBM’s Project Debater (IBM, 2019) also shares the same focus, however they are driven by speech and do not take advantage of debate specific datasets. There are a number of online debate hosting platforms that contain pro/con metadata (ProCon, 2019; iDebate, 2019). We use Kialo, which uniquely combines this metadata with a tree structure that models a long multi-turn debate, depending on the parsing strategy.

2 Methods

We define a framework for generating debates from a tree-structured corpus, as illustrated in Figure 1: (i) define and collect debate tree corpus, as described in Section 2.1; (ii) build debate paths, which are content/target pairs, as described in Section 2.2; (iii) train the debate encoding on all generated debate paths, as described in Section 2.3.

2.1 Debate Trees

A debate tree is a tree with one extra bit of storage for each node, its stance: either ‘Pro’ or ‘Con’ depending on whether the node supports or refutes

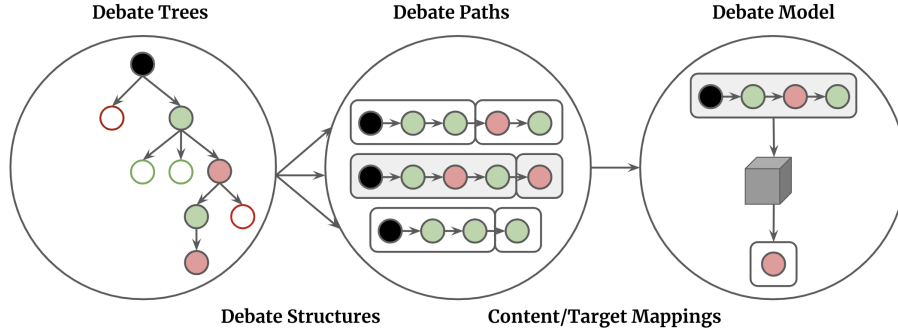


Figure 1: Our debate generation framework. We begin with debate trees on the far left, from which we extract all paths through the tree which correspond to meaningful debates, creating a dataset in the central bubble. On the far right, this new dataset is used to retrain an existing model to debate. In this paper, the retrained model is the one developed by Fan et al. (2018). In our diagram, each green bubble supports the point above it, while each red bubble argues against the point above it.

Parsing Strategy	Prompt	Response
Supportive Arguments	[Pro Con][Pro]*	[Pro]+
Contradicting Arguments	[Pro Con][Pro]*	[Con][Pro]*
Complex Arguments	[Pro Con][Pro]*	[Pro Con][Pro]*
Multi-Turn Debates	[Pro Con][Pro]*([Con][Pro]*)*	[Con][Pro]*

Table 2: Description of debate structures, expressed through regex notation.

its parent argument. A debate is a path on this debate tree, for example from the root to a leaf, as shown in Figure 2. An argument is defined as one node in tree.

We extract debate tree structures from a public debate site (Kialo, 2019) which allows crawling, and filter for English debates (Langdetect, 2014). When Kialo users referenced different parts of the debate tree, which rarely happened, we copy the text in the referenced node into the referring node.

2.2 Debate Paths

The debate tree can be parsed in different ways depending on the desired properties of the resulting corpus. We present four parsing strategies. For each, we provide a regex-like description corresponding to the prompt and the response of the debate, as described in Table 2.

The majority of paths through the tree would correspond to nonsensical debates. In order to model sensible and internally coherent debates in which each modeled participant has a clear viewpoint, we developed the following parsing strategies:

Supportive Arguments - a prompt and a response supporting it.

Contradicting Arguments - a prompt and a response refuting it.

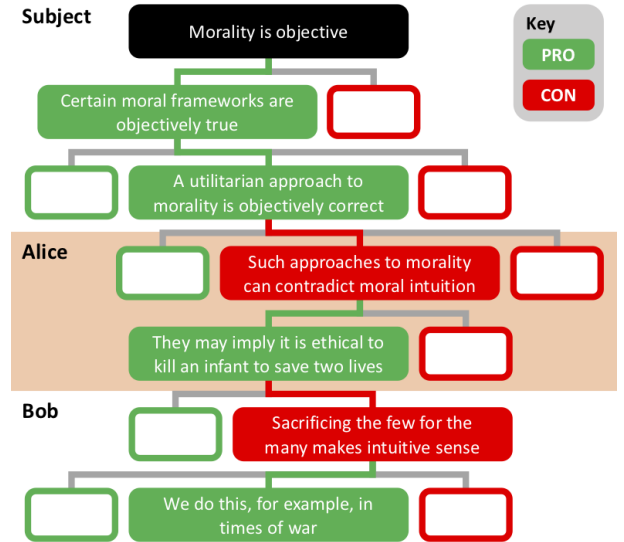


Figure 2: An example of a path through a Kialo debate tree. Green boxes indicate arguments that support their parents (Pro), while red boxes indicate arguments that contradict their parents (Con).

	Complex	Supportive	Contradictory	Multi-Turn
# Training examples	432,999	222,852	211,355	459,643
# Test examples	45,471	21,946	17,677	30,607
# Validation examples	38,779	20,810	14,485	35,916
Prompt dictionary size	20,535	14,691	16,690	50,626
Response dictionary size	28,313	21,907	19,534	55,359

Table 3: Description of datasets created from KIALO using each tree-traversal strategy.

Complex Arguments - a prompt and a response that may either support the prompt or refute it. This is the union of all arguments generated by the supportive and contradicting arguments approaches.

Multi-Turn Debates - a prompt containing a back-and forth debate between two speakers, separated by a specific token. Each participant is constrained to always disagree with the previous statement made by their opponent. This approach allows for the generation of debates of arbitrary length.

2.3 Debate Model

Once the debate paths are extracted, they are used to train a text generation model. In this work we retrained the model developed by Fan et al. (2018), but another model, such as the one developed by Devlin et al. (2018), could be substituted in its place with minimal changes.

To determine what would yield the best generation performance, we trained models with our complex arguments dataset under various conditions. We filtered for only Pro responses and Con responses, and compared results between Fusion (Sriram et al., 2018) and plain seq2seq models (Gehring et al., 2017).

We limited our vocabulary to tokens appearing 10 or more times in the training set and replaced words that don’t fit this criterion with an unknown token. We used named entity recognition (NER) (Honnibal and Montani, 2017) to limit the number of unknown tokens on our complex arguments datasets, to improve model coherence. Finally, we modeled an end of argument token to delimit arguments. We used 5% of the debates for our validation set, and 5% for our test set. A detailed description of our resulting dataset is shown in Tables 3. We implemented a custom generation approach to use previously generated responses as input prompts for real-time generation of new responses. This enabled generation of multi-turn debates using the seq2seq model trained on our de-

bate data.

3 Results

Our results demonstrate the ability to generate debates in real-time on complex topics, as shown in Table 4. Evaluating complex debates is a difficult task. Perplexity evaluation, shown in Table 5, is limited, and does not capture debate quality well. Therefore we use human evaluation of our generated debates (Kulikov et al., 2018). Specifically, we use Amazon Mechanical Turk (Amazon, 2019) for evaluating multi-turn debates, with 119 distinct turkers and a total of 526 evaluations times 4 criteria on a 1-4 scale.

We compare our generated debates with human debates, by comparing debate paths from the test set with debate sequences predicted by our model. We generated lengthy debates of between 5-15 back-and-forths and evaluated debates of length 10, with human debates of the same length, preventing any bias from the Turkers for longer or shorter debates. We asked humans to rank 50 human-created and 56 automatically-generated debates, where each debate was rated by 5-10 separate reviewers for a total of 2,104 answers as shown in Table 6. Reviewers were asked to evaluate the style, content, strategy, and overall quality of each debate, which are the same categories used to judge competitive debaters in the US National Debate Tournament (Speech and Competition, 2018). Our results demonstrate generated debates which are evaluated to be close to human level performance as shown in Table 6.

3.1 Implementation

Our model was trained and run on a virtual machine running Ubuntu 16.04 on 8 vCPUs and 2 NVIDIA Tesla K80 GPUs. The machine had access to 52 GB of memory and 100 GB of storage. Training time is 3 days on this machine. Given a new subject for debate, average computation time of a new argument is around 3 seconds, which is acceptable in real-time debates. In the spirit of

Subject:	All religions should be banned on a global scale.
Generated debate:	
Alice:	Religion is a global problem that cannot be solved in a bad way. Religion is a tool used to challenge human rights. It does not exist to be as a way to be a good thing for humans in order to use the world as a form of language and the existence of its own religion. For example, religion is not the only way to prevent a world from a different religion
Bob:	Religion is a function in which morality involves the existence of something, rather than just the existence of one another. Religion is a personal and philosophical concept
Alice:	Religion is a form of tradition. Simply because one group may be sacrificed or not. Everyone can make a person considered to be good at all
Bob:	Religion may be a good thing for many people to commit violence, but rather the life of one person
Alice:	Religion is not a way to measure a belief that it is inherently acceptable to kill a sandwich if the one needs it gets to live
Bob:	Not all people are the ones doing anything of anything. Religion is the gift of that they disagree about. This is the same as the people who value the community may value their lives through little reason not to process it or destroy people or other people to live
Alice:	Moral values are a way to value moral and ethical consequences, no one simply wants to do what is right, not worth how you want
Bob:	If people have a duty to help those people to value their life they will have to act in the interest of their place rather than forcing them to do so without harming their religious life

Table 4: Multi-turn debate generated from human prompt “All religions should be banned on a global scale.”

Model type	Strategy	Perplexity
seq2seq	No NER, Complex	73.16
seq2seq	No NER, Supportive	54.68
seq2seq	No NER, Contradicting	60.72
seq2seq	NER, Complex	80.87
seq2seq	NER, Supportive	58.82
seq2seq	NER, Contradicting	77.39
Fusion	No NER, Complex	70.41
Fusion	No NER, Supportive	43.06
Fusion	No NER, Contradicting	60.33
Fusion	NER, Complex	83.77
Fusion	NER, Supportive	77.04
Fusion	NER, Contradicting	79.98

Table 5: Average perplexity over all generated debates.

Criteria	Human	Generated
Style	3.273 \pm 0.722	3.154 \pm 0.868
Content	3.392 \pm 0.626	3.289 \pm 0.770
Strategy	3.412 \pm 0.596	3.133 \pm 0.821
Overall	3.512 \pm 0.596	3.193 \pm 0.824

Table 6: Mean and standard deviation of Mechanical Turk evaluations. Ratings were on a 1-4 scale.

reproducible research we make our data, models, and code available (Bolton et al., 2019).

3.2 Limitations

Unknown tokens present in the output were replaced with underscores representing an underline, and turkers were shown a note explaining that certain words would be omitted and replaced with an underline. Therefore, these missing words may have skewed the Turkers’ judgments on responses unknown tokens, such as in Table 7.

Source:	People have been led to do terrible things in the name of religion.
Response:	Many religions have the option of <unk>

Table 7: A generated response that suffered from a lack of coherence due to unknown tokens.

Our generated debates are extremely varied, ranging from short simple quips to complicated multi-sentence responses. This results in a high standard deviation of perplexities for each debate we generate. Whether or not the debate was complete (addressing important points) was not a criterion that we assessed.

4 Conclusion and Future Work

This work builds a dataset of varied arguments and responses from a tree-structured debate corpus by leveraging information about the nature of the arguments. This approach resulted in a large dataset of varied debate structures.

We present and evaluate a novel structured debate generation approach which brings us closer to realistic simulation of human debate. Our contribution is a multi-turn debate structure following a high level stance-based grammar.

Our framework is language model agnostic; in the future we would like to improve our work using large embeddings to experiment with this and other text generation models.

References

- Amazon. 2019. Mechanical Turk. <https://www.mturk.com/>.
- Eric Bolton, Alex Calderwood, and Niles Christensen. 2019. High Quality Real-Time Structured Debate Generation code. <https://bit.ly/2UjqOMI>.
- Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. *Annual Meeting of the Association for Computational Linguistics*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Mehdi Drissi, Olivia Watkins, and Jugal Kalita. 2018. Hierarchical text generation using an outline. *International Conference on Natural Language Processing*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *Annual Meeting of the Association for Computational Linguistics*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2019. *Strategies for structuring story generation*.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1243–1252.
- Matthew Honnibal and Ines Montani. 2017. Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.
- IBM. 2019. Project Debater. <https://www.research.ibm.com/artificial-intelligence/project-debater/>.
- iDebate. 2019. <https://www.idebate.org/>.
- Xu Jingjing, Zhang Yi, Zeng Qi, Ren Xuancheng, Cai Xiaoyan, and Sun Xu. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Conference on Empirical Methods in Natural Language Processing*.
- Kialo. 2019. Kialo: A platform for rational debate. <https://www.kialo.com/tour>.
- Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a search strategy in neural dialogue modelling. *arXiv preprint arXiv:1811.00907*.
- Langdetect. 2014. Language detection library ported from Google’s language-detection. <https://pypi.org/project/langdetect/>.
- Jeff Malpas and Donald Davidson. 2012a. The stanford encyclopedia of philosophy: Informal logic. <https://plato.stanford.edu/entries/logic-informal/>.
- Jeff Malpas and Donald Davidson. 2012b. The Stanford Encyclopedia of Philosophy: Negation. <https://plato.stanford.edu/entries/negation/>.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1.
- ProCon. 2019. <https://www.procon.org/>.
- Christopher Purdy, Xinyu Wang, Larry He, and Mark Riedl. 2018. Predicting generated story quality with quantitative measures. In *Artificial Intelligence and Interactive Digital Entertainment Conference*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *Conference on Empirical Methods in Natural Language Processing*.
- National Speech and Debate Competition. 2018. World Schools Debate Ballot. <https://www.speechanddebate.org/wp-content/uploads/World-Schools-Ballot-BW.pdf>.
- Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2018. Cold fusion: Training seq2seq models together with language models. *Interspeech*.
- Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. 2018. Plan-and-write: Towards better automatic storytelling. *arXiv preprint arXiv:1811.05701*.
- Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. *International Conference on Learning Representations*.