A Novel Corpus of Discourse Structure in Humans and Computers

Babak Hemmatian[1,2], Sheridan Feucht[1], Rachel Avram[1], Alexander Wey[1], Muskaan Garg[1], Kate Spitalnic[3]
Carsten Eickhoff[1], Ellie Pavlick[1], Bjorn Sandstede[1], Steven Sloman[1]
[1]Brown University, [2]University of Illinois at Urbana-Champaign, [3]University of Sussex
babak.hemmatian@gmail.com

Abstract

We present a novel corpus of 445 human- and computer-generated documents, comprising about 27,000 clauses, annotated for semantic clause types and coherence relations that allow for nuanced comparison of artificial and natural discourse modes. The corpus covers both formal and informal discourse, and contains documents generated using fine-tuned GPT-2 (Zellers et al., 2019) and GPT-3 (Brown et al., 2020). We showcase the usefulness of this corpus for detailed discourse analysis of text generation by providing preliminary evidence that less numerous, shorter and more often incoherent clause relations are associated with lower perceived quality of computer-generated narratives and arguments.

1 Introduction

Recent years have seen the massive growth and popularity of text-generating algorithms, from GPT-2 (Radford et al., 2019) with its 1.5B parameters to GPT-3 with its 175B (Brown et al., 2020). However, it is less clear which aspects of human discourse these models can or cannot capture. We present a novel corpus of human- and computer-generated text with detailed annotations of discourse elements and coherence to allow for more nuanced comparison of the two types of text.

2 Corpus Composition

Our corpus focuses on marijuana legalization discourse throughout 2008-2019, spanning a period of time throughout which general attitudes towards cannabis shifted, allowing our corpus to capture temporal changes in discourse style. The topical focus also allowed us to reduce sources of noise and focus instead on the discourse properties that distinguish human- and computer-produced content. The corpus contains 409 unique full-length documents, 445 unique annotations of them adding up to 26,986 clause labels, covering both formal news discourse and informal social media discussions. News articles from across the political spectrum (Washington Post and Huffington Post as Liberal and Fox News and Breitbart as Conservative sources) were sampled across the covered years from Common Crawl. Potentially relevant content was identified using original comprehensive regular expressions and then manually examined. Reddit discussions were carefully chosen from debate forums and other communities on Reddit (sampled from this comprehensive dataset).

This set of human-written documents is complemented by an almost equal number of corresponding computer-generated articles. 162 documents were generated using Grover (Zellers et al., 2019), a fine-tuned version of GPT-2 (Radford et al., 2019) for news article generation using prompt text and meta-data. Another 60 annotated documents were produced using the most powerful GPT-3 engine once it was made available during the course of corpus development (Brown et al., 2020). Each computer-generated document was produced using one of the human-written articles as a prompt, creating pairs of documents with similar content and style. Much trial and error was involved in prompt design to achieve the most coherent results. For GPT-3, various prompt lengths, top p, stop, and temperature parameter values were tested. For Grover, default parameters from the original work were eventually used (Zellers et al., 2019), but we identified influences on the quality of generations that have not been recognized by the original authors. For instance, inclusion of article titles in the prompt (human or automatically-generated) was necessary for output coherence.

3 Annotation Procedures

Documents were annotated in terms of quality according to definitions of narrative and argument adapted from Smith (2003). Trained assistants (blind to whether documents were human-written or algorithmically-generated) rated each document for narrative and argument presence/absence. The ratings also included narrative quality across four dimensions (plausibility, completeness, consistency and coverage; Yale, 2013) and argument quality across two dimensions (cogency and effectiveness; Wachsmuth et al., 2017), as well as expressed attitude, partisanship, and other document-level information. Wachsmuth et al.'s third measure of argument quality was excluded after beginning annotation with all three measures, due to high correlation with cogency and effectiveness.

Only 57% of documents generated by Grover contained the same amount of narrative-like discourse as their corresponding human-written prompts, whereas 68% of Grover-generated documents maintained the presence or absence of argumentation in the prompt. The algorithm also lagged far behind humans when averaged across all quality measures (mean human quality = 4.5 (narrative), 4.41 (argument) on a 5-point scale; mean Grover quality = 2.275 (narrative), 1.915 (argument)). GPT-3 approached humans in performance but was still significantly worse across all measures of document quality ($p < 0.01$ across dimensions; mean quality = 4.01 (narrative), 3.54 (argument)).

To pinpoint what might be causing the disparity in document quality between humans and computers, the corpus was manually annotated by trained assistants for: 1) Structural linguistic elements of the two discourse modes based on the framework proposed by Smith (2003) and developed for corpora by Friedrich (2017). Examples of clause types under this framework include basic states, bounded events and generic sentences. 2) A comprehensive set of coherence relations based on Wolf and Gibson (2005). Examples of relations between clauses in this framework include cause and effect, temporal sequence and contrast. Krippendorf's alpha for interrater agreement ranged [.45,.52].

We extended these previous frameworks to better distinguish the compositional linguistic properties making up each clause label (e.g., based on Govindarajan et al., 2019), and to account for incoherent content (e.g. repetition or intuitively meaningless relations) that may explain the difference in quality between computer and human discourse. More details about the annotation procedure and its evaluation can be found in Chapter 4 of Hemmatian, 2021. The annotated corpus along with metadata and links to the code used in analyses can be found here[1].

4 Preliminary Results

Ongoing analysis of the annotations reveals certain discourse aspects that the algorithms captured well, as well as other elements that differ wildly from their human counterparts. Topic distributions (based on LDA; Blei et al., 2003) were largely similar between paired human and computer articles, suggesting that the algorithms can capture word co-occurrence patterns. The clause type compositions of narrative and argument discourse modes in artificial text also closely match those of human documents (Smith, 2003; Friedrich, 2017).

More differences were found between humans and computers in coherence relations. For certain types like cause-effect (exemplified by "because"), contrast (exemplified by "but") and violated expectation (exemplified by "although"), more than a third of all Grover relations were incoherent. A correlation analysis between doc-level and clause-to-clause annotations showed that relations more commonly found in arguments showed a higher rate of incoherence than relations that were more frequent in narratives. The quality discrepancy was despite the fact that algorithms produced significantly fewer relations, particularly for the same categories, giving the model fewer chances to generated incoherent content. For instance, mean Grover document frequency of cause-effect relations was 1.71 (95 CI: [1.52,1.9]) compared with the human average of 3.56 (95 CI: [3.2,3.92]). For temporal sequence, a relation more commonly found in narratives, these means were much closer (Grover: 1.48 (95 CI: [1.35,1.61]); Human: 1.9 (95 CI: [1.71,2.08]). Computer-generated relations were also shorter in span across the board ($p < 0.01$), suggesting a less robust high-level document structure. GPT-3 showed a significant improvement over Grover, generating fewer incoherent relations, but the overall described patterns among coherence relation categories remained the same. For instance, cause-effect relations were less frequent than in human text (mean frequency of 1.5 compared with the human average of 3.56; $p < 0.001$), and up to half of instances for certain relation types like violated expectation (exemplified in the use of "although" or "but") were rated as incoherent.

These mechanisms may explain the quality discrepancy between humans and text-generation al-

[1] https://github.com/sfeucht/annotation_evaluation

gorithms. However, more broadly speaking, they could reflect either training regimens employed for model development or the greater abstraction of certain (often more argument-related) relations that makes them inherently more difficult to capture. The results may also reflect how models trained on text completion tasks are not incentivized to learn what might be more often left implicit in human text, such as commonsense cause-effect relations (Becker et al., 2017). More detailed analyses of discourse features that distinguish human- from computer-generated content, and subsequent disambiguation of these possibilities, requires further study. Therefore, we invite the computational discourse analysis community to aid us in further investigations of this novel corpus.

Acknowledgements

This work is supported by a seed research grant from the Office of Vice President for Research at Brown University (GR300190). The current extended abstract was presented at the 2nd Workshop on Computational Approaches to Discourse, as part of the Empirical Methods in Natural Language Processing 2021 conference. We thank anonymous reviewers for their comments on its earlier version. We also thank Naomi Shammash for her work as an annotator and early research assistant for this project.

References

Maria Becker, Michael Staniek, Vivi Nastase, and Anette Frank. 2017. Enriching argumentative texts with implicit knowledge. In *Natural Language Processing and Information Systems*, pages 84–96, Cham. Springer International Publishing.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. volume 3, pages 601–608.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Annemarie Friedrich. 2017. *States, events, and generics: computational modeling of situation entity types*. Ph.D. thesis, University of Saarland.

Venkata Govindarajan, Benjamin Van Durme, and Aaron Steven White. 2019. Decomposing generalization: Models of generic, habitual, and episodic statements. *Transactions of the Association for Computational Linguistics*, 7:501–517.

Babak Hemmatian. 2021. *Taking the high road: A big data investigation of natural discourse in the emerging U.S. consensus about marijuana legalization*. Ph.D. thesis, Brown University.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Carlota S. Smith. 2003. *Modes of Discourse: The Local Structure of Texts*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, United Kingdom.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 32:249–287.

Robert N Yale. 2013. Measuring narrative believability: Development and validation of the narrative believability scale (nbs-12). *Journal of Communication*, 63(3):578–599.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news.