Improving Text Generation via Neural Discourse Planning

Alexander Chernyavskiy
alschernyavskiy@gmail.com
National Research University Higher School of Economics
Moscow, Russia

ABSTRACT

Recent Transformer-based approaches to NLG like GPT-2 can generate syntactically coherent original texts. However, these generated texts have serious flaws. One of them is a global discourse incoherence. We present an approach to estimate the quality of discourse structure. Empirical results confirm that the discourse structure of currently generated texts is inaccurate. We propose the research directions to plan it and fill in the text in its leaves using the pipeline consisting of two GPT-based generation models. The suggested approach is universal and can be applied to different languages.

ACM Reference Format:
Alexander Chernyavskiy. 2022. Improving Text Generation via Neural Discourse Planning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3488560.3502214

1 INTRODUCTION

Natural Language Generation (NLG) task is one of the most challenging and important tasks in NLP. There are various types of NLG tasks: text summarization, machine translation, knowledge aggregation, etc. We consider tasks where the main goal is to construct a text that cannot be distinguished from a human-written text, by a human or a recognition system.

The most successful and universal models for solving NLP tasks are models based on the idea of transformers. Hence GPT [11] and its larger modifications, e.g. GPT-2 [12], successfully perform text generation tasks. However, they still have drawbacks. First of all, fragments in some generated texts do not cohere well with each other, despite the correct syntactic structure. Ko and Li [7] demonstrated that even the words that indicate discourse relations (such as "but" and "because") can be generated improperly, and proposed an auxiliary model to correct them. More problems arise at a higher level, associated with the consistency between sentences. In some cases, the model generates a completely incorrect discourse structure triggered by an inability to plan it. Thus, even the order of the discourse relations should be corrected.

We conducted experiments for GPT-2 and distinguished two types of its mistakes. Firstly, it does not generate well an overall discourse structure (RST-based, [8]). Accordingly, contradictions can be found in it. We fine-tuned GPT-2 on lower-cased movie reviews. Here are examples of mistakes in the generated texts.

WSDM '22, February 21–25, 2022, Tempe, AZ, USA

ACM ISBN 978-1-4503-9132-0/22/02.
https://doi.org/10.1145/3488560.3502214

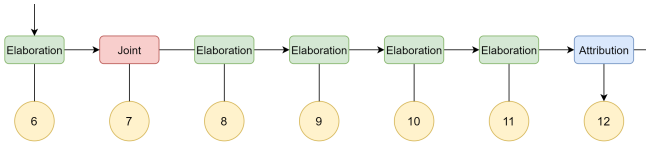

Figure 1: A part of the discourse tree for the generated text: *"... [named john]*[6] *[who survives a major accident]*[7] *[and is saved by a state of the art experimental operation]*[8] *[that turns him into a robotic machine-like agent]*[9] *[who has tools and contraptions of all sorts]*[10] *[built into his body at his use]*[11] *[when he says]*[12] *...". Arrows are drown from Nucleus to Satellites.

Let us consider the example demonstrated in Figure 1. The sentence has too many "Elaboration" and "Joint" rhetorical relations, which are default ones. Moreover, thought structure is not reflected in this discourse tree as it looks like a chain. Generally, genuine discourse trees are more balanced.

Apart from that, the "final" summary is in the middle of the text in some cases. It is not followed by the "end of sequence" token and continues by Elaboration. As a result, the text is duplicated and contradictions may arise.

In this paper, we present an automatic approach to estimate the quality of discourse structure and experimentally confirm that the discourse structure can be generated improperly in some cases.

Our main goal is to develop the model that can generate EDUs connected by discourse relations in the correct order and use the correct words to express it. To this end, we propose a pipeline consisting of two GPT-2-based generation models.

2 RELATED WORK

Puduppully et al. [10] considered the task of generating summaries for games. The output texts are long but obey a certain structure. The authors' model learns content plans from training data.

Most data-to-text datasets do not naturally contain content plans. These plans can be derived following an information extraction approach, by mapping the text in the summaries onto entities in the structured data, their values, relations and types. Similarly, Ciampaglia et al. [4] showed that we can leverage any collection of factual human knowledge for automatic fact checking.

At the same time, some neural methods can be used to plan content and structure without any knowledge bases. For instance, Peng et al. [9] proposed a method to generate text endings based on a pre-planned intent which is predicted due to an additional neural model.

Also, some researchers suggested planning the entire discourse structure or its approximation. Biran and McKeown [1] proposed neural text generation based on the selected discourse relations which can be chosen using n-grams. Ji et al. [6] suggested a similar

approach but predicted discourse relations using RNN. Harrison et al. [5] investigated an approach that allows generating text depending on the need of the "Contrast" relation. One of the main goals was that the model itself should be able to determine which items are suitable for contradistinction and which values are acceptable for them. Text generation benefits from this planning, but the approaches do not use modern methods, and the discourse is only partially planned.

Bosselut et al. [2] suggested an RL-based approach with rewards associated with the correctness of the discourse structure. However, due to the complexity of discourse evaluation, the authors trained the model only to generate the correct order of sentences.

Post-processing can also be used to correct discourse by analogy with correcting entity values. Ko and Li [7] considered the word-level discourse correction for GPT-2. The proposed approach predicts the masked discourse connective given the rest of the sentence. Thus, it improves consistency within sentences. The quality was verified due to the human-annotated relations. Firstly, it should be highlighted that this approach does not consider long discourse dependencies. Moreover, human annotations may be costly.

Our ideas allow to partially solve the problems mentioned above.

3 DISCOURSE EVALUATION

We suggest the discourse structure evaluation using a recursive neural network [3] denoted as RSTRecNN. This model was initially suggested for discourse-based text classification.

To evaluate discourse, RSTRecNN can be trained with an objective to distinguish real texts and texts generated by a generation model based on the prompts from the real texts. The classifier will pay more attention to the order of EDUs and the discourse relations between them than to the meanings of the words since the semantic embeddings will be close.

We conducted an experiment with lower-cased IMDB movie reviews from a Kaggle competition [1]. The base GPT-2 model was fine-tuned on 42,000 texts, and 1250 texts were used as real texts with generation prompts. The RSTRecNN model achieved the accuracy of 0.82 for this balanced dataset. Thus, the discourse structure for real and fake texts differs considerably.

This approach can also be used to compare generation models. A better model should generate texts with the lower accuracy of RSTRecNN since it should be more difficult for the classifier to distinguish generated texts using the discourse structure.

4 DISCOURSE PLANNING

Our major idea is to plan further discourse substructure based on the current context and generate the text in its leaves. To this end, we propose the generation pipeline: discourse structure planner and relation's text generator.

The first generator is proposed to plan further discourse substructure based on the current context. Its main advantage is planning the relations between future leaves. It takes the context and a previously generated structure as its input and generates a sequence of tokens that uniquely define the following raw discourse subtree (without the text in leaves). We consider *treeASstring* converting for the discourse subtrees of pre-defined depth and fine-tune GPT-2 on texts of the form "context ⟨SEP⟩ structure". We propose the "Global" relation to connect nodes that are connected at a higher level in a complete discourse tree, and "Empty" relation that adds a dummy node with empty text.

The second generator fills in the text in leaves of the discourse subtree produced by the structure generator. It supplements the structure step by step in a traversal (DFS) order from Nucleus to Satellites. At each step, the model objective is to generate a text span associated with some leaf. Conditioning the input of the generator to the current generated leaves (text spans) occurs by adding them to the current structure. At the same time, the generator uses ⟨MASK⟩ tokens to hide spans that are still not generated in the current structure.

Our approach is universal and is not just English language-specific, and we plan to apply it to other languages. Apart from that, we will experiment with dialogue discourse structure to improve the discourse coherence of the generated conversations.

[1] https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

ACKNOWLEDGMENTS

The publication was supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139. It was also supported in part through the computational resources of HPC facilities at NRU HSE.

REFERENCES

[1] Or Biran and Kathleen McKeown. 2015. Discourse Planning with an N-gram Model of Relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1973–1977. https://doi.org/10.18653/v1/D15-1230

[2] Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-Aware Neural Rewards for Coherent Text Generation. *CoRR* abs/1805.03766 (2018). arXiv:1805.03766 https://arxiv.org/abs/1805.03766

[3] Alexander Chernyavskiy and Dmitry Ilvovsky. 2020. *Recursive Neural Text Classification Using Discourse Tree Structure for Argumentation Mining and Sentiment Analysis Tasks*. 90–101. https://doi.org/10.1007/978-3-030-59491-6_9

[4] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis Mateus Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational Fact Checking from Knowledge Networks. *PLoS ONE* 10 (2015).

[5] Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn A. Walker. 2019. Maximizing Stylistic Control and Semantic Accuracy in NLG: Personality Variation and Discourse Contrast. *CoRR* abs/1907.09527 (2019). arXiv:1907.09527 http://arxiv.org/abs/1907.09527

[6] Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. *CoRR* abs/1603.01913 (2016). arXiv:1603.01913 http://arxiv.org/abs/1603.01913

[7] Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing Discourse Relations in Language Generation from GPT-2. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, Dublin, Ireland, 52–59. https://aclanthology.org/2020.inlg-1.8

[8] William Mann and Sandra Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organization. (01 1987).

[9] Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards Controllable Story Generation. In *Proceedings of the First Workshop on Storytelling*. Association for Computational Linguistics, New Orleans, Louisiana, 43–49. https://doi.org/10.18653/v1/W18-1505

[10] Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-Text Generation with Content Selection and Planning. In *AAAI*.

[11] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035

[12] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.