Correcting Texts Generated by Transformers using Discourse Features and Web Mining

Alexander Chernyavskiy, Dmitry Ilvovsky
HSE University
Moscow, Russia
alschernyavskiy@gmail.com
dilvovsky@hse.ru

Boris Galitsky
Oracle Inc.
Redwood Shores CA USA
bgalitsky@hotmail.com

Abstract

Recent transformer-based approaches to NLG like GPT-2 can generate syntactically coherent original texts. However, these generated texts have serious flaws: global discourse incoherence and meaninglessness of sentences in terms of entity values. We address both of these flaws: they are independent but can be combined to generate original texts that will be both consistent and truthful. This paper presents 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 correct it using discourse features during the fine-tuning procedure. The suggested approach is universal and can be applied to different languages. Apart from that, we suggest a method to correct wrong entity values based on Web Mining and text alignment.

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 and multimedia information construction such as music generation. 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 (Radford et al., 2018) and its larger modifications GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) 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 (2020) demonstrated that even the words that indicate discourse relations (such as "but", "after" 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. The model sometimes generates a completely incorrect discourse structure triggered by an inability to plan it. Even the order of the discourse relations should be corrected.

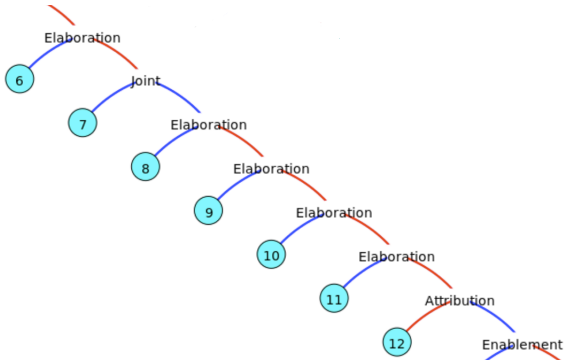

Figure 1: A part of the discourse tree for the generated text: *"... [named john]^{6} [who survives a major accident]^{7} [and is saved by a state of the art experimental operation]^{8} [that turns him into a robotic machine-like agent]^{9} [who has tools and contraptions of all sorts]^{10} [built into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

We conducted experiments for GPT-2 and distinguished two types of its mistakes. Firstly, it does not generate well an overall discourse structure (RST is described in Sect. 3). Accordingly, in some cases, contradictions can be found in it. We fine-tuned GPT-2 on lower-cased movie reviews. Here are examples of typical mistakes in the generated texts.

Let us consider the example demonstrated in Figure 1. The sentence has too many "Elaboration" and "Joint" rhetorical relations, which are default

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 36–43,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_006

ones. Moreover, thought structure is not reflected in this discourse tree as it looks like a chain. Generally, genuine discourse trees are more balanced.

Another illustrative example of generated text:

> *with enough fine performances from all of them , this is one of the best cult films ever made.* ⟨. . . ⟩ *it does have some nice gore and some pretty well chosen actors but it is definitely not one of the best cult films of all time .*

In this case, the core text idea is contained in the last sentence. The first part is used to elaborate it. However, there is a contradiction, and a contrast relation does not contribute to coherence.

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.

In addition, for specific areas like medical texts, GPT-2 uses incorrect values of entities since it does not utilize any knowledge base.

Our example is from the 19th century literature domain. The seed to GPT content generation is in bold.

> ***Pushkin wrote the originial chapters of Yevgenii Onegin****, then Alexander Pushkin sent letters to Nikolay Gogol, Mikhil Lermontov, and others. While in Saint Petersburgh, Pushkin was approached by Rodion Romanovich Krafft. Krafft wanted Pushkin to be an English translator of a German edition of his poem*

Here, 'then' part of the first sentence seems plausible. Then the reader proceeds to the invented person Krafft, and a random, implausible text starts. To continue impress the reader with a smooth flow of text, we need to find a real person like Krafft related to Pushkin via translation.

The bottom sentences need to be replaced with following:

> *In the 1960's, Vladimir Nabokov, the only writer to simultaneously hold the position as a giant of both Russian and English literature, published an English translation of Pushkin's masterpiece.*

Hence most entities from GPT text generation need to be replaced.

This paper presents empirical proof that GPT-2 generates wrong discourse structure in some cases and proposes ideas for the development of text generation in two directions:

1. Correction of the general consistency of the text, namely its discourse structure. We plan to investigate methods that allow the model to generate Elementary Discourse Units connected by discourse relations in the correct order and use the correct words to express it. Besides, we propose a method to estimate the overall quality of the discourse structure.

2. Correction of generated entity values using external knowledge bases, Web Mining and alignment.

These directions complement each other and can be used sequentially. So, firstly, we correct the discourse structure of the text using additional fine-tuning. Then, we utilize external knowledge to correct the exact specific entity values and get meaningful text.

Our paper is organized as follows. Firstly, we summarize related work and introduce base concepts. Then, we discuss our ideas, preliminary results, and propose directions for further research.

2 Related Work

2.1 General View

In this paper, we consider approaches that can generate unique coherent texts. One of the major problems in open-domain content generated by a deep learning based system is its meaninglessness. Although overall syntactic and logical structure, obtained via averaging of texts from the training dataset, looks plausible in most cases, and some individual phrases might make sense, almost each sentence is meaningless. The main advantage of such raw content is that it is **original.**

Recent approaches to NLG based on external knowledge bases provide good results (Liu et al., 2017; Freitag and Roy, 2018). At the same time, most research considers the only superficial description of a simple piece of structured data such as attribute-value pairs of fixed or very limited schema, like E2E (Novikova et al., 2017) and WikiBio (Lebret et al., 2016). For real-world complex databases, it is often more desirable to provide descriptions involving abstraction and a logical inference of higher generality about database tables and

records. For instance, readers should get a kind of a summary over a structured, relational or no-SQL database.

2.2 Knowledge-based Planning

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.

Ciampaglia et al. (2015) showed that we can leverage any collection of factual human knowledge, such as Wikipedia, for automatic fact checking. The authors follow the paradigm of epistemic closure, computationally building a support for statements by mining the connectivity patterns on a knowledge graph. The initial task is to compute the support of simple statements of fact using a large-scale knowledge graph obtained from Wikipedia. Generally speaking, fact checking can be seen as a special case of link prediction in knowledge graphs (Nickel et al., 2016).

2.3 Style and Discourse Correction

Another way is integration of plans into the training process of the generation system without any knowledge bases. This suffices to solve the second aforementioned problem associated with planning the global structure. Some researchers suggest ways to generate consistent long texts.

Peng et al. (2018) proposed a method to generate text endings based on a pre-planned intent (e.g. "happyEnding", "sadEnding", or "cannotTell") which is predicted due to an additional neural model.

Biran and McKeown (2015) proposed neural text generation based on the selected discourse relations which can be chosen using n-grams. Ji et al. (2016) suggested a similar approach but predicted discourse relations using RNN. Harrison et al. (2019) 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. The idea of using the intent is very important and can be used for discourse planning too, and we propose ways to integrate it in GPT fine-tuning.

Bosselut et al. (2018) suggested an RL-based approach with rewards associated with the correctness of the discourse structure. However, due to the complexity of assessing the correctness of the discourse, the authors trained the model only to generate the correct order of sentences. The sentence coherence was considered as an approximation of the discourse structure.

Post-processing can also be used to correct discourse by analogy with correcting entity values. Ko and Li (2020) 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. It should be highlighted that this approach does not consider long relations. Moreover, human annotations may be costly. Our ideas allow to partially solve it.

At the same time, the consistency of generated texts still remains not at a high level, and there are quite a few articles devoted to its investigation. In addition, there are even fewer papers devoted to the correction of the discourse itself.

3 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) proposed by Mann and Thompson (1987) allows to represent any text as a binary tree. Its nodes correspond to text spans and edges conform to discourse relations between them.

The tree is constructed step by step. At the first step, the leaves of the tree are Elementary Discourse Units (shortly EDUs). They contain introductory words, single thoughts, or clauses. Further, some of these nodes are connected via corresponding discourse relations like "Elaboration", "Joint" and "Summary", and form new nodes associated with bigger text spans. Then, the updated set of nodes is connected due to the corresponding discourse relations and so forth. In the end, the text span associated with the root node is the full entire text.

RST distinguishes two types of nodes: Nucleus and Satellite. Nodes of the first type contain key information necessary to understand meaning of the text. Nodes of the second type comprise supplementary information.

Figure 2 demonstrates an example of a discourse tree constructed by the open-source ALT document-level discourse parser (Joty et al., 2012) fot the following text:

"[Media accounts have portrayed past moves as cuts, as well.]^1 [Whats more,]^2 [when Walker intro-

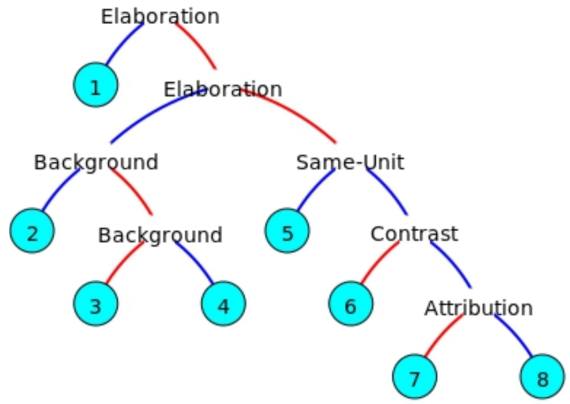

Figure 2: Discourse tree constructed by the ALT parser. Dark blue edges are drawn to Nucleus nodes and light red arrows are drawn to Satellite nodes.

duced his budget,][3] [he also described the changes as cuts.][4] [So,][5] [Walker made a generally accurate statement about the increase,][6] [but by dismissing the talk of cuts][7] [he left out a lot of context and important details on his budget move.][8]"

Here, EDUs are limited by brackets.

4 Discourse Correction

4.1 Methods

Our first idea is to utilize external discourse markers during GPT-2 fine-tuning. The model can be trained to generate them. In a sense, the generated marker will assist to plan next words during generation. In this case, it is necessary to ensure that the model learns planning, not labeling. In details, our suggestions can be described as follows.

Let the GPT-2 model be fine-tuned on a large dataset from the selected subject area. It can be fine-tuned once again using an additional smaller set $\{X_{\text{real\_ft}}\}$ of texts from the same area. So, the set of generated texts $\{X_{\text{fake}}\}$ will be constructed. These are texts that the base model generates with the unmodified training process.

At the same time, we can modify real texts using their discourse structure to add more useful information to the training process. For example, special tokens related to discourse relations can be added. One of the most obvious options is to utilize $\langle$R$\rangle$ tokens (e.g. $\langle$Contrast$\rangle$) before the corresponding connectives. The updated texts denoted as $\{X_{\text{real\_ft\_disco}}\}$ are used to fine-tune the model instead of $\{X_{\text{real\_ft}}\}$. Thus, another set of fake texts will be generated. The aforementioned special tokens should be removed from them to get raw texts $\{X_{\text{fake\_disco}}\}$. Also, the model tokenizer must treat any $\langle$R$\rangle$ token as one subtoken.

We construct fake texts using another correct set $\{X_{\text{real}}\}$. To make the texts more similar in terms of semantic embeddings, for each text from $\{X_{\text{real}}\}$ we generate the text in $\{X_{\text{fake}}\}$ and the text in $\{X_{\text{fake\_disco}}\}$ that have the same words in the beginning.

We propose a criterion to check the improvement of the discourse structure using a recursive neural network (Chernyavskiy and Ilvovsky, 2020) denoted as RSTRecNN. This model was suggested for discourse-based text classification.

Let two discriminative RSTRecNN models $\mathcal{M}$ and $\mathcal{M}_{\text{disco}}$ are trained to solve binary classification tasks $\{X_{\text{real}}\}$ vs $\{X_{\text{fake}}\}$ and $\{X_{\text{real}}\}$ vs $\{X_{\text{fake\_disco}}\}$ respectively. The classifier will pay more attention to the order of EDUs and to the discourse relations between them than to the words meanings since the semantic embeddings are close. Therefore, if the quality of $\mathcal{M}_{\text{disco}}$ is lower than that of $\mathcal{M}$, then it has become more difficult for the classifier to distinguish fake texts using its discourse structure. Thus, in this case the goal of discourse correction will be achieved.

In addition, $\mathcal{M}$ itself suffices to test the hypothesis that the discourse needs to be corrected for the base GPT-2 model. In this aspect, if the accuracy for $\mathcal{M}$ is close to 0.5, then the generated discourse structure is already good (since the dataset is balanced).

4.2 Experimental Details

We conducted experiments for IMDB movie reviews. We added 50,000 examples from a Kaggle competition [1] to the base dataset with 2000 texts from (Pang and Lee, 2004). As the preprocessing, we lowercased all texts.

The base GPT-2 model was fine-tuned on 32,400 texts, and we utilized 3,600 texts as the validation set. GPT-2 was fine-tuned for 3 epochs with a learning rate 5e-5. The final validation perplexity is 28.17. We used 10,000 texts as $\{X_{\text{real\_ft}}\}$ for the second fine-tuning. The ALT parser was applied for discourse labeling.

We did not utilize "Elaboration", "Joint" and "Same-Unit" relations since they are the most popular and do not make much sense. The discourse parser distinguished 19 unique discourse relations, and the relative frequencies of "Elaboration","Joint'

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

Model	Max Acc.	Mean $\pm$ Std Acc.
$\mathcal{M}$	0.822	0.807 ± 0.009
$\mathcal{M}_{\text{disco}}$	0.819	0.810 ± 0.006

Table 1: Model performance for two variants of the datasets.

and "Same-Unit" are 0.44, 0.12 and 0.11 correspondingly. Also, we did not use rare relations such as "Manner-Means", "TextualOrganization" and "Topic-Change" (the relative frequency is lower than 0.001). Broadly speaking, we considered only meaningful popular non-trivial relations.

We generated texts with length exceeding 300 since the discourse mistakes for long texts are more obvious. $\{X_{\text{real}}\}$ included 1250 texts. Nucleus sampling technique (Holtzman et al., 2019) was used because its results were superior to that of top-k sampling (Fan et al., 2018).

4.3 Results

Table 1 demonstrates the results for the models $\mathcal{M}$ and $\mathcal{M}_{\text{disco}}$. One can see that the accuracy for $\mathcal{M}$ is much higher than 0.5 Thus, the discourse structure for real and fake texts differs considerably.

The table shows that the maximal accuracy (over 6 runs) for $\mathcal{M}$ is higher than that for $\mathcal{M}_{\text{disco}}$. However, there is no significant difference between the models.

It is important to note that there are much fewer special tokens in raw $\{X_{\text{fake\_disco}}\}$ than in the labeled $\{X_{\text{real}}\}$. Moreover, the biggest part of the generated special tokens are quite obvious and in most cases stand in front of their indicators.

4.4 Future Work

There are several directions for further research:

- Customization of the loss function. For instance, we can modify loss weights for the classes associated with the special tokens.
- Modification of special tokens. It may be important to generate tokens associated with the beginning or end of EDUs. In this way, we will also experiment with $\langle \text{R}_{\text{start}} \rangle$ and $\langle \text{R}_{\text{end}} \rangle$ tokens instead of $\langle \text{R} \rangle$.
- Modification of the sampling process. We can check the global discourse coherence using RSTRecNN after generating new EDU at the stage of generation. In the case of bad structure, we will generate it again.
- Customization of the attention module in GPT-2. We plan to add discourse information in the attention module. Accordingly, new words will be generated using the current discourse structure during the training process.

5 Correction of Entity Values

5.1 Methods

Our second major contribution is to correct entity values using web or external knowledge bases. The entities that need to be corrected comprise titles, names and so forth.

Our intent is to take the meaningless raw generated content and cross-breed it with the one taken piece-by-piece from various sources, so that each sentence is not original but truthful. We borrow the structure and content flow from the generated text, and factoids are taken from true texts mined from the web to correspond to the generated sentences.

To obtain true sentences, we form a query from the generated sentence by retaining noun phrases and other significant phrases, and forming OR query. We then search against the whole web, a given web source such as Wikipedia, an intranet or a specific index containing authoritative documents. Iterating through search results, the true sentences which are the closest to the generated sentences are identified. Aligning the raw generated sentence and the identified true sentence, we observe which entities and values in the generated sentences are incorrect and substitute them.

Firstly, we determine what kind of linguistic data should be taken from the generated text, and which – from the true text. Sources such as syntax and discourse we attempt to use from the raw text (the text generated by base GPT or by discourse-based fine-tuned GPT), and once we determine that it is not possible, we obtain from the true text. Once we perform a substitution of a phrase from true to generated, we know which linguistic sources can be retained in the raw generated text. Table 2 demonstrates the partition of data sources.

We use a non-symmetric operation of alignment between generated and true text (Galitsky, 2020). This alignment occurs at the level of the whole text, paragraphs, sentences and phrases. To assess a similarity between texts, paragraphs, sentences and phrases, we apply a symmetric operation of generalization (Galitsky et al., 2012). However, to obtain a proper text aligning generated and true texts, the operation is not symmetric since we use distinct

Data	Generated sentences	True sentences
Source of the text	text generated by DL	real text obtained from sources like web
Syntactic flow	if possible	if required
Discourse flow	if possible	if required
Coreference structure	if possible	if required
Logical flow	if possible	if required
Idea	original "idea"	existing idea, if the original idea is too distant from the topic
Entities	except entities are most likely wrong and need to be substituted	correct entities
Other	actions can be retained, if confirmed by	phrases

Table 2: Merging linguistic data types from generated and true texts.

sources depending on the source type. An entity from a true sentence kills an entity from a generated sentence, but rest of the phrase is taken from generated sentence to retain the logical structure, discourse and coreference. Exploring the ways to identify a piece of true content to repair a flawed raw content, we observe that it is a complex multi-step process requiring conventional linguistic analysis at multiple layers and also a special substitution technique operating at various levels of abstraction.

5.2 Future Work

Other research directions include neural methods inspired by the fact-checking task. To check a given statement, some methods (Thorne et al., 2018; Nie et al., 2019) extract a corresponding justification or refutation from a big data corpus (e.g. Wikipedia texts). This extracted text contains the correct entity value and we can substitute the hidden value with it. This approach does not use the Web but requires a more complex search phase. For instance, we can use deep neural models like BERT (Devlin et al., 2018) to find relevant articles and paragraphs.

6 Conclusion

In this paper, we investigated the current transformer-based approaches to natural language generation. We proposed two challenging research directions: an improvement of overall discourse structure and a correction of entity values to construct meaningful texts.

These directions are independent but complement each other well. By learning how to solve both problems, we expect to be able to generate coherent long texts with meaningful, plausible thoughts. Our idea is to first generate discursively and syntactically coherent text using the custom fine-tuned GPT-2 model. Further, utilizing the non-learning technique based on web mining and text alignment, we replace values of wrong factoid entities with truthful ones.

We proposed a method to evaluate the quality of overall discourse structure and experimentally confirm that GPT-2 generates texts with a mistaken and inconsistent structure in some cases. We suggested some ideas to integrate additional knowledge about discourse into the GPT-2 fine-tuning and generation processes.

Apart from that, we suggested a way to correct wrong entity values in generated texts using web mining. We propose the partition of information types from two sources (generated texts and texts from external knowledge base) to apply a syntactic and semantic alignment.

We suggested some ways to develop a universal approach that will not be just English language-specific and can be applied to other languages. At the moment, there are open-source discourse parsers for Russian, German, Spanish, and we can utilize them without modifying the approach.

Acknowledgments

The article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'.

References

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*, pages 1973–1977, Lisbon, Portugal. Association for Computational Linguistics.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

Alexander Chernyavskiy and Dmitry Ilvovsky. 2020. *Recursive Neural Text Classification Using Discourse Tree Structure for Argumentation Mining and Sentiment Analysis Tasks*, pages 90–101.

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.

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.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. *CoRR*, abs/1805.04833.

Markus Freitag and Scott Roy. 2018. Unsupervised natural language generation with denoising autoencoders. In *EMNLP*.

Boris Galitsky. 2020. *Employing Abstract Meaning Representation to Lay the Last-Mile Toward Reading Comprehension*, pages 57–86.

Boris A. Galitsky, Josep Lluis de la Rosa, and Gábor Dobrocsi. 2012. Inferring the semantic properties of sentences by mining syntactic parse trees. *Data & Knowledge Engineering*, 81-82:21–45.

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.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *CoRR*, abs/1603.01913.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915. Association for Computational Linguistics.

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*, pages 52–59, Dublin, Ireland. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and M. Auli. 2016. Neural text generation from structured data with application to the biography domain. In *EMNLP*.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2017. Table-to-text generation by structure-aware seq2seq learning. *CoRR*, abs/1711.09724.

William Mann and Sandra Thompson. 1987. Rhetorical structure theory: A theory of text organization.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *AAAI*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. pages 201–206.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *CoRR*, cs.CL/0409058.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.