Correcting Texts Generated by Transformers using Discourse Features and Web Mining

Alexander Chernyavskiy, Dmitry Ilvovsky and Boris Galitsky

National Research University Higher School of Economics
Russia, Moscow

Oracle Inc.
Redwood Shores CA USA

alschernyavskiy@gmail.com *dilvovsky@hse.ru* *bgalitsky@hotmail.com*

September 3, 2021

Overview

1. Introduction
    - Motivation: coherent and truthful text generation
    - Contributions
2. Approaches
    - Discourse Tree Structure
    - Discourse Coherence Estimation
    - Discourse Correction
    - Correction of Entity Values
3. Conclusion

Motivation

Natural Language Generation (NLG) task is one of the most challenging and important tasks in NLP.

Our major goal is to develop an approach that can generate original texts as close as possible to human-written texts.

Current SOTA approaches are Transformer-based systems: GPT, GPT-2, GPT-3…

Serious **drawbacks**:

- global discourse incoherence
- meaninglessness of sentences in terms of truthfulness

Discourse Incoherence

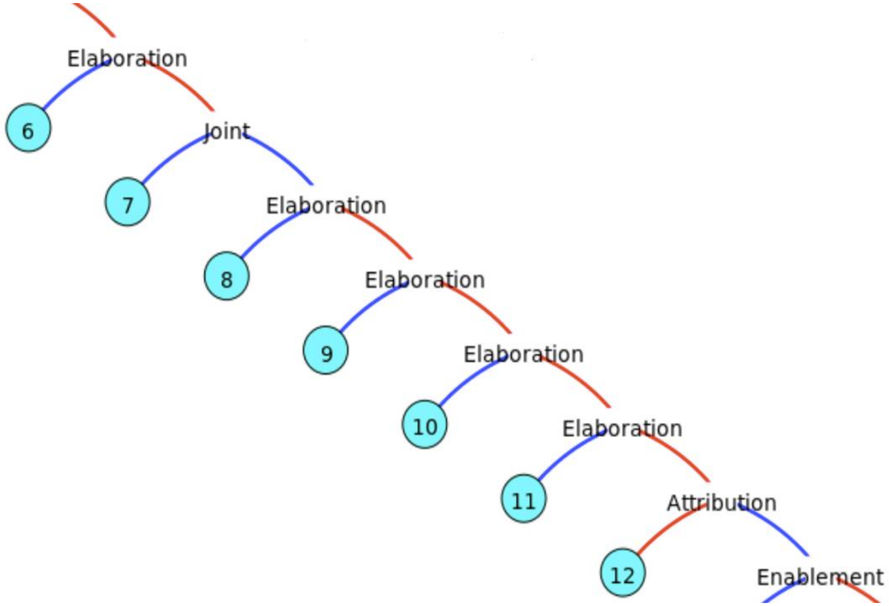
- Even discourse connectivities (such as “but”, “before” and “because”) can be generated improperly. [Ko and Li (2020)]

- More problems can be seen at a higher structural level.

- We conducted experiments for GPT-2 and generated some movie reviews. The discourse structure in the biggest part of them should be corrected.

Discourse Incoherence

- *"... [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} ..."*



The sentence has too many default rhetorical relations (Joint and Elaboration). Also, genuine discourse trees are more balanced.

- *"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"*

The "Summary" relation is generated in the first part and followed by the EOS token. The "Contrast" relation does not contribute to coherence.

Wrong Entity Values

For specific areas, GPT-2 uses incorrect values of entities.

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*

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

We need to find a real person like Krafft related to Pushkin via translation.

Contributions

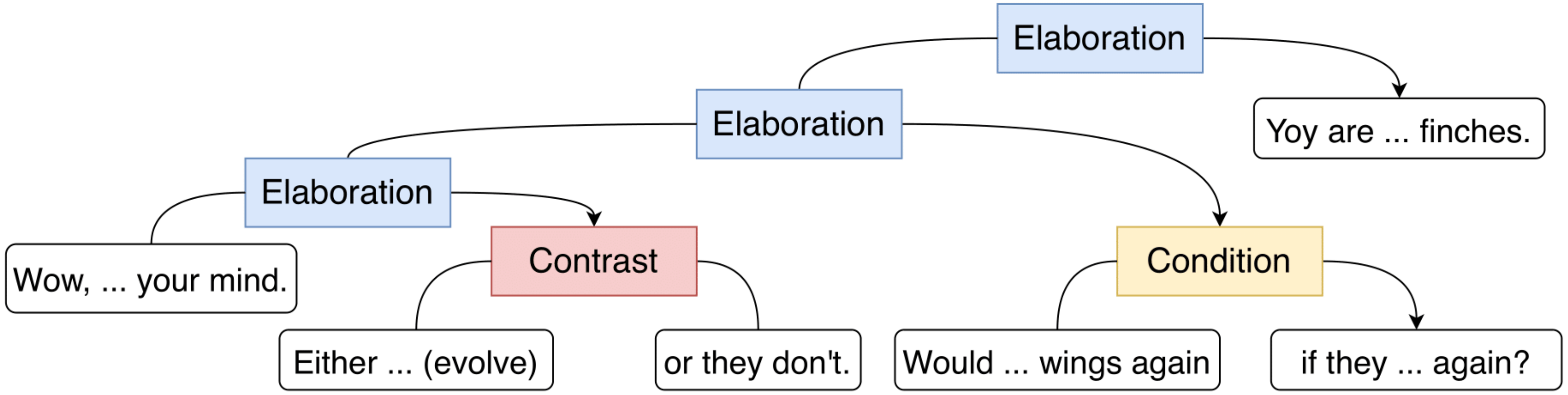
We address both flaws: they are independent but can be combined to generate original texts that will be both consistent and truthful.

1. Correction of the general discourse consistency of the text.
   A. We investigate methods that allow the model to generate text spans connected by discourse relations in the correct order and use the correct words to express it.
   B. 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.

Discourse Tree Structure

- Leaves are elementary discourse units (EDUs)
- Text spans are connected recursively by discourse relations
- There two types of vertices: “Nucleus” (crucial) and “Satellite” (supplementary)
- ALT Document-level Discourse Parser



Text: "Wow, make up your mind. Either populations change over time (evolve) or they don't. Would your wingless beetles be able to produce wings again if they were somehow beneficial again? You are starting to sound like Darwin and his finches."

Discourse Coherence Estimation

We propose an automatic criterion to check the improvement of the discourse structure using a recursive neural network (Chernyavskiy and Ilvovsky, 2020) denoted as RSTRecNN.

Training objective: distinguish human-written texts and texts generated by a model. Prompts from the generation should be taken from original texts to make semantic embeddings closer.

Comparison of generation approaches: lower classifier performance indicates the improvement in terms of the discourse structure (it has become more difficult to distinguish fake texts based on it)

Discourse Correction

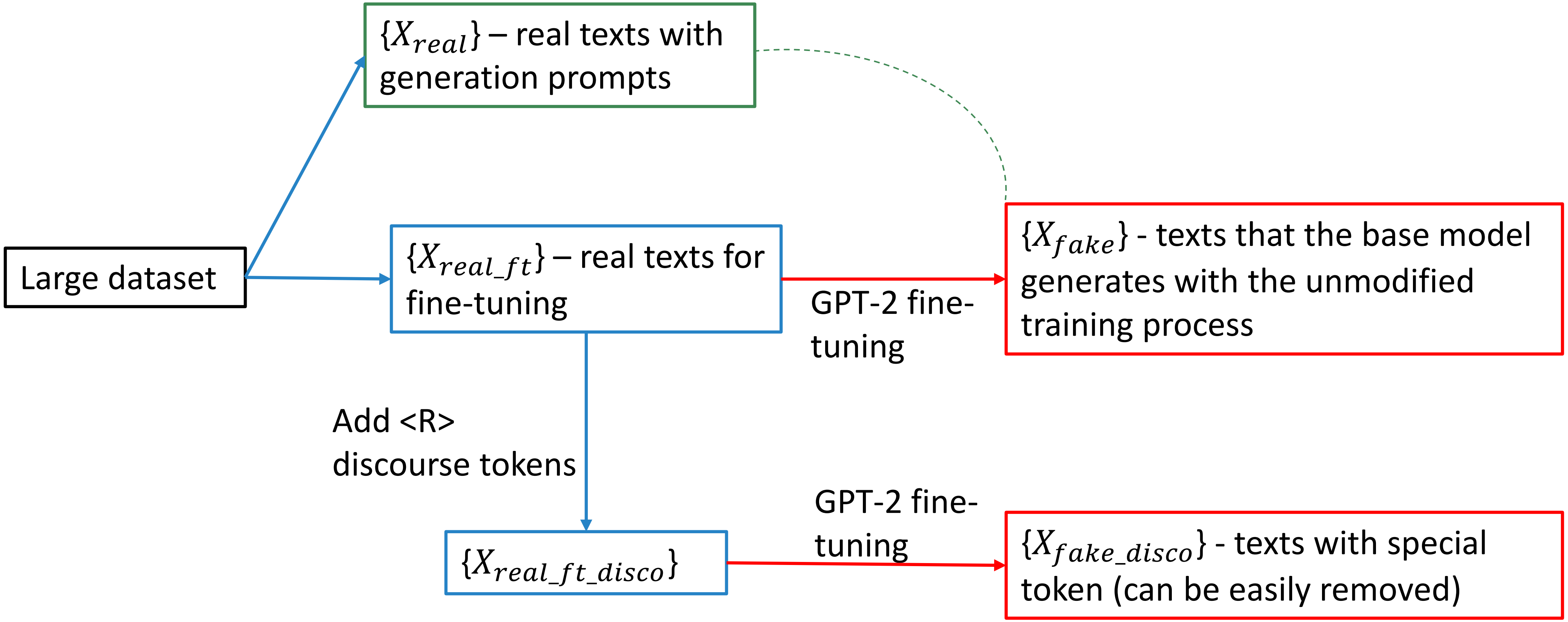
Idea: to plan discourse relations based on the current generated context and generate the following text based on it.

Simplest implementation: to utilize external discourse markers during GPT-2 fine-tuning.

Example:

"… <Background> as far as i know , gene siskel and i are the only ones <Attribution> willing to admit that we dug it . the first time i saw the cable guy , in theatres , i was in super critic - mode , and did nt really like it . <Contrast> however , due to the fact that…"

Discourse Correction



Large dataset

$\{X_{real}\}$ – real texts with generation prompts

$\{X_{real\_ft}\}$ – real texts for fine-tuning

GPT-2 fine-tuning

$\{X_{fake}\}$ - texts that the base model generates with the unmodified training process

Add <R> discourse tokens

$\{X_{real\_ft\_disco}\}$

GPT-2 fine-tuning

$\{X_{fake\_disco}\}$ - texts with special token (can be easily removed)

Discourse Correction

Let’s consider two discriminative RSTRecNN models that solve binary classification:

- $\mathcal{M}$ distinguishes $\{X_{real}\}$ vs $\{X_{fake}\}$
- $\mathcal{M}_{disco}$ distinguishes $\{X_{real}\}$ vs $\{X_{fake\_disco}\}$

Semantic embeddings are close ⇒ classifier will pay more attention to the order of EDUs and to the discourse relations between them.

$\mathcal{M}$ itself suffices to check does the discourse need to be corrected for the base GPT-2. If the accuracy for $\mathcal{M}$ is close to 0.5, then the generated structure is already appropriate.

Experimental Details

- Lowercased IMDB movie reviews
- The base GPT-2 model was initially fine-tuned on 32,400 texts, and on 10,000 texts with discourse markers.
- 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.
- Generation bases on the nucleus sampling technique.

Results

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

- The accuracy for $\mathcal{M}$ is much higher than 0.5. Thus, the discourse structure for real and fake texts differs considerably.
- The maximal accuracy (over 6 runs) for $\mathcal{M}$ is higher than that for $\mathcal{M}_{disco}$. However, this can not be said for the mean accuracy, and now no conclusion can be made about improving the discourse.

Future Work

- Customization of the loss function (modification of loss weights for the special tokens).
- Modification of special tokens. It may be important to generate tokens associated with the beginning or end of EDUs.
- Modification of the sampling process. At the stage of generation, we can check the discourse coherence using RSTRecNN after generating new EDU and regenerate it if needed.
- Splitting generation system into a pipeline: discourse structure planning and generation of relation’s texts.

Correction of Entity Values

Our intent is to take the meaningless raw generated content and cross-breed it with the one taken piece-by-piece from various sources.

- structure and content flow from the generated text
- factoids from true texts mined from the web

Correction of Entity Values

How to obtain ***true sentences***?

noun phrases and other significant phrases from generated context

OR query

search against the whole web, a given web source such as Wikipedia, an intranet or a specific index containing authoritative documents

select sentences which are the closest to the generated sentences

Correction of Entity Values

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 con-firmed by	phrases

Correction of Entity Values

We use an operation of **alignment** between generated and true text.

We coordinate it between two semantic graphs and respective syntactic trees accordingly: semantic and syntactic relations between the source and target graphs are honored.

Raw (~generated) text

She has potassium depletion; spironolactone is a potassium-sparing drug that will cause her to retain potassium

Green – synonymous
Red – substitution

True text

She has potassium depletion due to channelopathy that affects epithelial sodium channels; there is a choice of potassium-sparing drugs and amiloride acts via sodium channels

Future Work

- Conduct experiments for the current approach.

- Modify the search stage. For instance, we can use neural approaches to extract relevant sentences (inspired by fact-checking [Thorne et al., 2018; Nie et al., 2019]).

Conclusion

- We investigated two challenging research directions: an improvement of overall discourse structure and a correction of entity values to construct meaningful texts.
- We proposed a method to automatically evaluate the quality of overall discourse structure and experimentally confirmed that GPT-2 generates texts with a mistaken and inconsistent structure in some cases.
- We suggested some ways to develop a universal approach that can be **applied to other languages**. Now, there are open-source discourse parsers for Russian, German, Spanish, and we can use them without modifying the approach.

References

[1] Wei-Jen Ko and Junyi Jessy Li (2020)
Assessing discourse relations in language generation from GPT-2
Proceedings of the 13th International Conference on Natural Language Generation, pages 52–59, Dublin, Ireland. Association for Computational Linguistics(2015)

[2] 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

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

[4] Or Biran and Kathleen McKeown (2015)
Discourse planning with an n-gram model of relations
Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1973–1977, Lisbon, Portugal. Association for Computational Linguistics

References

[5] Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein (2016)
A latent variable recurrent neural network for discourse relation language models
CoRR, abs/1603.01913

[6] 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

[7] Alexander Chernyavskiy and Dmitry Ilvovsky (2020)
Recursive Neural Text Classification Using Discourse Tree Structure for Argumentation Mining and Sentiment Analysis Tasks
In ISMIS 2020, pages 90–101

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