Assessing Discourse Relations in Language Generation from GPT-2

Wei-Jen Ko
Department of Computer Science
The University of Texas at Austin
`wjko@utexas.edu`

Junyi Jessy Li
Department of Linguistics
The University of Texas at Austin
`jessy@austin.utexas.edu`

Abstract

Recent advances in NLP have been attributed to the emergence of large-scale pre-trained language models. GPT-2 (Radford et al., 2019), in particular, is suited for generation tasks given its left-to-right language modeling objective, yet the linguistic quality of its generated text has largely remain unexplored. Our work takes a step in understanding GPT-2's outputs in terms of discourse coherence. We perform a comprehensive study on the validity of explicit discourse relations in GPT-2's outputs under both organic generation and fine-tuned scenarios. Results show GPT-2 does not always generate text containing valid discourse relations; nevertheless, its text is more aligned with human expectation in the fine-tuned scenario. We propose a decoupled strategy to mitigate these problems and highlight the importance of explicitly modeling discourse information.

1 Introduction

Recent progress in NLP has been marked with the emergence of large-scale pre-trained models, e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and GPT-2 (Radford et al., 2019). Among these, GPT-2 is particularly suitable in natural language generation due to its underlying left-to-right language modeling objective. Indeed, GPT-based language models have shown impressive results for open-domain dialogue generation (Golovanov et al., 2019; Wolf et al., 2019; Zhang et al., 2020). This has motivated investigations into GPT-2's generated text (See et al., 2019; Wallace et al., 2019). In particular, using automatic metrics (e.g., cosine similarity, lexical diversity, sentence length), See et al. (2019) illustrated that GPT-2 has the ability to generate interesting and coherent text. However, analysis of GPT-2's outputs from deeper linguistic dimensions (e.g., discourse) has largely remained unexplored.

In this paper, we perform the first discourse analysis of GPT-2's outputs, under both organic and fine-tuned scenarios, with the goals of understanding model behavior and pointing towards ways of improvement. We chiefly focus on *discourse relations*, one of the most important linguistic devices for textual coherence. Discourse relations specify the relationships between text spans, for example:

> *Jazz is good, **but** my favorite is country music.*

The two clauses (also called *arguments*) are connected by a CONTRAST relation, as signaled by the connective *but*. Discourse relations are central in establishing textual coherence. For example, they create rhetorical connections between spans in the absence of anaphoric entity mentions (Lascarides and Asher, 2008). Cognitive experiments have repeatedly shown discourse relations to be highly influential in the mental processing of text (Meyer and Freedle, 1984; Horowitz, 1987; Millis et al., 1993; Sanders and Noordman, 2000). Spans joined with *incorrect* discourse connectives can seem logically incoherent although they are independently grammatical:

> *Jazz is good, **because** my favorite is country music.*

The importance of generating good discourse connectives are recognized in prior work in NLG (Biran and McKeown, 2015; Callaway, 2003).

We examine to what extent does GPT-2 generate texts that uphold *plausible* discourse relations, once a discourse connective (usually 1-2 tokens) is generated. We present a comprehensive analysis of discourse connectives in both *fine-tuned* generation—specifically, open domain dialogue generation—and *organic* generation directly from GPT-2. We find that GPT-2 generates valid discourse connectives when the relation can be inferred by humans

52

with high agreement, yet struggles to recover less obvious relations. Our manual analysis reveals the most common connective error is that the relations, signaled by the connectives, *do not hold* between the spans they connect. To this end, we propose a simple remedy: train a connective prediction model and replace incorrect connectives in a post-processing step. This method improves agreement between human and machine-generated connectives in both the fine-tuned and the organic scenarios. Collectively, our results highlight the importance of inferring discourse relations (Xue et al., 2015), and explicitly incorporating discourse information in language models (Ji et al., 2016), to increase their downstream efficacy.

2 Experimental Setup

Fine-tuned. We choose open-domain dialog generation as our *fine-tuned* scenario. The model conditions on a prompt (dialog turn) and generates a response (next turn). We use the PERSONACHAT (Zhang et al., 2018) data for the ConvAI2 challenge. We use 122,499 prompt-response pairs for training and 4,801 pairs for validation.

We fine-tune GPT-2 medium (345M parameters). For compatibility with GPT-2's pre-training, we concatenate the prompt and response (separated by a delimiter) during training. GPT-2 is fine-tuned for 3 epochs using Adam (Kingma and Ba, 2015) with a learning rate of 5e-5. The cross-entropy (language modeling) loss is only calculated for the response. At test-time, the model is conditioned on the prompt (and delimiter) and generates the response. Our approach is similar to Zhang et al. (2020) and we follow Ko et al. (2019) to encourage generation of informative responses.[1]

For decoding, we experimented with both top-k sampling (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2019), and picked the better performing one upon manual inspection of the validation data. We use top-k (k=10) in this scenario.

For quality assurance, we manually evaluate GPT-2's generated responses against SpaceFusion (Gao et al., 2019), a state-of-the-art RNN-based model, re-trained on PERSONACHAT. The evaluation is conducted on Amazon Mechanical Turk, where 5 annotators (per HIT) chose between GPT-2 and SpaceFusion responses. GPT-2 (45.5% chosen) largely outperforms SpaceFusion (16.9%

chosen). For the other 37.7%, the two are tied.

"Organic" generation. To determine to what extent GPT-2 understands the discourse functions of connectives without the effects of fine-tuning, we engage an *organic* scenario. In this case, we pick out utterances with explicit discourse relations in the dataset, and feed the partial utterance that approximates the first argument of an explicit discourse relation (the part before the discourse connective), along with the connective, into the GPT-2 model; we then let it continue to generate the rest of the utterance. We use PERSONACHAT to make the results more comparable to the fine-tuned scenario.[2] We again experimented with both nucleus sampling and top-k, and used nucleus sampling ($p = 0.9$) which performed better upon manual inspection.

3 Assessing explicit discourse relations

At a high level, our assessment strategy compares discourse connectives from GPT-2 outputs with human judgment, following existing strategies of discourse relation annotation, which ask annotators to insert connectives between text spans (Prasad et al., 2008; Scholman and Demberg, 2017; Yung et al., 2019). A discourse connective can be considered valid if humans would also insert a connective signaling the same discourse relation when the connective is masked.

Extracting sentences with discourse connectives. We follow prior work (Braud and Denis, 2016; Ma et al., 2019) in the use of heuristics to extract sentences with discourse connectives, using a list of 11 connectives most frequently observed in PERSONACHAT: *after, and, because, before, but, if, since, so, though, when, while*. Specifically, a clause (using verbs as approximations) needs to appear before and after the connective; the connective cannot be immediately followed by a punctuation; and only *and* and *but* can follow a period. We remove instances of *so* immediately followed by an adjective or adverb. Upon manual inspection of a random sample of 133 extracted sentences, 100% of them contain an explicit discourse relation.

In the PERSONACHAT training set, ~11% of the responses contain one of the connectives. In contrast, the fine-tuned model generates a connective

---

[1] Ko et al. (2019) used a linguistic metric which performed better than mutual information also used in Zhang et al. (2020).

[2] We do not explicitly perform quality assurance for this scenario as we do not fine-tune GPT-2. Details of language modeling performance are discussed in Radford et al. (2019).

	after	and	because	before	but	if	since	so	though	when	while
PERSONACHAT	1.4	40.7	4.2	1.1	28.5	4.4	2.8	4.8	1.1	8.8	2.1
Fine-tuned	0.5	45.7	1.7	0.4	35.9	1.6	2.6	3.7	0.2	5.3	2.4
Organic	0.5	51.4	4.4	1.0	22.1	5.7	1.5	5.8	0.7	5.1	1.8

Table 1: % of sentences with a particular discourse connective, of all sentences that contain a connective.

26% among all responses, and the organic one 15%. The increase in percentage is likely because connectives are frequent words in the corpus. Table 1 shows the relative frequencies of these connectives. Notably, the distribution of connectives is skewed, with *and* and *but* appearing much more often than other connectives, a characteristic similar to other collected examples of discourse relations in the conversation domain (Ma et al., 2019).

Annotating discourse relations. To assess if GPT-2 generate valid discourse connectives, we compare relations signaled by these connectives with relations that humans judge to hold given the rest of the sentence, as in a masked language modeling task. Specifically, for each output sentence that contains a discourse connective, we mask the connective[3] and show the rest of the sentence to annotators (in the case of dialogue generation, we also show the prompt). They are asked to fill in the blank with a connective that most naturally expresses the relation between the arguments, or NONE if they think the two segments are not related. This type of insertion is used previously to crowdsource discourse relations (Yung et al., 2019; Scholman and Demberg, 2017). To reduce label sparsity, we group the connectives into the four top-level discourse relations in the Penn Discourse Treebank (Prasad et al., 2008) (*contingency, contrast, expansion, temporal*), and the annotators are asked to choose a group if it contains the connective they think most appropriately fills the blank. To further help annotators, we included unambiguous synonyms of connectives to anchor the relations more. For ambiguous connectives in our list, we put them in all possible relations they signal. The specific groupings are listed below:

- because, therefore, if, so, since (CONTINGENCY)

- but, although, though, however, whereas, while (CONTRAST)

---

[3]The workers saw an underlined blank space for the mask. If multiple connectives exist, we only consider the first one in this work.

	Fine-tuned	Organic
5	40.9	27.7
4	27.5	25.0
3	21.3	30.8

Table 2: % of sentences where the discourse relation is agreed by $n \in \{3, 4, 5\}$ annotators.

	Fine-tuned	Organic
contingency	6.4	12.5
temporal	5.1	6.2
contrast	35.1	27.1
conjunction	52.5	53.0
no relation	0.9	1.1

Table 3: % of annotated majority relations.

- before, after, when, since, while (TEMPORAL)

- and, in addition (EXPANSION)

We also give the NONE option if the annotator could not find a suitable connective or that the two text spans are not related.

We use Amazon MechanicalTurk to crowdsource annotations for 1.2K sentences with discourse connectives each for the organic and fine-tuned scenarios. Each sentence is annotated by five workers. As quality control, we only allow workers in the US that have completed more than 500 hits with an acceptance rate of >98%.

Table 2 shows the percentage of sentences whose discourse relation is agreed upon by 5, 4, and 3 workers; Table 3 shows the frequency distribution of majority relations (one that is agreed by ≥ 3 workers). For the fine-tuned case, 89.7% of the sentences have a majority relation; inter-annotator agreement measured by Krippendorff's alpha is 0.508, indicating moderate agreement (Artstein and Poesio, 2008). This shows that in most cases, readers are able to infer a discourse relation between the spans of text given, and they do so consistently. Similarly in the organic case, 83.5% of the sentences have a majority relation. However, relations agreed by ≥ 4 workers are much fewer; Krippendorff's alpha is also at a lower value of 0.382. After adjudicating 70 examples with no majority, we find

54

	Fine-tuned	Organic
≥ 3	81.5	74.9
5	94.0	92.6
4	75.6	79.2
3	64.3	53.2

Table 4: % of connectives in generated texts that are consistent with human annotation, stratified by the # of annotators agreeing on the relation.

	Fine-tuned		Organic	
	GPT-2	predicted	GPT-2	predicted
≥ 4	0.781	0.828*	0.839	0.883*
≥ 3	0.760	0.789	0.726	0.766*

Table 5: Consistency between human annotated and predicted discourse relations, measured in macro-F1 of the four relation types. ($\geq n$): $\geq n$ annotators agree on a relation. (*): $p < 0.05$ on a bootstrapping test.

Figure 1: Confusion matrix for human labeled relations vs. generated connectives (after grouping into relations). Darker color indicates more instances. Vertical axis: human annotated relation. Horizontal axis: GPT-2.

that the cause of lower inter-annotator agreement is likely due to the fact that more than one relation can often hold, and in other cases, the quality of the generated text is low.

Assessment results. Table 4 shows the percentage of sentences where the connective in the generated text agrees with the majority relation annotated by humans; we also show the results stratified by how many people agree on the relation. For the connectives *since* and *while* which can signal two relations, we count the model as correct if either relation is annotated by humans. The results reveal that a wrong connective could be a prominent source of error in GPT-2 generation, though the fine-tuned model agrees better with humans. Notably, for relations that humans agree more consistently, the models also generate correct relations more often. This hints that GPT-2 captures obvious, unambiguous relations better. Figure 1 shows a confusion matrix comparing human labeled relations (where at least 3 annotators agree) with GPT-2 generated ones.

4 Fixing discourse connectives

As a first step to fix erroneous connectives, we propose a post-processing technique that does not require retraining a model or modifying model structure: replacing generated discourse connec-

tives with ones from a connective prediction model. This task is related to discourse relation classification (e.g., Xue et al. (2015), Nie et al. (2019)), yet there is no annotated corpora on the dialog domain. While Ma et al. (2019) mined discourse relations from conversations, using their data yielded inferior performance in preliminary experiments.

Connective prediction model. We train a model to predict the masked discourse connective given the rest of the sentence, or NONE if no relation. For training, we extract 1 million sentences from Reddit that contain discourse connectives, using the heuristics in Section 3. We restrict the length of sentences to be 7-25 tokens, similar to that in PersonaChat. The model is fine-tuned on the pre-trained BERT-base-uncased model (Devlin et al., 2019), where the text before the connective is used as sentence A, and text after the connective is used as sentence B. We add an additional classification layer taking the learned [CLS] representation as input. To obtain training data for the NONE class, we add 300K synthesized examples with sentence A and sentence B sampled from different posts, approximating the absence of discourse relations.

The model is fine-tuned for 3 epochs on Reddit using a learning rate of 5e-6. The classification accuracy on the validation set of PERSONACHAT is 0.743 and macro-F1 is 0.649. In the organic setting, we directly apply this model to predict the masked connective. In the fine-tuned setting, to obtain a better model in the domain of PERSONACHAT, we fine-tune the model for 1 epoch on the training set of PERSONACHAT. The classification accuracy improved by 3% and macro-F1 by 5%.[4]

Post-processing results. With this connective prediction model, we replace connectives in generated outputs with the predicted ones. We evaluate whether the predicted connectives align better with

---

[4]Note that this improvement does not translate to a better model for the organic scenario, since GPT-2's output without fine-tuning does not fall in the PERSONACHAT domain.

	Fine-tuned		Organic	
	GPT-2	predicted	GPT-2	predicted
≥ 4	86.8	89.2*	86.6	91.4*
≥ 3	81.5	82.9	74.9	80.4*
≥ 2	84.1	85.9*	80.0	84.5*

Table 6: Consistency between human annotated and predicted discourse relations, measured in *accuracy*. ($\geq n$): calculated on all sentences that $\geq n$ annotators agree on a relation. (*): $p < 0.05$ on a binomial test.

human judgments, after collapsing to discourse relation types. We see the NONE prediction (4.4% for fine-tuned and 17.5% for organic) as an indicator that the sentence is not coherent, and resample from the model for a new sentence. These cases are not included in the results. Appendix A shows several examples illustrating connectives in the generated text and those predicted by the classifier.

Table 5 and Table 6 show the consistency between a connective in the sentence and its corresponding human labeled discourse relation after post-processing, measured by macro-F1 and accuracy respectively. We stratify results according to the agreement among human annotators. We also show the accuracy of cases where ≥ 2 annotators agree to account for the possibility of multiple valid relations. For both fine-tuned and organic scenarios, the predicted connective aligns closer to human labels than those generated by GPT-2.

Figure 2 compares the prediction between GPT-2 and the connective predictor for post-processing (Fig. 2(a)). It illustrates the types of relations that the connective model replaced correctly (Fig. 2(b)) and incorrectly (Fig. 2(c)). This shows that the better performance of the model is not due to simply preferring the most frequent class.

The improvement is notably more substantial for the organic case, an indication that fine-tuning GPT-2 nudges the model very close to what the connective prediction model learns. The overall improvement is likely due the connective prediction model having access to text before and after the connective, while the initial language generation model does not. This finding points to future work on considering stronger discourse-related signals (Ji et al., 2016) and stronger models for inferring relations.

5 Conclusion

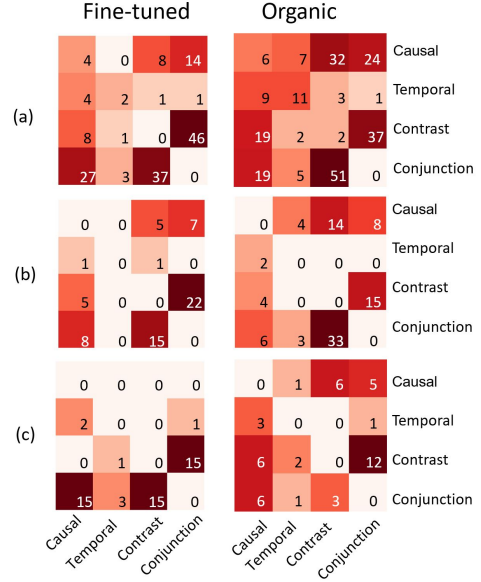This work presents an assessment of discourse relations in organic and fine-tuned language generation

Figure 2: Confusion matrix for GPT-2 (vertical axis) vs. connective prediction model (horizontal axis). Darker color indicates more instances. (a): all *changed* connectives; (b): sentences that the GPT-2 connectives are inconsistent with human labels, but the connective prediction model gave correct predictions; (c): sentences that the GPT-2 connectives are consistent with human labels, but the connective prediction model gave incorrect predictions. Changed connectives in the same relation class are also included.

from GPT-2. We find that the understanding of discourse connectives are present in these models but are limited, especially when the relation requires more inference. We present a post-processing strategy to replace generated connectives, such that they align better with human expectation.

Acknowledgements

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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.

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 203–213.

Charles B. Callaway. 2003. Integrating discourse markers into a pipelined natural language generation architecture. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 264–271.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1229–1238.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *Proceedings of the Eighth International Conference on Learning Representations*.

Rosalind Horowitz. 1987. Rhetorical structure in discourse processing. In *Comprehending oral and written language*, pages 117–160. Academic Press.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations*.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-informed specificity and semantic plausibility for dialogue generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3456–3466.

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Mingyu Derek Ma, Kevin Bowden, Jiaqi Wu, Wen Cui, and Marilyn Walker. 2019. Implicit discourse relation identification for open-domain dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 666–672.

Bonnie JF Meyer and Roy O Freedle. 1984. Effects of discourse type on recall. *American Educational Research Journal*, 21(1):121–143.

Keith K Millis, Arthur C Graesser, and Karl Haberlandt. 1993. The impact of connectives on the memory for expository texts. *Applied Cognitive Psychology*, 7(4):317–339.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *Language Resources and Evaluation Conference*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Ted JM Sanders and Leo GM Noordman. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, 29(1):37–60.

Merel Scholman and Vera Demberg. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 843–861.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2153–2162.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. In *NeurIPS 2018 CAI Workshop*.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 1–16.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

A Example sentences

We show several examples below for both fine-tuned and organic scenarios. We list the text that GPT-2 generated (with the connective bolded), and the connective that our classifier predicted (in the subsequent line).

A.1 Fine-tuned

- *GPT-2*: I do work out at the gym **but** not as often.
  Connective classifier: but
  (In this case, GPT-2 produced a plausible connective, and the classifier also predicted the same connective.)

- *GPT-2*: My husband is a detective **so** he loves my family .
  Connective classifier: and
  (In this case, GPT-2 did not produce a plausible connective, and the connective classifier was able to correct it.)

- *GPT-2*: I 'm a housewife , **but** i also take care of my children
  Connective classifier: but
  (In this case, GPT-2 did not produced a plausible connective, neither did the classifier.)

A.2 Organic

- *GPT-2*: It was hard for me to get into college **and** I 'm still in a wheelchair.
  Connective classifier: because
  (In this case, GPT-2 did not produce a plausible connective, and the connective classifier was able to predict a more plausible one.)

- *GPT-2*: I agree . they insist that **while** they will not pursue civil or criminal action , that they have agreed to withdraw their complaints.
  Connective classifier: while
  (In this case, GPT-2 produced a plausible connective, and the classifier also predicted the same connective.)