Improving Textual Cohesion by Combining Pre-Trained Models

Yifan Bai yifan.bai@mail.mcgill.ca 260562421
Benjamin LeBrun benjamin.lebrun@mail.mcgill.ca 260864955
Krystal Xuejing Pan xuejing.pan@mail.mcgill.ca 260785873

Abstract

Evidence has shown that *cohesion*, the elements of a text that give it a quality of unity, are critically related to reader understanding. (McNamara et al., 2014) Despite this importance, language models struggle to consistently produce cohesive text in the task of natural language generation. (Pishdad et al., 2020) With the goal of generating text exhibiting greater levels of cohesion, we propose a hybrid language model which uses BERT as an encoder and GPT-2 as a decoder. Using a prompted text generation task, we assess the model for cohesion across three dimensions. Overall, we do not find evidence to suggest that a hybrid model produces more cohesive text than a vanilla GPT-2.

1 Introduction

Pre-trained transformer-based (Vaswani et al., 2017) language models have had a significant impact on the field of natural language processing (NLP). Indeed, perhaps the most influential iterations of these models, OpenAI's GPT-2 (Devlin et al., 2019) and Google's BERT (Radford et al., 2018), have in recent years achieved remarkable success across a variety of NLP tasks.

At the same time, language models struggle to consistently generate text that achieves human-level *cohesion*. (Cho et al., 2019) Cohesion, defined as the elements of a text that give it a quality of unity, has been shown to be critically related to reader understanding. (McNamara et al., 2014) In this project, we aim to address this important textual characteristic by augmenting GPT-2 with aspects of BERT with the goal of building a model which generates more cohesive text when compared to GPT-2 alone.

More precisely, we combine BERT's bidirectional encoder-only architecture, which optimally encodes textual representations, with GPT-2's decoder-only architecture, which has been shown to perform well in the language modelling task. We hypothesize that this augmentation will produce a hybrid model with superior contextual representations compared to GPT-2 alone all while retaining GPT-2's language modelling capabilities. This, we expect, will lead to generated text that is more cohesive than text generated by GPT-2.

We test this hypothesis using a prompted text generation task. We evaluate GPT-2's and our hybrid model's generations for explicit markers of textual cohesion using three of the measures outlined in Coh-Metrix (McNamara et al., 2014), a classical tool in computational linguistics. Additionally, we use ROUGE to evaluate the quality of each model's generation with respect to the ground truth text.

2 Related Work

Combining language models Ever since the release of GPT and BERT, there has been word done attempting to integrate a bidirectional encoder with an auto-regressive decoder. BART (Lewis et al., 2020), a transformer-based pre-trained model, combines a BERT-like bidirectional encoder with a GPT-2-like auto-regressive decoder, but is nonetheless different from a hybrid GPT-2/BERT model. Similarly, (Rothe et al., 2019) developed a sequence-to-sequence model that is compatible with combinations of BERT, GPT-2 and RoBERTa checkpoints. However, they did not evaluate their models on prompted generation tasks, nor did they assess for variation in cohesion.

Cohesion and coherence A wide range of research in NLP has tackled the problem of modelling textual *coherence*, including discriminative neural models of coherence (Xu et al., 2019), generative neural models (Li and Jurafsky, 2017), and classical entity-grid discourse representations

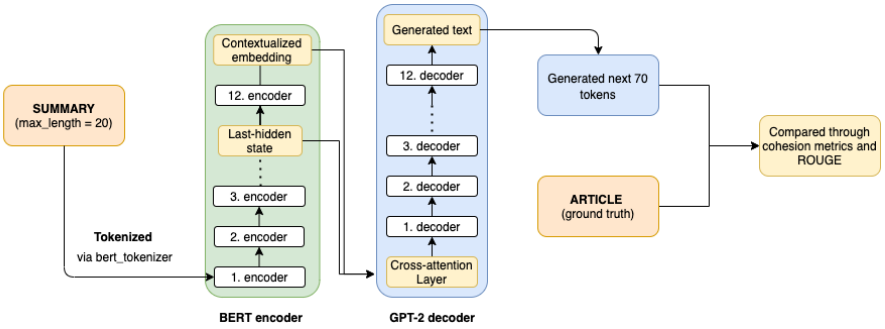

Figure 1: The hybrid model pipeline with BERT uncased and GPT-2_{small}.

(Barzilay and Lapata, 2008), just to name a few. Other work has focused on encoding properties of coherence into the training signal. (Cho et al., 2019)

Though, while intrinsically related, coherence should not be conflated with cohesion. Whereas coherence is a discourse property that is concerned with the logical and semantic organization of a passage, cohesion measures the use of linguistic devices to tie together local textual units. Indeed, some have even claimed that coherence can be thought of as lying in the mind of the reader, whereas cohesion lies in the text. (McNamara et al., 2014)

Common evaluation techniques in coherence modelling include sentence order discrimination tasks or correlation with human ratings of coherence, neither of which explicitly measures the linguistic aspects of cohesion in text. Since we are interested in these "cohesive cues", we opt for simpler and more transparent measures of cohesion drawn from Coh-Metrix. We expect these metrics to assess cohesion at a granular level, specifically, across our three dimensions of study: lexical cohesion, syntactic complexity, and referential cohesion.

3 Methods

3.1 Data

We use two data sets consisting of summary and article pairs: Daily Mail (Hermann et al., 2015) and Reddit `tl;dr` posts (Syed et al., 2018). The Daily Mail corpus consists of 220k samples, with a maximum summary length of 150 tokens and a maximum article length of 1223 tokens. Reddit, on the other hand, consists of 3.05 million samples, with maximum summary length of 324 tokens and maximum article length of 405 tokens. We conducted an exploratory analysis on the summary and

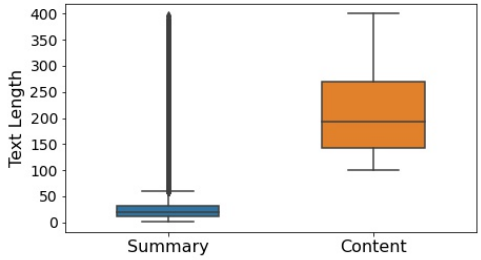

Figure 2: Box plot of Reddit corpus summary and content lengths.

content lengths of the Reddit corpus. Since the corpus is over 12GB in size, we only trained a portion of it.

Figure 2 shows that the 25^{th} to 75^{th} percentile of summary lengths are 15 and 40, respectively, while content lengths of 150 and 260 define the 25^{th} to 75^{th} percentile. The medians are 25 and 190 for summary and content lengths, respectively. Truncating the corpus based on these values, we are left with a total of 1 million instances. Attempting to remain close to the median value, we feed the models summaries of length 20 and generate a 70-token sequence that is approximately $\frac{1}{3}$ of the median content lengths and $\frac{1}{2}$ of the first quartile. In order to keep the comparisons similar, we used the same setup for the Daily Mail corpus, as the converse would not be possible. Note that since we are constrained by data set size, we do not truncate this corpus.

3.2 Model

In terms of computational resources, we fine-tuned on Google Colab and GCP's Tesla T80 with 12 GB of memory. For GPT-2, we used the 124M parameter `GPT-2 small` model, which takes approximately 490 MB on disk, and consumes 9 GB of GPU memory. Therefore, we could only run

batch size of 1. For each model, we train 10 epochs with `PyTorch AdamW` optimizer, and setting the warm-up steps to 10000 and learning rate to 0.0002, as recommended by default.

We used the default GPT-2 settings, which consists of a vocabulary size of 50,257 and a maximum sequence length of 1,024. (Radford et al., 2018) We truncated each post's content length to match this size. The model has 12 decoder layers, uses Byte-Pair Encoding (BPE), and, given an input prompt, returns a probability distribution over the vocabulary. To fine-tune the model, we input a concatenated vector consisting of the summary and article content. During the generation process, we prompted the model with 20 tokens from the summary text and sampled the next 70.

As previously mentioned, we also build a hybrid model which leverages BERT's contextual embedding in an attempt to provide a causality-infused representation. As our encoder, we used Huggingface's `bert-base-uncased` variant, which was trained on a large data set of English text, and consists of 12 layers, 768 hidden states, 12 attention heads, 110M parameters and vocabulary size of 30,522. We fed the summary into it and took the cross-attention of the last hidden layer and the contextual embedding, and then fed fine-tuned GPT-2's decoders via `HuggingFace PretrainedEncoderDecoder` module, as depicted in Figure 1.

3.3 Evaluation

We pre-process the data by removing punctuation and special/infrequent characters. Although GPT-2 does not require texts to be lower cased, we needed to do this to use uncased BERT. We then feed a truncated summary text of length 20, as well as the full content text. In prediction stage, we prompt the model with 20 tokens from the summary text and generate the next 70 tokens. We keep the first as the most likely result. We also use a training/test split of 80/20, and fine-tune the models with training split and generate predictions with test split.

We evaluate cohesion across three dimensions outlined in (McNamara et al., 2014): lexical diversity, syntactic complexity, and referential cohesion. To evaluate the relevance of the generation with respect to the ground truth, we also use the text summarization metric ROUGE.

In all cases, we parse and tokenize samples using `Spacy`.[1] Furthermore, since our samples consist of a maximum of 70 tokens, we trim the ground truth samples so that they are under 70 tokens in length all while maintaining sentences boundaries.

MTLD Lexical diversity measures the variety of unique words (*types*) in a text as a function of the total number of words *(tokens)*. If the number of types in a text is equal to the number of tokens, lexical diversity is at a maximum. In such cases, the text is likely to be very low in cohesion, as new words need to be constantly integrated into the discourse context. (McNamara et al., 2014) When lexical diversity is lower, however, words are used multiple times across the text, and cohesion is therefore assumed to be higher. (McNamara et al., 2014)

A well known index for lexical diversity is the type-token ratio (TTR), which is simply the number of types divided by the overall number of words. However, the TTR is correlated with text length, which makes it unfavourable for comparisons of varying lengths. To overcome this confound, we use MTLD (McCarthy and Jarvis, 2010), which is the mean length of sequential word strings in a text that maintain a given TTR value (.72[2]). We consider a lower MTLD score, that is, lower lexical diversity, to be indicative of greater lexical cohesion.

SYMED Texts which employ a more uniform and consistent syntactic structure tend to be easier to process and comprehend. (McNamara et al., 2014) To measure syntactic consistency, we calculate the minimal edit distance (MED) of the parts of speech between adjacent sentences. To compute MED, we use the `editdistance`[3] implementation of Levenshtein distance. Essentially, this measure computes how much modification is required for two adjacent sentences to have the same syntactic structure. For a given sample, we take the mean MED normalized by sum of the two sentence lengths for all sentence bigrams. We expect texts with a lower syntactic MED to exhibit greater cohesion.

ROVLP Referential cohesion, or coreference, is a linguistic cue which aids readers in making making connections between sentences. (McNamara

[1] `https://spacy.io/`

[2] This value is recommended in (McCarthy and Jarvis, 2010).

[3] `https://pypi.org/project/editdistance/`

Corpus	Model	MTLD	SYMED	ROVLP	ROUGE
		mean (sd)	mean (sd)	mean (sd)	F1 score
Reddit	GPT-2	75.37 ± 20.87	$\mathbf{1.1089 \pm 0.3637}$	$\mathbf{0.0439 \pm 0.0414}$	$\mathbf{96.46}$
	Hybrid	75.34 ± 20.97	1.2176 ± 0.4565	0.0416 ± 0.0397	74.51
	True text	80.03 ± 21.77	1.1248 ± 0.3737	0.0448 ± 0.0416	N/A
Daily Mail	GPT-2	94.94 ± 24.91	$\mathbf{1.0585 \pm 0.3767}$	0.0362 ± 0.0369	41.65
	Hybrid	$\mathbf{90.31 \pm 22.86}$	1.2720 ± 0.5610	0.0357 ± 0.0360	$\mathbf{64.98}$
	True text	89.14 ± 22.05	0.8592 ± 0.1636	0.0468 ± 0.0443	N/A

Table 1: Cohesion metrics mean and standard deviations and ROUGE evaluation results. Mean values closest to ground truth are denoted in **bold**. For cases in which the distributions in question did not differ in a statistically significant manner, neither value is bolded.

et al., 2014) We can assess local referential cohesion by calculating the overlap between words of adjacent sentences. Specifically, for a sentence bigram (s_1, s_2) consisting of adjacent sentences, we calculate the overlap as the number of content words (adjectives, nouns, adverbs, and verbs) which occur in both s_1 and s_2, normalized by the sum of the lengths of s_1 and s_2. For a given sample, we calculate the mean overlap across all sentence bigrams.

ROUGE Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004) is a set of metrics used to evaluate generated text with respect to a corresponding ground truth. While ROUGE is typically used to evaluate automatic summarization or machine translation output, we use the unigram F1 score to evaluate similarity between the ground truth article and the generated text.

Interpreting results While an increase or decrease in the values of each cohesion metric is associated with an increase or decrease in overall cohesion, it should be stressed that this directionality is not stringent. For example, while lower values of MTLD are considered to be indicative of greater cohesion, this relationship does not apply across the entire spectrum of values, for a text with extremely low lexical diversity will likely be incomprehensible. Therefore, we consider a model's output to be maximally cohesive when its cohesion values match those of the ground truth article.

Finally, we note that samples consisting of only a single sentence will be excluded from the calculation of sentence bigram based measures. Furthermore, when comparing metric distributions, we remove outliers by excluding values which are not in the range $[Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)]$, where Q_i is the i^{th} quartile and IQR is the interquartile range of the distribution.

4 Results

Table 1 presents the evaluation results for each model across both data sets. We find statistically significant differences in the distributions of cohesion metrics values in four of the six comparisons (Mann Whitney U test, $p < 0.05$).[4] In three of the four cases, mainly syntactic complexity across both corpora and referential coherence across the reddit corpus, we find that the original GPT-2 model produces text which scores closer to the ground truth than the hybrid model.

However, for lexical diversity across the daily mail corpus, we find that the hybrid model does indeed produce text scoring closer to the ground truth than GPT-2. While this result indicates that the hybrid model may produce more lexically cohesive text, the overall evaluation outcome offers evidence to refute our hypothesis that a hybrid model will produce text which is more cohesive than GPT-2.

In the case of ROUGE, table 1 indicates that neither model performs best across both corpora. Indeed, we find that GPT-2 obtains a near perfect unigram ROUGE score on the reddit corpus, whereas the hybrid model outperforms GPT-2 on the daily mail corpus. Given the bidirectionally of these results, we cannot conclude that one model is necessarily better than the other at producing text that is relevant to its prompt.

5 Discussion & Conclusion

Discussion Given the results described in the previous section, we do not find evidence to suggest

[4]For the MTLD comparison across the reddit corpus, we use an independent-samples t-test ($p < 0.05$) since the distributions had equal variances (Levene test, $p < 0.05$). In all other cases, we use a Mann Whitney U test, as the distributions did not have equal variances.

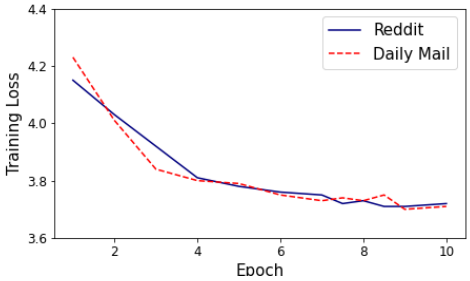

Figure 3: GPT-2 Training Loss for both corpora.

that a hybrid BERT/GPT-2 model generates more cohesive text. While it is not possible for us to know exactly why this is the case, we postulate three reasons. Firstly, the hybrid model architecture only uses BERT's contextualized embedding for the prompt, which is only a portion of the information used in the decoding process. Secondly, we expect this to be a case of *garbage-in, garbage-out*. That is to say, since the embedding is only a representation of the data, the model itself cannot augment its quality. In other words, the representation extracted by BERT is upper-bounded by the quality of the data. Finally, it is not clear whether an augmented contextual representation will even lead to increased cohesion.

In addition, we observe high degree of similarity between generated texts from GPT-2 model and the ground truth. This is coupled with the fact that the texts generated from the hybrid model, which, although have some information loss through BERT's encoding process, are still able to generate texts that closely mimic the ground truth. This was confirmed at fine tuning stage, where the model saturated within very few epochs, and that the improvements were small, as shown in Fig 3. This could be attributed by the fact that OpenAI had trained all variants of GPT-2 models on a wide range of publicly available data, and that our own fine-tuning data sets were included in the original training set. It also demonstrates the importance of preventing data set contamination, for without separate test sets, it is not possible to evaluate a model's ability to generalize.

Conclusion In this project, we have attempted to augment the GPT-2 decoder block using a BERT encoder with the goal of increasing cohesion of generated text. We compared the hybrid model to a vanilla GPT-2 model in a prompted text generation task across the Reddit tl;dr and Daily Mail data sets. Overall, we do not find evidence to suggest that our hybrid model leads to increased levels of cohesion. Future work should further explore the relationship between cohesion and contextualized representations.

6 Statement of Contributions

During this project, all three members have shared tasks equally, including programming, researching literature and documentations, brainstorming ideas and writing up proposal and final reports.

Although each individual member has taken more specific ranges of tasks where it was deemed fit, communications have been conducted rigorously and weekly meetings were held to discuss progress and work on issues together. Therefore, the authors attest herein that equal and fair contributions are made by each member.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. 2019. Towards coherent and cohesive long-form text generation. In *Proceedings of the First Workshop on Narrative Understanding*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pages 1693–1701. Curran Associates, Inc.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209,

Copenhagen, Denmark. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Leila Pishdad, Federico Fancellu, Ran Zhang, and Afsaneh Fazly. 2020. How coherent are neural models of coherence? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6126–6138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *CoRR*, abs/1907.12461.

Shahbaz Syed, Michael Voelske, Martin Potthast, and Benno Stein. 2018. Dataset for generating tldr.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762. Cite arxiv:1706.03762Comment: 15 pages, 5 figures.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Florence, Italy. Association for Computational Linguistics.