BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence

Zhecheng Sheng[*], **Tianhao Zhang**[*], **Chen Jiang**[*], **Dongyeop Kang**

University of Minnesota, Minneapolis, MN
{sheng136, zhan7594, jian0649, dongyeop}@umn.edu

Abstract

Measuring the coherence of text is a vital aspect of evaluating the quality of written content. Recent advancements in neural coherence modeling have demonstrated their efficacy in capturing entity coreference and discourse relations, thereby enhancing coherence evaluation. However, many existing methods heavily depend on static embeddings or focus narrowly on nearby context, constraining their capacity to measure the overarching coherence of long texts. In this paper, we posit that coherent texts inherently manifest a sequential and cohesive interplay among sentences, effectively conveying the central theme, purpose, or standpoint. To explore this abstract relationship, we introduce the "BBScore," a novel reference-free metric grounded in Brownian bridge theory for assessing text coherence. Our findings showcase that when synergized with a simple additional classification component, this metric attains a performance level comparable to state-of-the-art techniques on standard artificial discrimination tasks. We also establish in downstream tasks that this metric effectively differentiates between human-written documents and text generated by large language models under a specific domain. Furthermore, we illustrate the efficacy of this approach in detecting written styles attributed to diverse large language models, underscoring its potential for generalizability. In summary, we present a novel Brownian bridge coherence metric capable of measuring both local and global text coherence, while circumventing the need for end-to-end model training. This flexibility allows for its application in various downstream tasks.

Introduction

Text coherence is the quality of a written or spoken text that makes it internally consistent, logical, and easily understandable to the reader or listener. It encompasses the flow and connectivity of ideas, the logical progression of information, and the smooth transition between sentences and paragraphs. Evaluating the coherence of text requires the evaluator to differentiate between a set of individual utterances and naturally flowing discourses such as co-references or grammatical relationships (Wang and Guo 2014; Kehler 2022).

In linguistic theory, the evaluation of text coherence can be performed at both the local and global levels (Agar and Hobbs 1982; Glosser and Deser 1992; Kintsch and van Dijk

---

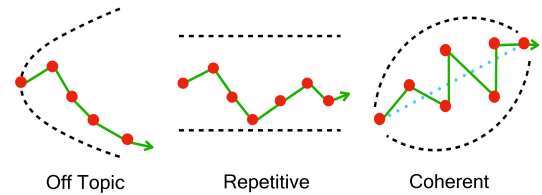[*]These authors contributed equally.

Figure 1: Utterance flow in latent space

1978). Local coherence focuses on the interrelation of information within a specific unit of discourse and its connection to the preceding unit. Conversely, global coherence refers to the degree to which the various components of discourse maintain a cohesive representation of the main topic. Despite the emergence of neural coherence models, most prior works focused on static representations of their local neighboring texts. (Tien Nguyen and Joty 2017; Mesgar and Strube 2018; Mesgar, Ribeiro, and Gurevych 2021; Jeon and Strube 2022)

This paper aims to capture *global coherence* from text, a sequential and cohesive relationship among the sentences or words, representing the main idea, goal, opinion, and other pertinent aspects regarding a coherent text. This relationship may vary across different situations, individuals, and even over time for the same individual. To simplify this concept and enhance its accessibility, we assume that this relationship remains consistent within a fixed domain or dataset. Also, we consider a coherent text should encompass the following properties: (1) A main idea should be kept throughout the text (2) Sentences near the beginning and ending part of the article should emphasize the main idea and (3) Sentences near the middle part can go a little bit far from the main idea, but still be controlled.

We present a novel metric that can capture global text coherence using the Brownian bridge framework, which has been used to analyze animal movements (Horne et al. 2007) or long-text generation (Wang et al. 2023). Brownian bridges serve as continuous-time stochastic models for movement, where the probability of being in a particular area is conditioned on the starting and ending locations, the time elapsed between those points, and the mobility or speed of movement (Horne et al. 2007; Chow 2009). More specifically, we as-

sume each step of this movement can be viewed as a sentence or word within the text, and the trajectory of the movement can represent different levels of coherence. Three such examples are given in Figure 1. In this paper, our contributions and findings are summarized as follows:

- We introduce σ_m in the Brownian bridge as a parameter that characterizes the text style.
- We design a flexible score measurement based on the Brownian bridge that can be used to evaluate the overall coherence of a text and naturally fit into any downstream tasks.
- We show our score can capture both global and local coherence and can effectively differentiate large language models generated and human-written text.

Related Work

Pre-neural models have predominantly relied on entity-based methods (Barzilay and Lapata 2008; Elsner and Charniak 2011) to model coherence in text, as they closely tie to linguistic theory. This class of methods builds upon the Centering theory (Grosz, Joshi, and Weinstein 1995), which assumes that each sentence contains a centered entity that focuses the attention of the entire discourse. These methods provide insights into how information is connected at the local level but often fall short of capturing global coherence. With recent advancements in neural models such as the Transformer (Vaswani et al. 2017), efforts have been made to improve these methods using deep network architectures (Guinaudeau and Strube 2013; Tien Nguyen and Joty 2017; Adewoyin, Dutta, and He 2022), integrating various lexical information to enhance performance.

Previous studies have revealed that Transformer language models struggle to effectively capture coherence structures (Deng, Kuleshov, and Rush 2022). To address this limitation and guide language models in learning text dynamics, representation learning methods have incorporated auto variational encoders (Bowman et al. 2016) or contrastive learning techniques (Gao, Yao, and Chen 2022), which aim to model sentence embeddings by leveraging information from neighboring utterances. However, these approaches yield static representations that overlook the dynamics of the entire text. In a recent study, Wang et al. (2023) proposed an innovative method utilizing the Brownian bridge to model long-range text dependencies. The proposed method exhibited superior performance in generating long and coherent discourses compared to the baseline GPT2 model. In this paper, we extend their findings by introducing a likelihood measurement and conducting experiments to assess its efficacy in capturing both global text coherence and local coherence.

The Shuffle test (Barzilay and Lapata 2008) has been widely utilized as a common practice for assessing coherence models. This artificial task requires the model to distinguish between an original document and its randomly shuffled counterpart. Laban et al. (2021) proposes that it should be treated as a probe—an evaluation task that allows models to be assessed without explicit supervision. Notably, they argue that the Shuffle test is not limited to local coherence analysis and can be extended to longer sequences, such as in the k-Blocked Shuffle test (Van Dijk 1985; Hearst 1997), where the block size can provide insights into a model's ability to discern global coherence or identify main topics and subtopics.

Preliminary

Brownian Bridge   A Brownian bridge is a stochastic process that modifies the standard Brownian motion by requiring an extra constraint (Chow 2009; Revuz and Yor 2013). A standard Brownian bridge $\{B(t)|t \in [0,1]\}$ is a Gaussian stochastic process such that $E[B(t)] = 0$ for all $t \in [0,1]$ and $E[B(t_1)B(t_2)] = \min(t_1, t_2) - t_1 t_2$ for all $t_1, t_2 \in [0,1]$.

However, in our paper, we modified the standard Brownian bridge with a constant called diffusion coefficient and denoted it as σ_m^2 to encode the information of the domain-specific property. Similar methods are also used to model animal mobility (Horne et al. 2007) and human mobility (Krumm 2021), but to the best of our knowledge, it is the first to use this assumption for evaluating text coherence. Formally, this modified Brownian bridge $\{S(t)|t \in [0,T]\}$ follows a normal distribution $S(t) \sim \mathcal{N}(\mu(t), \sigma^2(t)\mathbf{I})$, for $t \in [0,T]$, where

$$\mu(t) = a + \frac{t}{T}(b - a), \quad \sigma^2(t) = \frac{t(T - t)}{T}\sigma_m^2.$$

$S(0) = a \in \mathbb{R}^n$ and $S(T) = b \in \mathbb{R}^n$ are the starting and ending location respectively, and $n \in \mathbb{N}$ is a fixed constant.

Brownian Encoder via Contrastive Learning   To map utterances from the observable space into the designed latent space, we followed the practice from the previous work (Wang et al. 2023). The weights from pretrained GPT2 (Radford et al. 2019) are frozen while at the top of its hidden states, a trainable multi-layer perceptron is appended. The training details of the encoder are as follows. Given a sequence of data points $X = \{x_1, \ldots, x_N\}$, we randomly draw batches $\mathcal{B}$ consisting of randomly sampled triplets $(x_{i_1}, x_{i_2}, x_{i_3})$ with $1 \le i_1 < i_2 < i_3 \le T$ and optimize the encoder by minimize the corresponding contrastive objective function:

$$\mathcal{L} = \mathbb{E}_X \left[-\log \frac{\exp(d(x_{i_1}, x_{i_2}, x_{i_3}; f_\theta))}{\sum_{(x_{i_1}, x_{i_2'}, x_{i_3}) \in \mathcal{B}} \exp(d(x_{i_1}, x_{i_2'}, x_{i_3}; f_\theta))} \right]$$

with $d(x_{i_1}, x_{i_2}, x_{i_3}; f_\theta)$ defined by:

$$-\frac{\|f_\theta(x_{i_2}) - (1 - \frac{i_2 - i_1}{i_3 - i_1})f_\theta(x_{i_1}) - \frac{i_2 - i_1}{i_3 - i_1}f_\theta(x_{i_3})\|_2^2}{2\sigma^2},$$

where f_θ denotes the encoder and $\sigma^2 = \frac{(i_2 - i_1)(i_3 - i_2)}{i_3 - i_1}$. The loss function can be viewed as maximizing towards the triplets sampled from a Brownian bridge process while minimizing the function for triplets sampled from a different sequence. We implemented the encoder training step based on code from (Wang et al. 2023). As one can find that during the training, the diffusion coefficient σ_m^2 does not show up in the σ^2, and instead, this information is encoded into the function, f_θ. After the training is finished, we can approximate
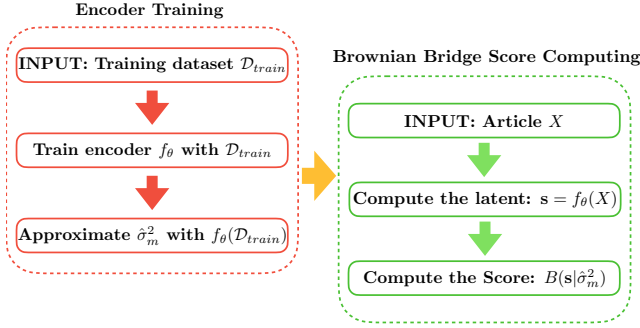
Figure 2: Procedure to generate the Brownian bridge score, $B(\mathbf{s}|\hat{\sigma}_m^2)$.

the diffusion coefficient σ_m^2 based on the training set and the encoder f_θ which will be discussed in the following section.

Proposed Methodology: Brownian Bridge Score for Text Coherence

In this section, we introduce how to approximate the diffusion coefficient and how to design our Brownian bridge score (BBScore). First of all, as shown in Figure 2, we display the general procedure to use our BBScore: Given a domain-specific dataset $\mathcal{D}_{train}$ and train the encoder f_θ with this dataset. Then, approximate the diffusion coefficient σ_m^2 with all the latents $f_\theta(\mathcal{D}_{train})$. After this, we can compute the BBScore of any given article which is assumed to be in the same domain as the training dataset $\mathcal{D}_{train}$. The overall procedure to compute the BBScore is shown in the green dash box. Next, we will introduce how to approximate σ_m^2 and construct BBScore $B(\cdot|\sigma_m^2)$.

The approximation and construction are both inspired by the likelihood of Brownian bridge. Given a Brownian bridge sequence $\mathbf{s} = (s_1, s_2, \cdots, s_{T(\mathbf{s})})$, where $s_i \in \mathbb{R}^n$, for $i = 2, 3, \cdots (T(\mathbf{s}) - 1)$ and $T(\mathbf{s})$ is the length of $\mathbf{s}$, it satisfies: $s_i \sim N(\mu_i, \sigma_i^2 \mathbf{I})$, where $\mathbf{I}$ is the $n \times n$ identity matrix and

$$\mu_i = s_1 + \frac{i-1}{T(\mathbf{s})}(s_{T(\mathbf{s})} - s_1), \ \sigma_i^2 = \frac{(i-1)(T(\mathbf{s})-i)}{T(\mathbf{s})}\sigma_m^2.$$

To estimate σ_m^2 with the training dataset $\mathcal{D}_{train}$, we consider the likelihood: for $1 < i < T(\mathbf{s})$, the likelihood of location s_i is:

$$\mathcal{L}_i = \frac{1}{2\pi\sigma_i^2}\mathrm{Exp}(-\frac{1}{2\sigma_i^2}||s_i - \mu_i||^2).$$

Then the product of the log-likelihood of $\mathcal{L}_i$ is

$$\ln(\prod_{i=2}^{T(\mathbf{s})-1}\mathcal{L}_i) = -\sum_{i=2}^{T(\mathbf{s})-1}(\ln(\alpha_i(\mathbf{s})\sigma_m^2) + \frac{\beta_i(\mathbf{s})}{\sigma_m^2}), \quad (1)$$

where for $i = 2, 3, \cdots T(\mathbf{s}) - 1$

$$\alpha_i(\mathbf{s}) = \frac{2\pi(i-1)(T(\mathbf{s})-i)}{T(\mathbf{s})-1}, \beta_i(\mathbf{s}) = \frac{(T(\mathbf{s})-1)||s_i - \mu_i||^2}{2(i-1)(T(\mathbf{s})-i)}.$$

Next, by setting the derivative of Eq (1) to be zero, we can find the best approximation of σ_m^2 with maximum likelihood: $\hat{\sigma}_m^2(\mathbf{s}) = \frac{1}{T(\mathbf{s})-2}\sum_{i=2}^{T(\mathbf{s})-1}\beta_i(\mathbf{s})$. Therefore, we can approximate the σ_m^2 with the training set:

$$\hat{\sigma}_m^2 = \sum_{\mathbf{s} \in f_\theta(\mathcal{D}_{train})}\hat{\sigma}_m^2(\mathbf{s}).$$

And to further simplify our method, we use the average of n dimension of the approximated $\hat{\sigma}_m^2$ as the diffusion coefficient used in the BBScore. Besides, we show the necessity of such approximation by a simple sensitivity analysis, e.g. the standard Brownian bridge $\sigma_m^2 = 1$ cannot capture the coherence of the article (See Supplementary material). Next, we define the BBScore which is inspired by the log-likelihood:

$$B(\mathbf{s}|\hat{\sigma}_m^2) = \frac{|\sum_{i=2}^{T(\mathbf{s})-1}\ln(\alpha_i(\mathbf{s})\hat{\sigma}_m^2) + \frac{\beta_i(\mathbf{s})}{\hat{\sigma}_m^2}|}{T(\mathbf{s}) - 2}. \quad (2)$$

Based on the likelihood (Eq (1)), this measure should show the following relation with text coherence: the more the text is coherent, the smaller the value of $B(\mathbf{s}|\hat{\sigma}_m^2)$ will be.

Experiments

Research Questions

Throughout our experiments, we address two research questions as follows:

- While designed for global coherence, does BBScore capture both global and local coherence effectively in a synthetic setting?
- Can BBScore detects texts where the departure from desired coherence are not manipulated and recognize their differences ?

Dataset

In this paper, our focus is on the WikiSection dataset (city) as introduced in (Arnold et al. 2019). This dataset comprises Wikipedia articles related to global cities, sourced from Wikipedia dumps. Each segment of text within these articles is meticulously labeled with its corresponding section class and word count.

For our research, we make use of the pre-processed Wiki-Section data provided by Wang et al. (2023). This curated dataset encompasses 2,165 articles in the training set and an additional 658 articles in the test set, all classified into 4 distinct section classes. Unlike the approach taken by Wang et al. (2023), we deliberately exclude any section-related information during the training of the encoder. This decision is aimed at preventing the inadvertent disclosure of supplementary information and enhancing the model's overall applicability.

We then alter the orignal text to create material for each experiment. For the purpose of artificial tasks, we permute the original corpus to create various shuffled versions. As for downstream tasks, we employ large generative language models to complete each section. This is achieved by providing the model with the respective section title and the first few sentences as a prompt.

Additional details regarding the creation of the data are elaborated within the context of each specific experimental setup.

Methods for Comparison

We compare our proposed BBScore against the following coherence metrics:

Entity Grid   Barzilay and Lapata (2005) is the most recognized entity-based approach. For each input document, it creates a two-way contingency table to track the entity appearance in each sentence. The issue with this count-based approach is the grids can easily become sparse as the length of the document increases or when the text is informal (Lai and Tetreault 2018). The algorithm struggles to identify meaning entity transitions across sentences. This issue persists even in the subsequent neural version of Entity Grid. We include these methods in the comparison as the benchmark to validate the efficacy of BBScore. We use Stanford's CoreNLP to annotate the documents and the implementation in coheoka library[*] to obtain the Entity Grid score.

Unified Coherence   Moon et al. (2019) is a neural-based entity-grid method that incorporates sentence grammar, inter-sentence coherence relations, and global coherence patterns into a common neural framework, and it reaches state-of-the-art results in many artificial tasks. However, this model architecture is trained end-to-end with a pairwise ranking loss, and in our opinion, these pairwise constraints can help the model reach a good performance in some tasks, but limits its application and generalization, since in the real world not all the incoherent article is an exact shuffled version of a coherent one.

BBScore   Figure 2 shows a framework for the computation of BBScore. For example, in the WikiSection task, we train the encoder with the original training dataset and approximate the diffusion coefficient with the latents of the training dataset. The score is supposed to represent the deviation from the latent Brownian trajectory from training data.

BBScore + Classifier   We use global BBScore and local moving window BBScore with window sizes (defined in Supplementary material) being 1, 2, 4, and 8 as features showing how coherent the text is. A three-layer perceptron then takes those features as input and outputs predictions for designed tasks.

Artificial Tasks

Global Discrimination   The evaluation of coherence can be effectively assessed through the Shuffle test (Barzilay and Lapata 2008; Moon et al. 2019). This evaluation method involves randomly permuting sentences within a discourse to create an incoherent document and compared to the original one. By employing this task, the overall coherence of the text can be effectively evaluated. To ensure the robustness and sensitivity of our proposed coherence scoring mechanism, we also incorporated the k-Blocked Shuffle test, which introduces varying levels of incoherence into the documents.

---

[*]https://github.com/kigawas/coheoka

	Global		Local	
	Data	#Pairs	Data	#Pairs
Train	$\mathcal{D}_{b=1}$	36,574	$\mathcal{D}_{w=1}$	20,762
	$\mathcal{D}_{b=2}$	22,738	$\mathcal{D}_{w=2}$	26,165
	$\mathcal{D}_{b=5}$	17,110	$\mathcal{D}_{w=3}$	27,069
	$\mathcal{D}_{b=10}$	11,356	$\mathcal{D}_{w=1,2,3}$	73,996
Test	$\mathcal{D}_{b=1}$	8,600	$\mathcal{D}_{w=1}$	6,255
	$\mathcal{D}_{b=2}$	4,280	$\mathcal{D}_{w=2}$	7,924
	$\mathcal{D}_{b=5}$	3,837	$\mathcal{D}_{w=3}$	8,176
	$\mathcal{D}_{b=10}$	2,193	$\mathcal{D}_{w=1,2,3}$	22,355

Table 1: Statistics for the shuffle task dataset

In this test, we first trained the Brownian encoder using the WikiSection training set and then mapped both the original and shuffled documents into the latent space, generating their respective BBScores.

To create the dataset, we generated a single random shuffle for each document and constructed a mixture of raw and shuffled documents, forming the basis of a binary classification task. We used BBScores to discern between coherent and incoherent text without additional supervision. To further align with established methodologies (Joty, Mohiuddin, and Tien Nguyen 2018; Moon et al. 2019; Jeon and Strube 2022), we extended our experimentation to encompass a larger number of shuffles. Specifically, we generated 20 random shuffles for each document, both within the training and test sets. Furthermore, to maintain the integrity of the evaluation process, any documents that matched their original copy were omitted from the analysis. The statistics of the resultant dataset for 20 shuffle global discrimination tasks are shown in Table 1.

Local Discrimination   Detecting incoherent documents in the global discrimination task is relatively straightforward; however, assessing local coherence at the document level presents a more challenging task. To evaluate the effectiveness of our coherence scoring approach for local coherence, we constructed datasets using local permutation windows.

Following the methodology outlined in Moon et al. (2019), we randomly sampled a set of windows, denoted as w, from the entire text. Each window consisted of a fixed size of three sentences, ensuring there was no overlap between any of the windows. Subsequently, we randomly shuffled the order of sentences within each window. We repeated the shuffle procedure for 20 times, excluding any duplicate sample.

We generated three separate datasets for the local discrimination task, namely $\mathcal{D}_{w=1}$, $\mathcal{D}_{w=2}$, and $\mathcal{D}_{w=3}$. Each dataset had a different window size, which determined the degree of evaluation for local coherence. Figure 3 demonstrates examples of incoherence caused by local shuffles.

Downstream Task

As noted in a previous study (Mohiuddin et al. 2021), there is a limited correlation between model performance on artificial tasks and their performance on downstream tasks. Since BBScore is not restricted by a well-defined structure, such as pairwise input, we further explore its potential application

| When the post office opened in 1845, it changed the name to Woodberry . It was first named Sandtown for the white sand that covered the ground . By 1854, the spelling was officially changed to Woodbury . | It was first named Sandtown for the white sand that covered the ground . When the post office opened in 1845, it changed the name to Woodberry .By 1854, the spelling was officially changed to Woodbury . |

Figure 3: Example window pair for local discrimination task. Left is from the locally shuffled article and the right is from the original article.

Figure 4: Left: BBScore distribution with different size of shuffle blocks. Right: Test set BBScore-based AUC with different size of shuffle blocks

areas. We assess our method on two novel downstream tasks: a task aimed at distinguishing LLM-generated articles from human-written ones and another task to detect which LLM the given input texts are from.

AI Differentiation For this task, we use the WikiSection dataset as human-written texts. To obtain LLM-generated texts with similar structures as the human-written ones, we use the first few sentences in each section of the WikiSection article as prompts, ask LLM to complete each section, and then combine those sections into one generated article. We do not compare to Moon et al. (2019) in this task because it requires the text length to be aligned between the pair. We still include the Entity Grid as the baseline.

LLM Detection In order to exam the potential generalizability of BBScore, we form a domain detection task where texts generated by different large language models can be viewed as different domains based on our assumption. Our primary inquiry revolves around the sensitivity of $\hat{\sigma}_m^2$ in distinguishing the employed large language model during the text generation process when provided with input texts.

Results

Global Discrimination Task

We calculate the BBScore for the original training and test set as well as the ones with various shuffle windows, and from their distribution we can conclude coherent text ought to have a lower BBScore, which implies that they are closer to the hypothetical Brownian bridge process. We also rank the scores from all the articles and compute the AUC score for coherent and incoherent separation. The general performance of our model on shuffled tasks is shown in Figure 4. The AUC for Block 1 shuffle reaches 0.835 with the raw BBScore and remains at 0.738 for Block 10 shuffle where the incoherence is harder to detect.

Next, we calculate the pairwise accuracy score for each document with scores derived from different methods, and the result is shown in Table 2. As the block size increases in the block shuffle task, the task becomes more challenging. This is because the shuffled article exhibits greater local coherence, but overall, it still lacks sufficient global coherence. The performance of the basic Entity Grid score on the test
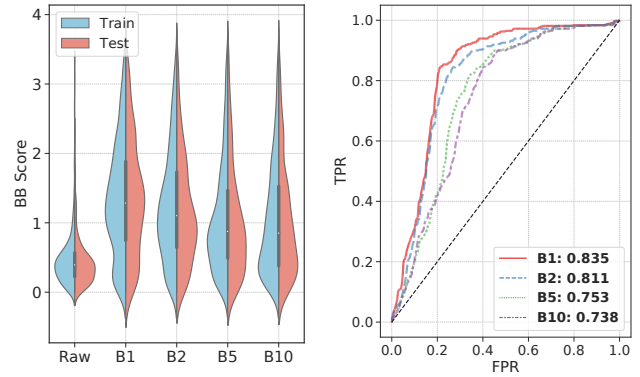
set shows significant variation with changes in block sizes. Consequently, to successfully accomplish this task, the metric should consider both local and global coherence while maintaining a balance between them.

From the result of the basic BBScore, we can find that with an encoder trained on the original Train dataset and diffusion coefficient approximated with latents of the Train dataset, its performance is still very stable even if the task becomes harder. After we add a simple linear classifier after the BBScore, its performance exceeds (Moon et al. 2019) and becomes better when the block size increases. Therefore, we conclude the BBScore has a better evaluation of global coherence while still capturing local coherence. To further explore and verify the ability of BBScore on local coherence evaluation, we implement the local coherence discrimination task in the following section.

Local Discrimination Task

We compute pairwise scores for each document and its locally shuffled counterpart. The accuracy reported is based on whether the model can assign a lower BBScore to the original document compared to the locally shuffled one in each pair. The results can be found in Table 3. As expected, the Unified Coherence Model (Moon et al. 2019) demonstrates strong performance across different locally shuffled datasets, given that their model incorporates sentence alignment and deliberately learns to differentiate between document pairs. Their model includes a dedicated module for capturing inter-sentence relations and modeling local coherence. In contrast, our approach focuses more on global coherence rather than supervising local windows. Nonetheless, our supervised model, BBScore+CLF, demonstrates the fundamental ability to evaluate local coherence.

AI Generated Text Differentiation

In this section, we mainly focus on two different training strategies: (1) Pairwise discrimination; and (2) General AI discrimination.

Methods	Train				Test			
	$\mathcal{D}_{b=1}$	$\mathcal{D}_{b=2}$	$\mathcal{D}_{b=5}$	$\mathcal{D}_{b=10}$	$\mathcal{D}_{b=1}$	$\mathcal{D}_{b=2}$	$\mathcal{D}_{b=5}$	$\mathcal{D}_{b=10}$
ENTITYGRID (Barzilay and Lapata 2008)	79.17	86.20	74.51	60.70	85.73	82.79	75.81	64.65
UNIFIEDCOHERENCE (Moon et al. 2019)	**99.75**	98.60	97.10	96.21	**99.73**	97.86	96.90	96.09
BBSCORE	76.29	75.12	73.04	73.12	83.39	80.71	79.36	78.66
BBSCORE + CLF	99.12	**98.95**	**98.68**	**98.54**	98.92	**98.46**	**98.25**	**98.50**

Table 2: Global Discrimination Task Results on WikiSection

Methods	Train				Test			
	$\mathcal{D}_{w=1,2,3}$	$\mathcal{D}_{w=1}$	$\mathcal{D}_{w=2}$	$\mathcal{D}_{w=3}$	$\mathcal{D}_{w=1,2,3}$	$\mathcal{D}_{w=1}$	$\mathcal{D}_{w=2}$	$\mathcal{D}_{w=3}$
ENTITYGRID (Barzilay and Lapata 2008)	61.25	55.37	62.94	65.44	60.18	53.04	60.83	66.67
UNIFIEDCOHERENCE (Moon et al. 2019)	90.84	84.02	83.64	89.59	87.00	77.47	82.98	87.87
BBSCORE	57.85	47.17	55.84	63.10	57.66	50.29	60.15	64.06
BBSCORE + CLF	69.68	61.69	69.00	75.04	67.12	55.20	67.86	75.66

Table 3: Local Discrimination Task Results on WikiSection. $\mathcal{D}_{w=1,2,3}$ represents the joint set of all three other datasets.

For the pairwise discrimination task, we design a data pair, e.g. (original doc, language model generated doc) as an input to test whether we can distinguish which one was generated by humans. An example is shown in Figure S1 (See Supplementary material), and in this pair, the document from WikiSection has a smaller BBScore than the AI-generated document, so we can successfully detect the AI-generated one. Later, we also train a simple classifier with the BBScore on this pairwise classification task, and the complete result is listed in Table 4, where we use six different large language models to generate the document.

General AI discrimination task is a harder problem. With any document as an input, we train BBScore+CLF to predict whether humans or the AI write this document, the classification accuracy are shown in Table 5. We also compute the AUC score as shown in Figure 5. The AUC corresponds to different large language models that do not vary a lot which implies our model can finish the general AI discrimination task in a stable way.
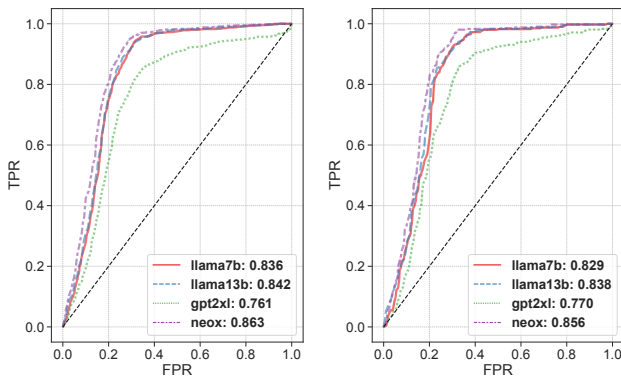

Figure 5: Left and Right correspond to Train and Test BBScore-based AUC with different LLMs, respectively.

LLM Detection

For this task, we initiate by training a Brownian encoder utilizing a composite corpus consisting of articles generated by diverse LLMs and the original articles. Subsequently, we derive estimates for $\sigma^2_{m_k}$ from the training set corresponding to LLM k. Then we calculate $\sigma^2_m(\mathbf{s})$ conditional on the input text $\mathbf{s}$ obtained from the test set. Moving forward, we evaluate the distributional similarity between $\hat{\sigma}^2_m(\mathbf{s})$ and $\sigma^2_{m_k}$, and normalize the outcomes for each $\mathbf{s}$ to ensure the values are on the scale from 1 to 2. The outcomes, as calculated by the Wasserstein distance, are illustrated in Figure 6. In the results, we observe that 5 out of the 6 test texts are in close proximity (among the top 2) to their respective source LLMs, except for the text generated by GPT-NeoX. This illustrates that the comparison of the distribution of σ^2_m effectively distinguishes a narrower range of LLMs to be considered. Given that the input texts pertain to the same subject and variations introduced by different LLMs can be subtle, we find this technique valuable in identifying subdomains in the condition of training with an extensive collection of text and can potentially be generalized into other domains. To examine the generalizability in terms of out of domain performance, we also test the WikiSection encoder on the GCDC dataset (Lai and Tetreault 2018) (See Supplementary material).

Discussion and Limitations

In this paper, we propose the BBScore as a simple yet effective metric for measuring text coherence. Unlike previous entity-based approaches that rely on linguistic insights, the BBScore naturally learns text coherence through a contrastive objective during training. According to (Grosz and Sidner 1986), three factors contribute to discourse coherence: the organization of discourse segments, the intention or purpose of the discourse, and attention or focused items. While well-organized discourse segments are common in formal writing, the abundance of informal text poses challenges in conveying

Methods	Train						Test					
	♣	♠	♦	♥	♢	♡	♣	♠	♦	♥	♢	♡
ENTITYGRID	54.43	36.99	18.64	25.23	43.50	25.86	59.57	36.17	25.88	32.19	39.40	22.85
BBSCORE	76.96	85.07	82.21	82.42	73.07	75.29	77.75	83.74	81.16	81.82	75.63	75.13
BBSCORE + CLF	**83.21**	**85.51**	**88.04**	**88.03**	**92.19**	**91.89**	**83.01**	**86.17**	**86.18**	**86.73**	**92.02**	**91.89**

Table 4: AI Generated Text Differentiation Results on WikiSection. ♣: GPT2XL ♠: GPT-NeoX ♦: LLaMA-7B ♥: LLaMA-13B ♢: LLaMA2-7B ♡: LLaMA2-13B

Dataset	LLM					
	♣	♠	♦	♥	♢	♡
Train	77.33	83.42	84.55	82.80	85.12	86.80
Test	76.44	82.65	82.53	81.20	84.53	85.35

Table 5: General AI Discrimination Task Results of BBScore + CLF on WikiSection. ♣: GPT2XL ♠: GPT-NeoX ♦: LLaMA-7B ♥: LLaMA-13B ♢: LLaMA2-7B ♡: LLaMA2-13B

the main topic or intention through entity transitions. In such cases, explicit guidance for learning text coherence becomes more difficult, necessitating a self-supervised approach.

In Section , we demonstrate that BBScore could be calculated at the sentence level, and by manipulating the sentence window sizes, we can generate a combination of features that capture coherence at different discourse units. Despite its simple network structure, BBScore achieves state-of-the-art results on global discrimination tasks in standard coherence assessment experiments, highlighting its potential for evaluating global text coherence. Additionally, even without a specific structure for capturing local coherence, our model still performs well in evaluating local text coherence. We believe that the diffusion coefficient encodes important domain-specific features that aid in text coherence evaluation. More specifically, we show it can be leveraged for Human-AI discrimination and used to compare generation qualities of different models with respect to coherence. From the LLM detection task results, we find when training with an inclusive corpus, $\hat{\sigma}_m$ can potentially be utilized to identify sub-domains. Our results also demonstrate that while Entity Grid is a simple and effective measure in artificial settings, it falters in downstream tasks.

While our approach has yielded positive results, we also identify several limitations. In this paper, our focus has been on a single dataset to showcase the usability of BBScore. Although we have demonstrated its ability to discern different LLM-generated content by training with a cross-domain corpus, the obtained results are not entirely satisfactory and still require further robustness validation. In light of this, we acknowledge that the assumption we have made, linking σ_m^2 to specific domain/style, is rather restrictive in its general applicability. Future endeavors will involve expanding the parameter space by introducing higher-dimensional variance estimates. Furthermore, given BBScore's demonstrated
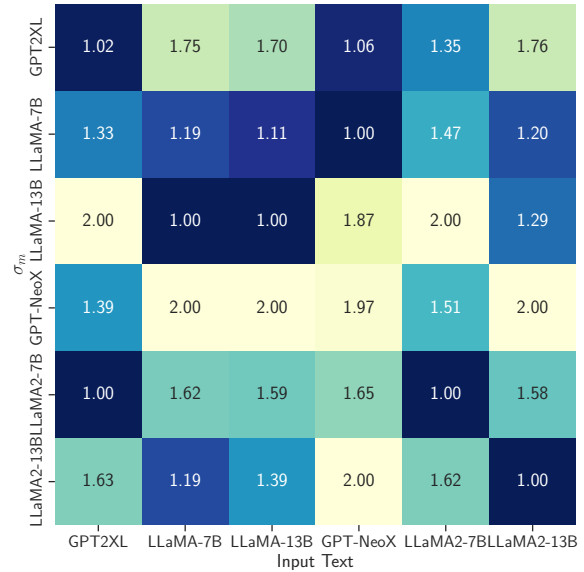

Figure 6: Normalized Wasserstein Distances. Each column represents the input texts are generated by a specific model.

capacity to differentiate between LLM-generated text and human-authored text, we aspire to establish BBScore as a more general and well-defined metric for comparing various LLMs in our forthcoming research.

Conclusion

Overall, BBScore presents a novel perspective on text coherence and demonstrates its efficacy on artificial tasks involving deliberately induced incoherence. Additionally, we illustrate the practical utility of BBScore in natural contexts, where unintentional deviations from desired coherent text occur. Serving as an intermediate computed score, BBScore holds the potential to become a valuable feature in numerous real-world applications, including tasks like Human-AI discrimination. In contrast to the intricate architectures employed in other coherence models (e.g., Multi-layer LSTM), our approach utilizes a simple three-layer perceptron for classification tasks, devoid of any crafted loss function. Remarkably, this approach attains comparable, and in some cases, even superior results. Further refinement of our existing method promises to be an intriguing avenue for future research.

References

Adewoyin, R.; Dutta, R.; and He, Y. 2022. RSTGen: Imbuing Fine-Grained Interpretable Control into Long-FormText Generators. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1822–1835. Seattle, United States: Association for Computational Linguistics.

Agar, M.; and Hobbs, J. R. 1982. Interpreting discourse: coherence and the analysis of ethnographic interview. *Discourse Processes*, 5: 1–32.

Arnold, S.; Schneider, R.; Cudré-Mauroux, P.; Gers, F. A.; and Löser, A. 2019. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *Transactions of the Association for Computational Linguistics*, 7: 169–184.

Barzilay, R.; and Lapata, M. 2005. Modeling Local Coherence: An Entity-Based Approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 141–148. Ann Arbor, Michigan: Association for Computational Linguistics.

Barzilay, R.; and Lapata, M. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1): 1–34.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. arXiv:1511.06349.

Chow, W. C. 2009. Brownian bridge. *Wiley interdisciplinary reviews: computational statistics*, 1(3): 325–332.

Deng, Y.; Kuleshov, V.; and Rush, A. M. 2022. Model Criticism for Long-Form Text Generation. arXiv:2210.08444.

Elsner, M.; and Charniak, E. 2011. Extending the Entity Grid with Entity-Specific Features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 125–129. Portland, Oregon, USA: Association for Computational Linguistics.

Gao, T.; Yao, X.; and Chen, D. 2022. SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv:2104.08821.

Glosser, G.; and Deser, T. 1992. A comparison of changes in macrolinguistic and microlinguistic aspects of discourse production in normal aging. *Journal of Gerontology: Psychological Sciences*, 47: 266–272.

Grosz, B. J.; Joshi, A. K.; and Weinstein, S. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2): 203–225.

Grosz, B. J.; and Sidner, C. L. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3): 175–204.

Guinaudeau, C.; and Strube, M. 2013. Graph-based Local Coherence Modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 93–103. Sofia, Bulgaria: Association for Computational Linguistics.

Hearst, M. A. 1997. Text Tiling: Segmenting Text into Multiparagraph Subtopic Passages. *Computational Linguistics*, 23(1): 33–64.

Horne, J. S.; Garton, E. O.; Krone, S. M.; and Lewis, J. S. 2007. Analyzing animal movements using Brownian bridges. *Ecology*, 88(9): 2354–2363.

Jeon, S.; and Strube, M. 2022. Entity-based Neural Local Coherence Modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7787–7805. Dublin, Ireland: Association for Computational Linguistics.

Joty, S.; Mohiuddin, M. T.; and Tien Nguyen, D. 2018. Coherence Modeling of Asynchronous Conversations: A Neural Entity Grid Approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 558–568. Melbourne, Australia: Association for Computational Linguistics.

Kehler, A. 2022. Coherence Establishment as a Source of Explanation in Linguistic Theory. *Annual Review of Linguistics*, 8: 123–142.

Kintsch, W.; and van Dijk, T. A. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85: 363–394.

Krumm, J. 2021. Brownian Bridge Interpolation for Human Mobility? In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '21, 175–183. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386647.

Laban, P.; Dai, L.; Bandarkar, L.; and Hearst, M. A. 2021. Can Transformer Models Measure Coherence In Text: Re-Thinking the Shuffle Test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 1058–1064. Online: Association for Computational Linguistics.

Lai, A.; and Tetreault, J. 2018. Discourse Coherence in the Wild: A Dataset, Evaluation and Methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 214–223. Melbourne, Australia: Association for Computational Linguistics.

Mesgar, M.; Ribeiro, L. F. R.; and Gurevych, I. 2021. A Neural Graph-based Local Coherence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2316–2321. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Mesgar, M.; and Strube, M. 2018. A Neural Local Coherence Model for Text Quality Assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4328–4339. Brussels, Belgium: Association for Computational Linguistics.

Mohiuddin, T.; Jwalapuram, P.; Lin, X.; and Joty, S. 2021. Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3528–3539. Online: Association for Computational Linguistics.

Moon, H. C.; Mohiuddin, T.; Joty, S.; and Xu, C. 2019. A Unified Neural Coherence Model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2262–2272. Hong Kong, China: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multi-task Learners. *OpenAI Blog*, 1(8): 9.

Revuz, D.; and Yor, M. 2013. *Continuous Martingales and Brownian Motion*, volume 293. Springer Science & Business Media.

Tien Nguyen, D.; and Joty, S. 2017. A Neural Local Co-herence Model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1320–1330. Vancouver, Canada: Association for Computational Linguistics.

Van Dijk, T. A. 1985. Semantic discourse analysis. *Handbook of discourse analysis*, 2: 103–136.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.

Wang, R. E.; Durmus, E.; Goodman, N.; and Hashimoto, T. 2023. Language modeling via stochastic processes. arXiv:2203.11370.

Wang, Y.; and Guo, M. 2014. A short analysis of discourse coherence. *Journal of Language Teaching and Research*, 5(2): 460.