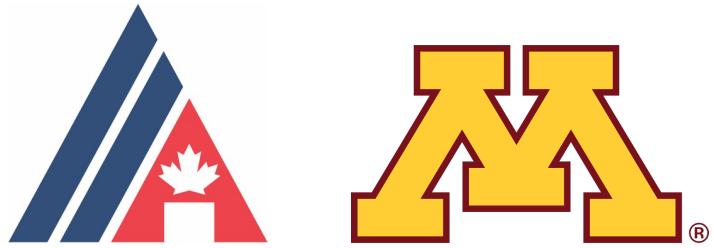


BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence

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

University of Minnesota, Twin Cities

* Equal contributi

Background

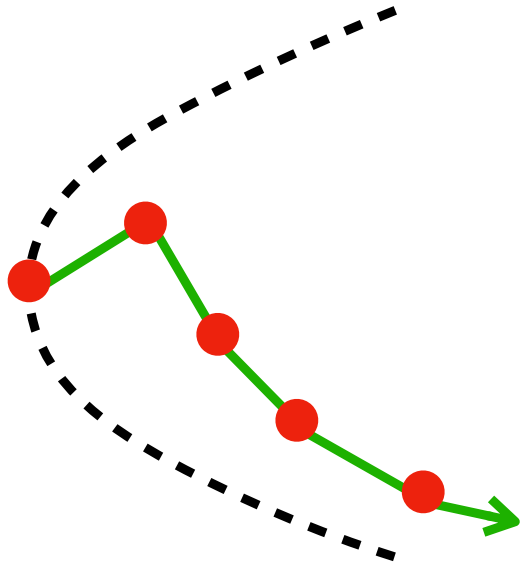
- Measuring text coherence is an essential aspect to assess Human written or machine generated text.

- Coherence can be measured at Global and Local level according to linguistic literature
  - Global coherence: maintain a cohesive representation of main topic
  - Local coherence: interrelation of information within unit

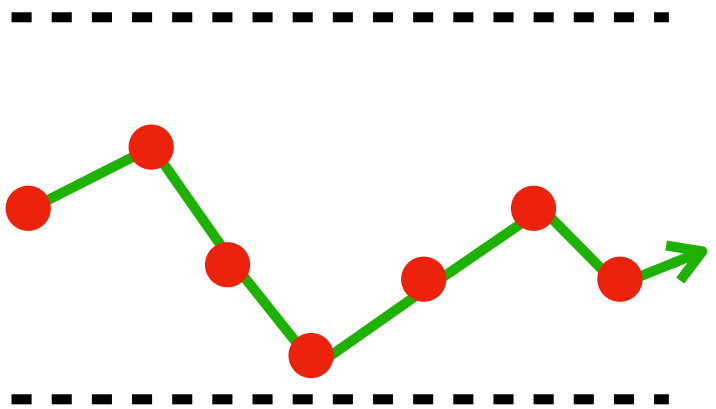
Our work

- Previous neural coherence modeling work relies on static representations and follows closely with linguistic features such as entity overlap.
- Instead, we hypothesis there exists a **sequential** and **cohesive** relationship among sentences in a latent space that represents main goal or opinion of the text.
- Based on the hypothesis, we introduce a novel metric to capture global coherence of input text.

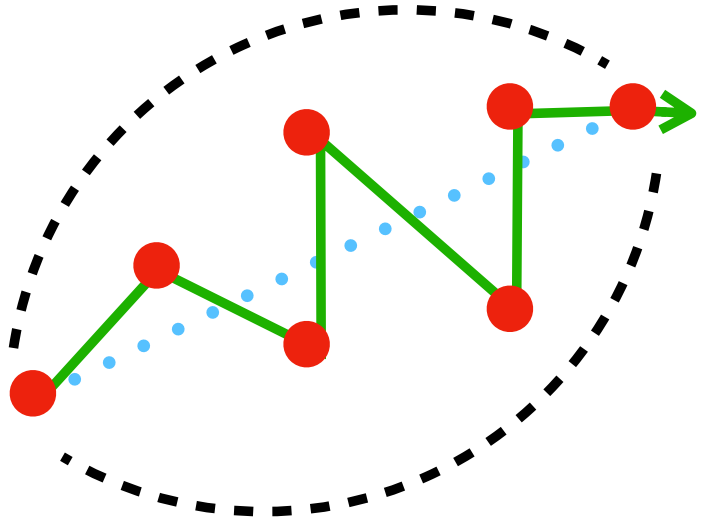
1) Main idea should be kept over the course
2) Sentences at the beginning and the end should emphasis the main idea
3) Sentences in the middle can depart from the main idea a little bit but still under control



A. Off Topic



B. Repeated

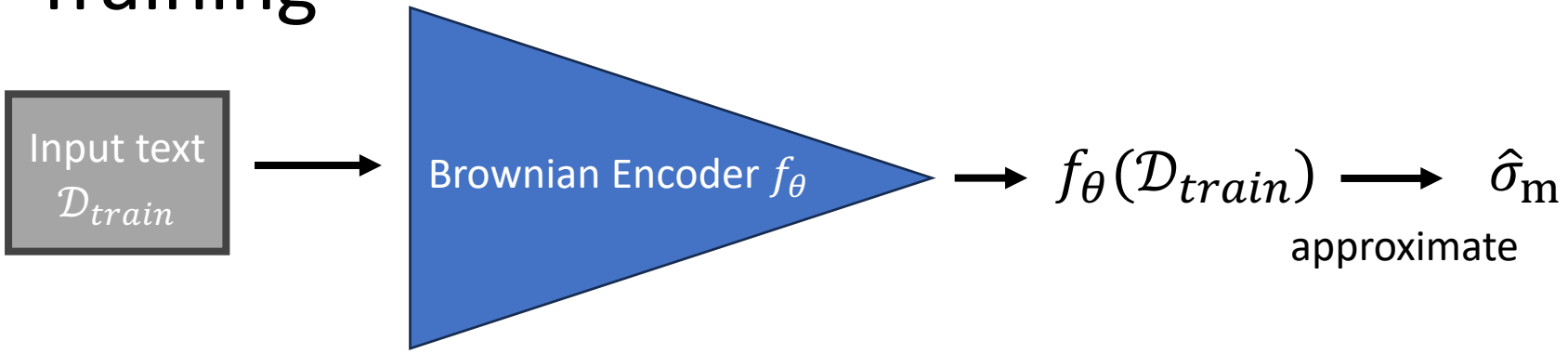


C. Coherent

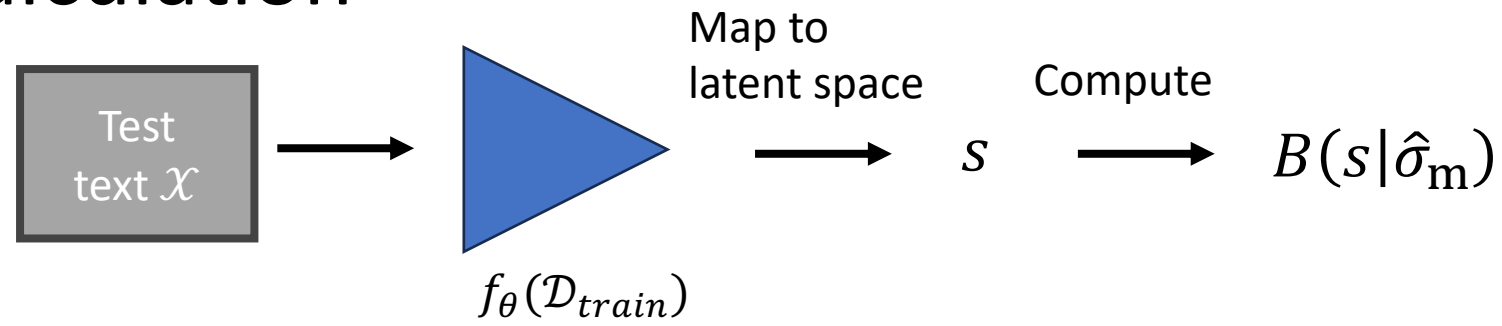
Methodology

- Given a Brownian bridge sequence $s = (s_1, s_2, \dots, s_T)$, the goal is to approximate the diffusion coefficient σ_m in the Brownian bridge formulation that we assume have coded domain-specific property.
- We can derive the MLE of $\sigma_m(s)$. Then for the training set we take the average value for all $\sigma_m(s)$.
- BBScore is defined as $B(s|\hat{\sigma}_{\mathrm{m}}) = \frac{|ln(\prod_{i=2}^{T(s)-1} \mathcal{L}_i)|}{T(s)-2}$, where $\mathcal{L}_i$ is the likelihood function for sentence i and contains $\hat{\sigma}_{\mathrm{m}}$. $T(s)$ is the length of sequence s.

1. Encoder Training



2. Score Calculation



Research Question

- Q1. Can BBScore capture both global and local coherence in synthetic settings?
- Q2. Can BBScore detects text where the departure from desired coherence are not manipulated? And can BBScore respect the difference?

Data

Wikisection (2165 articles in training ; 658 articles in test)

Task

RQ1

Artificial Task

1. Global Discrimination
2. Local Discrimination

RQ2

Downstream Task

1. AI-Human Differentiation
2. LLM detection

Results - Artificial Tasks

Table 2 & 3 in the paper

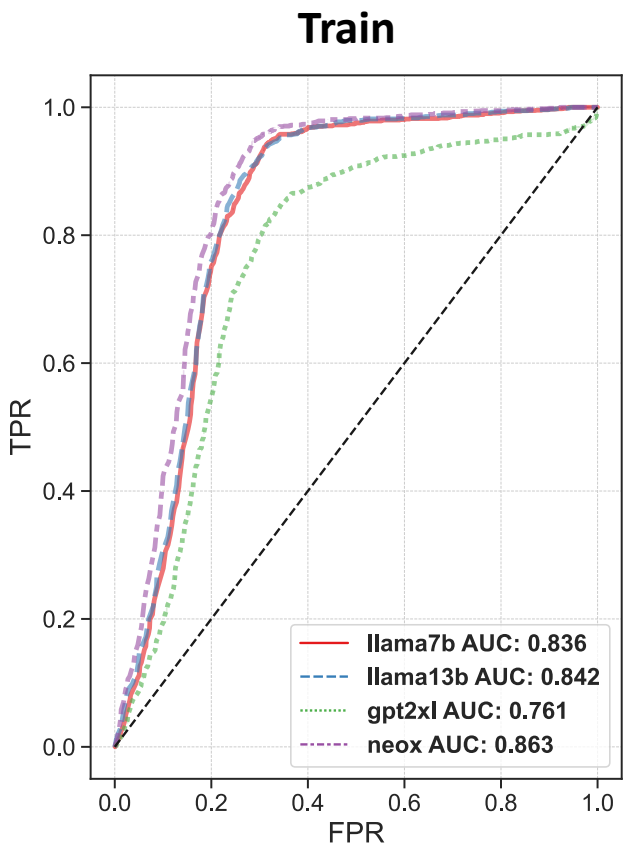
Methods	Train				Test			
	$\mathcal{D}_{b=1}$	$\mathcal{D}_{b=2}$	$\mathcal{D}_{b=5}$	$\mathcal{D}_{b=10}$	$\mathcal{D}_{b=1}$	$\mathcal{D}_{b=2}$	$\mathcal{D}_{b=5}$	$\mathcal{D}_{b=10}$
EntityGrid (Barzilay and Lapata 2008)	79.17	86.20	74.51	60.70	85.73	82.79	75.81	64.65
UnifiedCoherence (Moon et al. 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**

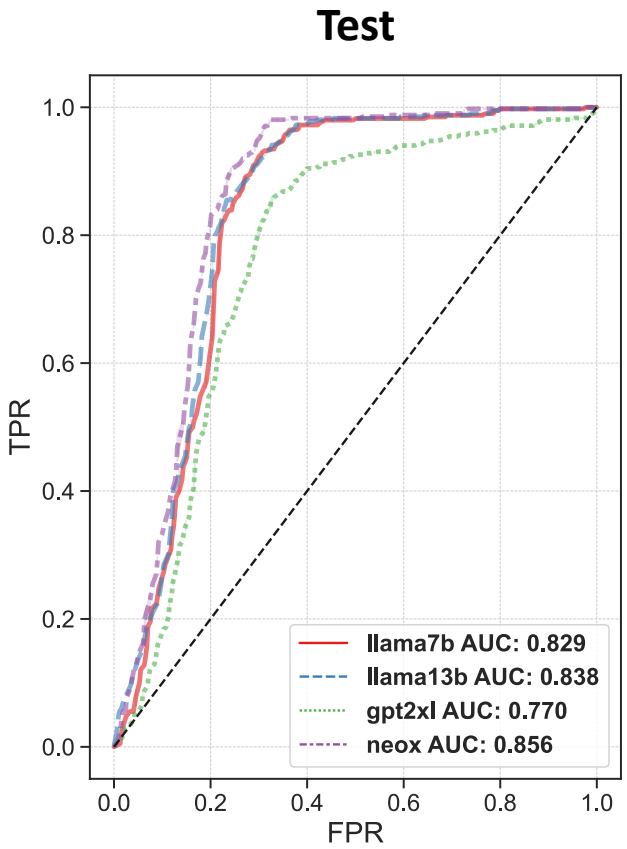
Table 2: Global Discrimination Task Results on WikiSection

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.

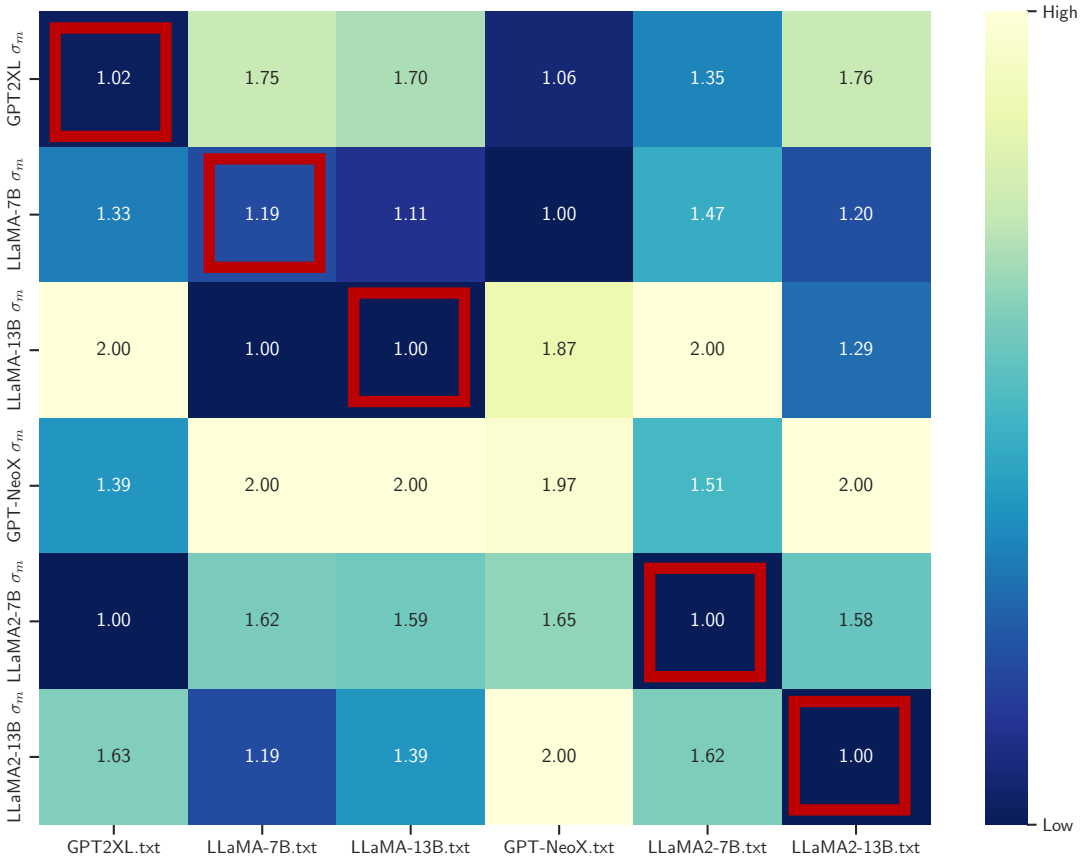
Results - AI-Human Differentiation





ROC Curve for human written text classification

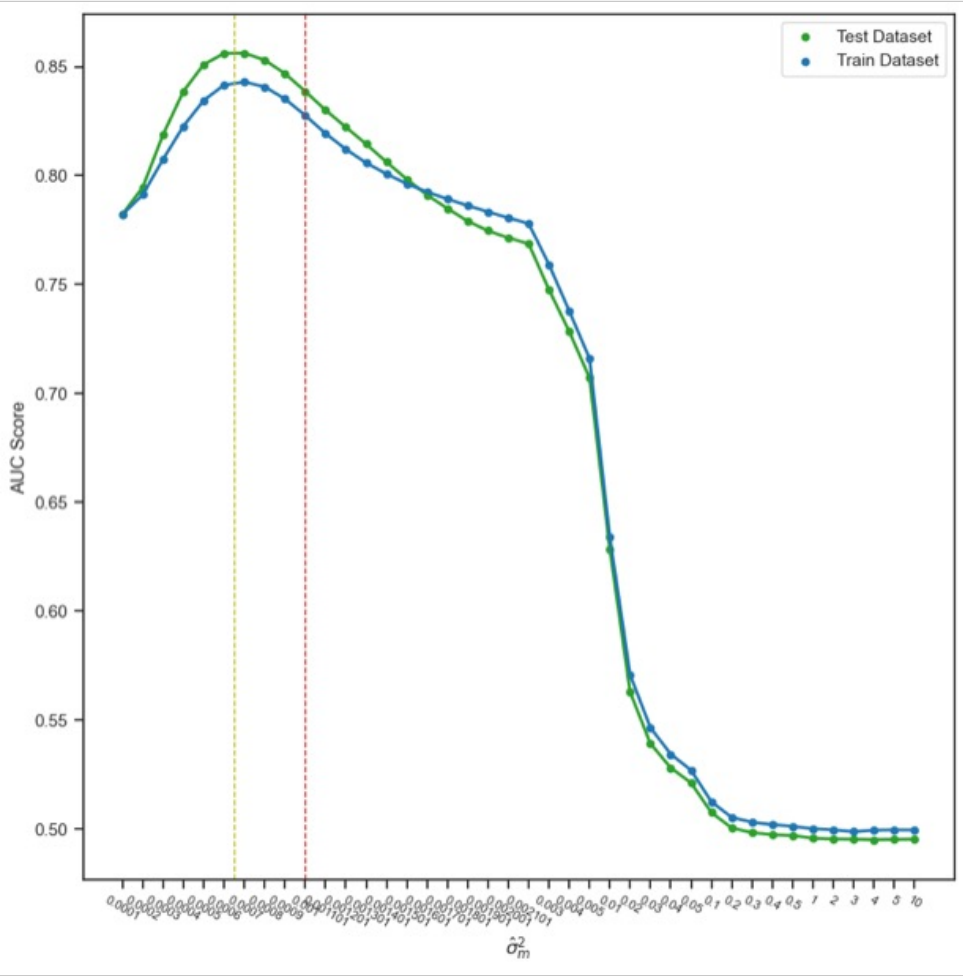
Results - LLM Detection



Normalized Wasserstein distance

Each column represent the text are generated by a specific large language model and each row represents the estimated σ_m for each model

Sensitivity Check



Performance in the block-1 global discrimination task under different $\hat{\sigma}_{\mathrm{m}}$. Olive line indicates the $\hat{\sigma}_{\mathrm{m}}$ value that yields the best performance possible; Red line indicates the $\hat{\sigma}_{\mathrm{m}}$ estimated from training data.

Take away

- Comparing to previous methods, BBScore presents a more flexible way to model text coherence.
- BBScore relies on joint likelihood function and the estimation of $\hat{\sigma}_{\mathrm{m}}$.
- BBScore shows excellent performance on global discrimination task (by design) and is also able to capture local text permutation.
- BBScore can be leveraged to discriminate generated text with Human written ones and even identify the written styles of different text generation processes (i.e. LLMs) under the same domain.