CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Low Resource With Contrastive Learning

西安交通大學 XI'AN JIAOTONG UNIVERSITY

Queen Mary University of London

Xiaoming Liu [1,†] Zhaohan Zhang [1,2,†] Yichen Wang [1,†] Hang Pu [1] Yu Lan [1] Chao Shen [1]

[1]Xi'an Jiaotong University [2]Queen Mary University of London [†]Equal Contribution

Motivation and Introduction

The emergence of Large Language Models (LLMs) brings broad concern about the malicious usage of machine-generated text (MGT). Effective MGT detectors are urgently needed.

Defects on existing detectors:

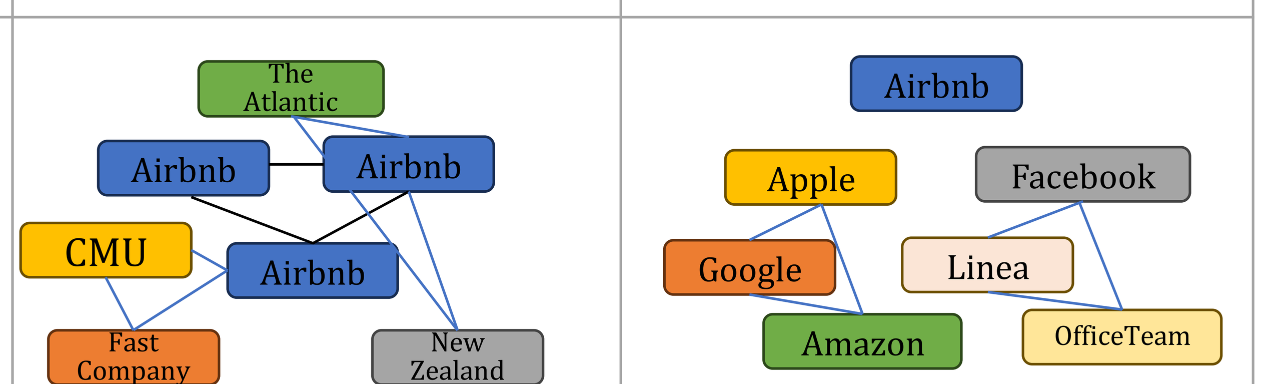
- Treat input documents as flat sequences of tokens while ignoring high-level linguistic representation of text structure
- Performance constrained by the amount of available annotated data

Our contributions:

- We model the text coherence with entity consistency and sentence interaction while statistically proving its distinctiveness in MGT detection, and we further introduce the linguistic feature at the input stage
- We introduce contrastive learning and improved contrastive loss into the MGT detector to alleviate data dependence
- We surprisingly find that MGTs originated from up-to-date language models could be easier to detect than those from previous models in our experiments

How to find hidden cameras in your Airbnb, and anywhere else		
	Human-Written Text (HWT)	**Machine-Generated Text (MGT)**
Sequence	**S1**: In recent months, there's been a number of alarming reports of Airbnb hosts installing hidden cameras in their properties but not disclosing them to the guests staying there. **S2**: Back in January, Fast Company reported on a computer science professor at Carnegie Mellon University who discovered two hidden cameras recording him and his family in an Airbnb. **S3**: And just last month, The Atlantic reported on a New Zealand family who was renting an Airbnb in Ireland and found they were being live-streamed from a hidden security camera.	**S1**: Anyone who finds a video of someone on Airbnb will probably fall under the new category of hidden cameras, which can be found only in a large part of every Airbnb listing, and you're never alone. **S2**: Apple, Google, and Amazon combined to find the most hidden camera listings in December 2018. **S3**: The electronics giant's Facebook, the mapping app and the mobile messaging company Linea formed an OfficeTeam unit that can find the video even if someone's not using them, and can track real-time activity. ...

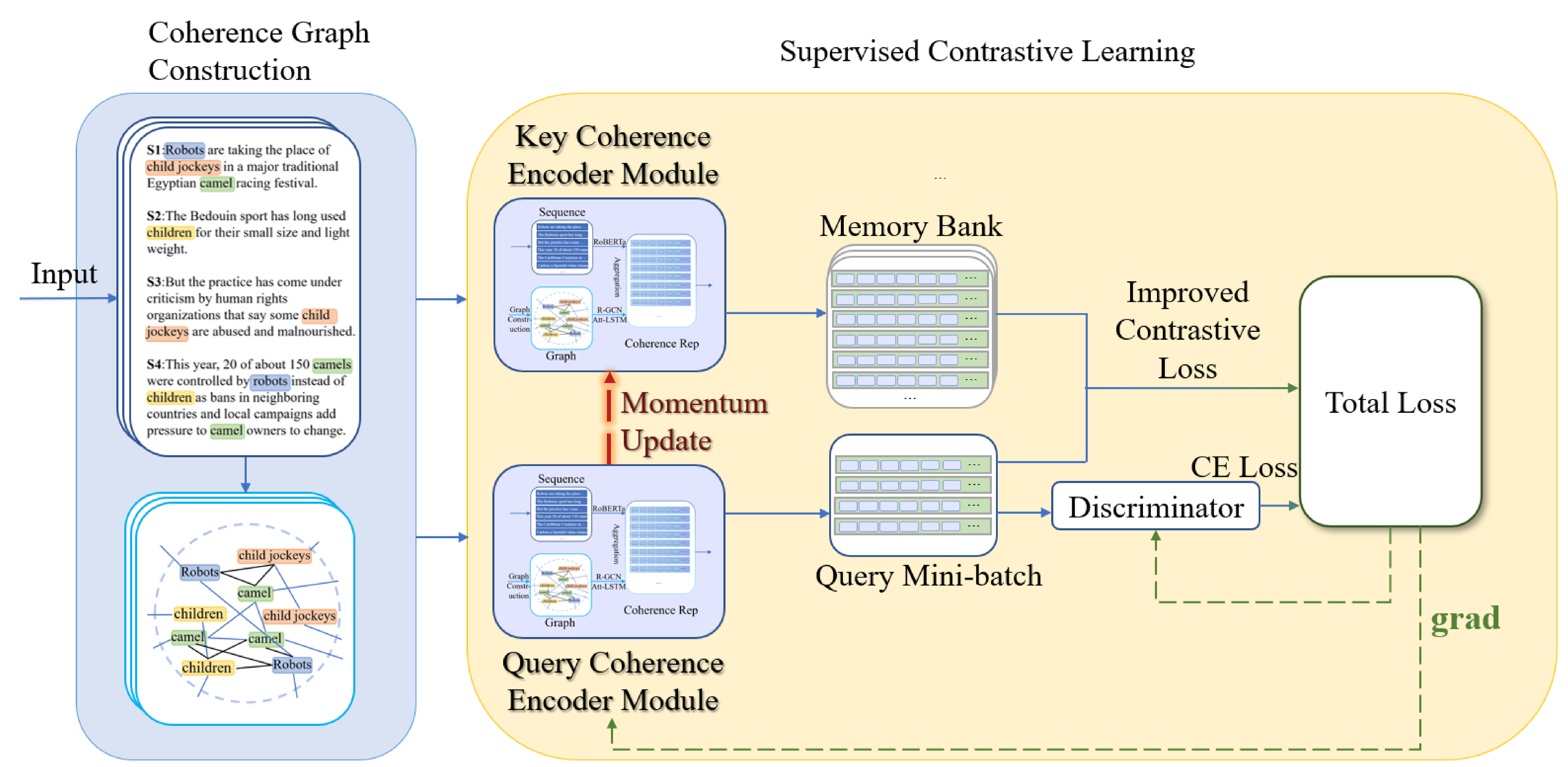
Graph



Coherence Modeling based on Centering Theory[1]

"Coherence of texts could be modeled by sentence interaction around center entities." We build a coherence graph, treat entities as nodes and co-occurrence relationship of entities as edges.

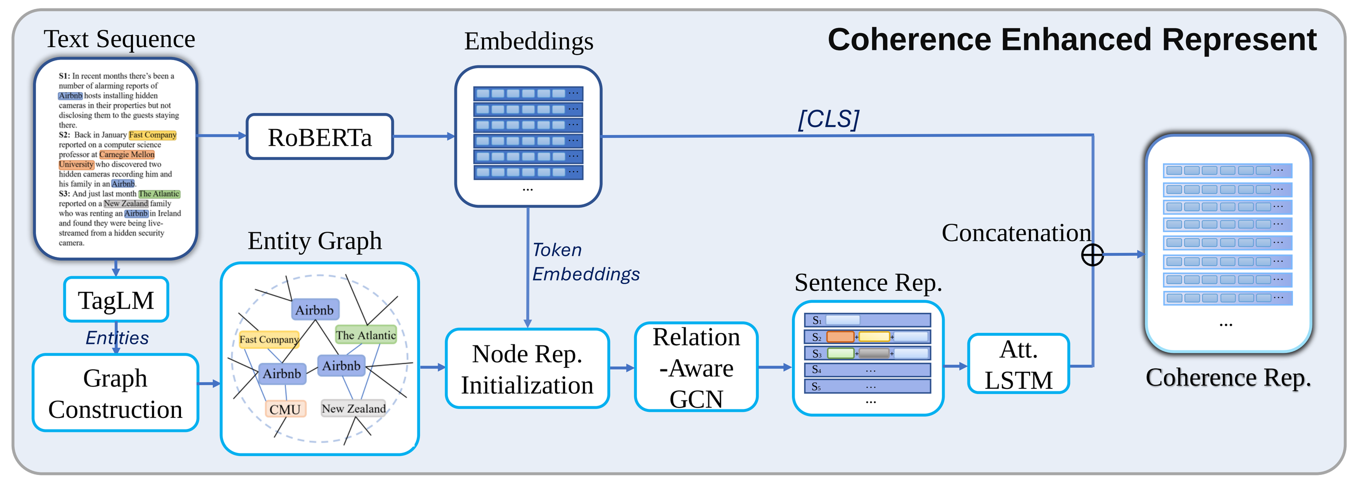
CoCo Methodology



CoCo consists of two complements: Coherence Graph Construction and Contrastive Learning.

Encoder Design: Coherence Graph Construction

We propose an innovative coherence encoder module (CEM), which is utilized to integrate coherence information into a semantic representation of text by propagating and aggregating information from different granularity via graph, to encode a coherence enhanced representation.



Toward Low Resource Scenario: Contrastive Learning

To instance compactness and class separability in low-resource settings, we utilize MOCO[2] as backbone and come up with an improved contrastive loss (ICL) for dynamically adjusting the weight of negative pair similarity according to the hardness of negative samples.

$$\mathcal{L}_{\text{ICL}} = \sum_{j=1}^{M} \mathbb{1}_{y_i=y_j} \log \frac{S_{ij}}{\sum_{p\in\mathcal{P}(i)} S_{ip} + \sum_{n\in\mathcal{N}(i)} rf_{in} S_{in}}, \qquad (1)$$

$$S_{ij} = \exp(D_q^i D_k^j/\tau), rf_{ij} = \beta \frac{D_q^i D_k^n}{\mathrm{avg}(D_q^i D_k^{1:|\mathcal{N}(i)|})},$$

where $\mathcal{P}(i)$ is the positive set in which data has the same label with q_i and $\mathcal{N}(i)$ is the negative set in which data has a different label from q_i. D_k is the key module representations and D_q is the query module representations.

Experiment and Analysis on Comparison, Ablation and Robustness

We conduct our main experiments on two public datasets and two self-constructed GPT-3.5 datasets[3], against seven baselines and SOTA. Also, an ablation study and robustness test are implemented. More additional experiments are in the paper. Here are some **key findings:**

- CoCo surpasses the state-of-the-art methods in MGT detection in both settings
- GROVER Dataset is the hardest to detect while GPT-3.5 datasets are surprisingly easy
- Coherence graph and contrastive learning module both contribute orthogonally
- CoCo shows comparable robustness to perturbations to some extent

Dataset	GROVER				GPT-2			
Size	Limited Dataset (500 examples)		Full Dataset		Limited Dataset (500 examples)		Full Dataset	
Metric	ACC	F1	ACC	F1	ACC	F1	ACC	F1
GPT2	0.5747 ± 0.0217	0.4394 ± 0.0346	0.8274 ± 0.0091	0.8003 ± 0.0141	0.5380 ± 0.0067	0.4734 ± 0.0182	0.8913 ± 0.0066	0.8839 ± 0.0078
XLNet	0.5660 ± 0.0265	0.4707 ± 0.0402	0.8156 ± 0.0079	0.7493 ± 0.0073	0.6551 ± 0.0083	0.5715 ± 0.0095	0.9091 ± 0.0091	0.9027 ± 0.0111
RoBERTa	0.6621 ± 0.0133	0.5895 ± 0.0231	0.8772 ± 0.0029	0.8171 ± 0.0048	0.8223 ± 0.0088	0.7978 ± 0.0085	0.9402 ± 0.0039	0.9384 ± 0.0044
DualCL	*0.5835 ± 0.0857*	*0.4628 ± 0.1076*	*0.7574 ± 0.0855*	*0.6388 ± 0.1300*	*0.6039 ± 0.1367*	*0.5435 ± 0.0903*	*0.8023 ± 0.1120*	*0.8046 ± 0.1530*
CE+SCL	0.6870 ± 0.0142	0.5961 ± 0.0197	0.8782 ± 0.0044	0.8202 ± 0.0057	0.8355 ± 0.0046	0.8127 ± 0.0067	0.9408 ± 0.0006	0.9390 ± 0.0009
GLTR	0.3370	0.4935	0.6040	0.5182	0.7755	0.7639	0.7784	0.7691
DetectGPT	0.5910	0.4258	0.6142	0.5018	0.7941	0.6982	0.7939	0.7002
CoCo	**0.6993 ± 0.0119**	**0.6125 ± 0.0159**	**0.8826 ± 0.0018**	**0.8265 ± 0.0036**	**0.8530 ± 0.0019**	**0.8410 ± 0.0018**	**0.9457 ± 0.0004**	**0.9452 ± 0.0004**
Dataset	**GPT-3.5 Unmixed**				**GPT-3.5 Mixed**			
Size	Limited Dataset (500 examples)		Full Dataset		Limited Dataset (500 examples)		Full Dataset	
Metric	ACC	F1	ACC	F1	ACC	F1	ACC	F1
GPT2	0.9023 ± 0.0095	0.8920 ± 0.0073	0.9917 ± 0.0056	0.9905 ± 0.0042	0.8898 ± 0.0094	0.8914 ± 0.0084	0.9910 ± 0.0046	0.9910 ± 0.0033
XLNet	0.9107 ± 0.0068	0.9037 ± 0.0064	0.9620 ± 0.0043	0.9634 ± 0.0068	0.8925 ± 0.0106	0.8922 ± 0.0089	0.9513 ± 0.0052	0.9505 ± 0.0039
RoBERTa	0.9670 ± 0.0084	0.9681 ± 0.0077	0.9928 ± 0.0035	0.9913 ± 0.0040	0.9565 ± 0.0103	0.9583 ± 0.0092	0.9923 ± 0.0017	0.9901 ± 0.0024
CE+SCL	0.9823 ± 0.0053	0.9703 ± 0.0070	0.9944 ± 0.0023	0.9943 ± 0.0031	0.9628 ± 0.0077	0.9686 ± 0.0062	0.9932 ± 0.0017	0.9905 ± 0.0038
GLTR	0.9255	0.9287	0.9350	0.9358	0.9175	0.9181	0.9210	0.9212
DetectGPT	0.9220	0.8744	0.9245	0.8991	0.8980	0.8814	0.9113	0.9041
CoCo	**0.9889 ± 0.0044**	**0.9791 ± 0.0062**	**0.9972 ± 0.0015**	**0.9957 ± 0.0020**	**0.9701 ± 0.0069**	**0.9735 ± 0.0086**	**0.9932 ± 0.0019**	**0.9937 ± 0.0028**

Model	**ACC**	**F1**
CoCo (Plain)	0.7697	0.6428
CoCo (Sentence Nodes)	0.7733	0.6379
CoCo (Coherence)	0.7777	0.6463
CoCo (Coherence+LSTM)	0.7787	0.6471
CoCo (Coherence+LSTM+SCL)	0.7827	0.6609
CoCo	0.7843	0.6684

Model	RoBERTa		CoCo	
Metric	Acc	F1	Acc	F1
Original	0.6635	0.5901	**0.6993**	**0.6125**
Delete	0.5736 (-0.0899)	0.5545 (**-0.0356**)	**0.6363 (-0.0630)**	**0.5703** (-0.0422)
Repeat	0.6320 (-0.0315)	0.5743 (-0.0158)	**0.6732 (-0.0261)**	**0.6004 (-0.0121)**
Insert	**0.6325 (-0.0310)**	0.4881 (**-0.1020**)	0.6286 (-0.0707)	**0.4970** (-0.1155)
Replace	0.5554 (-0.1081)	0.4814 (**-0.1087**)	**0.6367 (-0.0626)**	**0.5023** (-0.1102)
Average	0.5984 (-0.0651)	0.5246 (**-0.0655**)	**0.6437 (-0.0556)**	**0.5425** (-0.0700)

Preliminary Explore on the Detectable Feature in GPT-3.5 Dataset

We probe the statistical interpretation behind the GPT-3.5 dataset and try to answer the question: *Why the MGTs by GPT-3.5 are relatively easy to detect?* We count the N-gram coverage of the supporters in Transformers-Interpret and the token coverage from the Statistic Cue.

N-gram Coverage	MGT	HWT
γ_1	0.6659	0.6377
γ_2	0.4250	0.3630
γ_3	0.2883	0.2076
γ_4	0.2019	0.1372
γ_5	0.1425	0.0935

Token	Productivity	Coverage
according	0.6923	0.3126
where	0.6842	0.1998
they	0.6316	0.3837

- More consecutive spans of tokens act as an indicator for MGT than HWT
- No existing vulnerability in our dataset since trade-off between productivity and coverage
- The Easy-to-detect nature of GPT-3.5 texts might originate from language patterns

References

[1] Grosz B J, Sidner C L. Attention, intentions, and the structure of discourse[J]. Computational linguistics, 1986.

[2] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.

[3] CoCo GPT-3.5 Machine-Generated Text Datasets,https://huggingface.co/datasets/ZachW/MGTDetect_CoCo