CoUDA: Coherence Evaluation via Unified Data Augmentation

Dawei Zhu [*,η,δ] **Wenhao Wu** [*,η,δ] **Yifan Song** [η,δ] **Fangwei Zhu** [η,δ]
Ziqiang Cao [π] **Sujian Li** [η,δ,λ]
[η] School of Computer Science, Peking University
[δ] National Key Laboratory for Multimedia Information Processing, Peking University
[π] Institute of Artificial Intelligence, Soochow University
[λ] Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University

Abstract

Coherence evaluation aims to assess the organization and structure of a discourse, which remains challenging even in the era of large language models. Due to the scarcity of annotated data, data augmentation is commonly used for training coherence evaluation models. However, previous augmentations for this task primarily rely on heuristic rules, lacking designing criteria as guidance. In this paper, we take inspiration from linguistic theory of discourse structure, and propose a data augmentation framework named CoUDA. CoUDA breaks down discourse coherence into global and local aspects, and designs augmentation strategies for both aspects, respectively. Especially for local coherence, we propose a novel generative strategy for constructing augmentation samples, which involves post-pretraining a generative model and applying two controlling mechanisms to control the difficulty of generated samples. During inference, CoUDA also jointly evaluates both global and local aspects to comprehensively assess the overall coherence of a discourse. Extensive experiments in coherence evaluation show that, with only 233M parameters, CoUDA achieves state-of-the-art performance in both pointwise scoring and pairwise ranking tasks, even surpassing recent GPT-3.5 and GPT-4 based metrics. [1]

1 Introduction

Coherence is a vital aspect of communication that evaluates the structure and organization of discourse (Halliday and Hasan, 1976; Grosz and Sidner, 1986). Consequently, models capable of evaluating coherence of the given text are widely applicable in both discourse generation and assessment. While recent large language models show strong performance in various tasks (Brown et al., 2020), they have not presented superiority in coherence evaluation compared with the fine-tuning based models (Fu et al., 2023). Considering both computational efficiency and evaluation performance a good evaluation metric should possess, in this paper, we focus on modeling coherence via a fine-tuning based lightweight model.

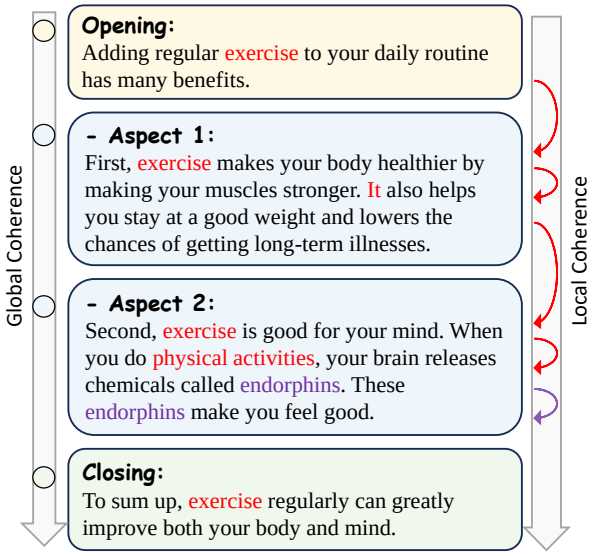

Figure 1: Example for global coherence and local coherence in a discourse. Globally, the discourse is well-structured, with a opening sentence to introduce the argument, five sentences to give evidence from two aspects, and a closing sentence for conclusion. Locally, the focused items, which is denoted in Red and Purple, transfers smoothly from sentence to sentence.

Due to the scarcity of human-annotated data, data augmentation techniques are commonly employed in training a coherence evaluation model (Li and Jurafsky, 2017; Jwalapuram et al., 2022). As human-written discourses naturally possess coherence and can thus serve as positive samples, previous research has focused on constructing negative samples, primarily through rule-based methods such as swapping or shuffling sentences (Barzilay and Lapata, 2008; Shen et al., 2021; Jwalapuram et al., 2022). However, as these methods

[*] Dawei Zhu and Wenhao Wu contribute equally to this paper. Prof. Sujian Li is the corresponding author.
[1] https://github.com/dwzhu-pku/CoUDA

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 967–978
June 16-21, 2024

are heuristically inspired without any design criteria as guidance, they suffer from weak correlation with human judgements (Mohiuddin et al., 2021). This brings up the research question: To effectively model coherence, can we find reasonable criterium as guidance to design augmentation strategies?

According to Grosz and Sidner (1986), discourse coherence is mainly determined by two aspects: the organization of discourse segments (i.e. *global coherence*), and the transition of attention or focused items (i.e. *local coherence*). Examples for these two aspects of coherence are presented in Figure 1. This inspires us to the designing criteria that a good data augmentation strategy should uniformly cover these two aspects of coherence. Following the criteria, we propose a **Co**herence evaluation framework via **U**nified **D**ata **A**ugmentation, namely **CoUDA**, which unifies both global and local aspects of coherence throughout training and inference phase.

CoUDA involves global and local augmentation to capture the corresponding aspects of coherence. Regarding global augmentation, we construct negative samples through shuffling, which disrupts the original order of the sentences to induce global incoherence. For local augmentation, our target is to construct negative samples that contain sentences incoherent with the context. While prior rule-based methods, such as swapping a sentence with another from a different text (Shen et al., 2021), can also introduce local incoherence, their constructed samples often lack diversity and complexity, potentially failing to capture nuanced aspects of local coherence. To address this, we propose a novel generative augmentation strategy that involves post-pretraining a generative model, and applying two controlling mechanisms to manipulate the difficulty of generated samples. By sampling from a generative model, and applying difficulty control, we construct high-quality negative samples to disrupt local coherence. Finally, in inference phase, we design a unified scoring strategy to incorporate both aspects of coherence for overall assessment.

While previous research on coherence evaluation has traditionally adhered to a pairwise ranking setup, we have pioneered a pointwise coherence scoring setting that we believe is more relevant in real-world scenarios. On SummEval(Fabbri et al., 2021), our CoUDA exhibits remarkable improvements in pointwise scoring compared to prior methods, including GPT-4-based metrics. Despite not being specifically tailored for pairwise ranking, our model outperforms previous ranking models on both the INSTED-CNN and INSTED-WIKI datasets (Shen et al., 2021). Furthermore, CoUDA is a lightweight model with only 233M parameters. To sum up, our contributions are as follows:

- We propose CoUDA, a data augmentation framework inspired by linguistic theory of discourse structure, which uniformly models both global and local coherence aspects of a discourse.
- We propose a novel generative augmentation strategy, which utilizes the power of the pretrained language model via post-pretraining and two mechanisms for sample difficulty control.
- Comprehensive experiments in coherence evaluation show CoUDA with only 233M parameters achieves SOTA performance, even surpassing GPT-3.5 and GPT-4 based metrics.

2 CoUDA Framework

In this section, we introduce our CoUDA framework, as illustrated in Figure 2. First, we use global and local augmentation to create negative samples that have relatively poor global coherence and local coherence, respectively. To be specific, we use sentence shuffling for global augmentation, and design a generative strategy for local augmentation. Our generative strategy involves post-pretraining a generative model, and applying two controlling mechanisms to control the difficulty of generated samples. Then we combine the constructed negative samples with the original discourses, which serves as positive samples, to train our metric model for coherence/incoherence classification. In inference phase, we utilize a unified scoring strategy to incorporate global and local coherence for overall assessment.

2.1 Preliminaries

Task Formulation. Given a discourse that contains multiple sentences $D = \{s_1, s_2, \ldots, s_n\}$, the goal of a coherence evaluator f_θ is to assess its degree of coherence by a logit score $f_\theta(D) \in [0, 1]$ (the higher the better). Ideally, $f_\theta(D) = 1$ represents that D is perfectly coherent, while $f_\theta(D) = 0$ indicates the opposite. Different from previous work that additionally relies on references (Zhao et al., 2022) or source inputs (Zhong et al., 2022), we evaluate coherence in this more concise framework that solely takes the discourse as the input.

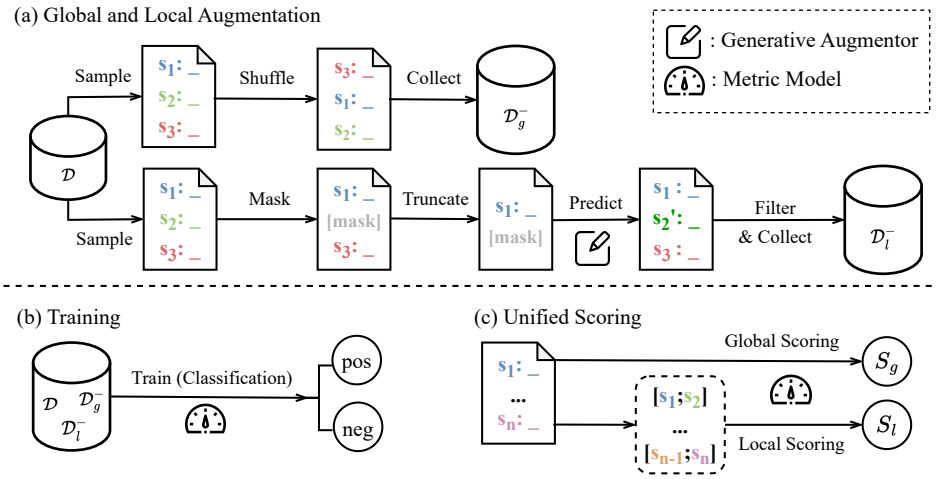

Figure 2: Overview of our proposed CoUDA framework. (a): First, we use global and local augmentation to create negative samples $\mathcal{D}_g^-$ and $\mathcal{D}_l^-$, respectively. (b): Then, we combine $\mathcal{D}_g^-$ and $\mathcal{D}_l^-$ with the original discourses $\mathcal{D}$ to train our metric model via coherence/incoherence classification. (c): In inference phase, our metric model scores the whole discourse for global score S_g, and scores each consecutive sentence pairs for local score S_l. S_g and S_l are combined to produce the final coherence score.

That is more appropriate for evaluation as coherence is an intrinsic quality of a discourse.

Data Augmentation. Data augmentation aims to artificially create additional training samples by manipulating existing data. For a discriminative setting, we need both positive and negative samples for training. In terms of coherence evaluation, since a natural discourse D is intrinsically coherent, we focus on applying data augmentation to construct negative samples, i.e. incoherent samples D^-. Afterwards, the created incoherent discourses D^- and the original discourse D respectively serve as negative and positive samples to train f_θ.

In the following, we introduce how we construct our two types of negative samples via *global augmentation* and *local augmentation* in details.

2.2 Global Augmentation

To construct samples that have relatively poor global coherence, we disrupt the original appropriate organization of sentences in D. Concretely, we shuffle the order of sentences in D to effectively disrupt its global coherence. As illustrated in Figure 2(a), by shuffling $D = \{s_1, s_2, s_3\}$, we can construct a negative sample $D_g^- = \{s_3, s_1, s_2\}$.

2.3 Local Augmentation

Local augmentation aims to construct samples with relatively poor local coherence using the original discourse D. Intuitively, we can realize it by replacing a sentence $s_k \in D$ with a substitute s'_k that is incoherent with the leftover discourse $D\backslash s_k$. This is based on the insight that, through such a replacement, s'_k will decrease local coherence of the discourse by introducing an incoherent transition of attention between sentences.

Subsequently, the most important question is how to find a suitable s'_k in practice. However, most prior studies introduce such incoherent elements via heuristic rules, resulting in s'_k that has **very weak relevance or even irrelevant** with the remaining discourse $D\backslash s_k$. For example, INSteD (Shen et al., 2021) obtains s'_k by extracting sentence of the highest n-gram overlap with s_k from another discourse. As a result, their introduced local augmentation samples are too easy to train a powerful coherence evaluator.

To construct samples with a higher level of local incoherence, we propose to construct s'_k in a generative way. Specifically, we train a generative augmentor G to reconstruct s_k based on $D\backslash s_k$ and use its generated sentence $s'_k \sim G(s_k|D\backslash s_k)$ to replace s_k. The strong performance of pretrained generation model will ensure s'_k to meet the basic standard of fluency and relevance with regard to $D\backslash s_k$. Meanwhile, due to the intrinsic limitation of autoregressive generation, the reconstructed s'_k will frequently be incoherent with $D\backslash s_k$, making it possible to construct negative samples in a generative way. To further ensure that s'_k conveys the local incoherence we expect, we design two controlling mechanisms during the inference of G.

These two mechanisms, *context truncation* and *coherence filtering* , constraints s'_k to be neither too strong (perfectly coherent with $D \backslash s_k$) nor too easy (the incoherence that is too obvious). Overall, by replacing s_k with s'_k, we construct a much stronger negative sample, which conveys high-level local incoherence while maintaining the basic relevance and fluency with $D \backslash s_k$. In the following, we will introduce our *generative augmentor*, *context truncation* and *coherence filtering* in details.

Generative Augmentor. Given discourse D, we uniformly sample s_k from D's non-opening and non-closing sentences. Next, we train a text generation model G by learning to reconstruct s_k based on the leftover discourse $D \backslash s_k$. Following recent popular text generation paradigm, this can be done by selecting G as a transformer-based sequence-to-sequence model and maximizing the likelihood of $G(s|D \backslash s_k)$ autoregressively.

We also notice that Gap Sentences Generation (GSG), the pretraining task of PEGASUS (Zhang et al., 2020), takes the similar form of reconstructing sentences. But we cannot directly apply PEGASUS as G, because GSG is specially designed for summarization, which requires predicting multiple salient sentences in the discourse. By contrast, our sentence reconstruction task aims to capture the coherence relation between an arbitrary sentence s and the leftover discourse. In practice, we also find s'_k generated by PEGASUS often serve as summaries of the leftover discourse, rather than being coherent with it. Thus, instead of directly applying PEGASUS, we leverage this similarity of tasks and use PEGASUS for initialization. In this way, we inherit the effectiveness of pretrained model.

After the generative augmentor is trained, we use it to predict s'_k with two controlling mechanisms:

Context Truncation. Due to the strong generation ability of generative augmentor, s'_k may be highly coherent with $D \backslash s_k$, which is not the negative sample we expect. To ensure s'_k to convey the local incoherence, we develop a context truncation mechanism to restrict the model's generation to only partially coherent with the context. Specifically, given $D \backslash s_k = \{s_1, s_2, ..., [mask], ..., s_n\}$ with s_k masked, we randomly choose to truncate the context before or after the mask token, i.e., the input for our generative augmentor is either $\{s_1, s_2, ..., [mask]\}$ or $\{[mask], ..., s_n\}$. Take the former as an example, without information from subsequent text, the model is only able to generate predictions that are coherent with preceding text.

Coherence Filtering. In addition to context truncation, we also perform coherence filtering to remove negative samples that are too easy. We utilize UniEval (Zhong et al., 2022) to score the coherence of each sample and eliminates samples with coherence scores below a filtering threshold δ.

2.4 Training and Unified Scoring

Training. We combine the original discourses with negative samples constructed via global and local augmentation to train our metric model, as illustrated in Figure 2(b). We utilize the classification setup, based on findings (Steen and Markert, 2022) that indicate its superior performance in coherence evaluation and downstream tasks, as opposed to the commonly used pairwise ranking setup. Specifically, we train our metric model to distinguish each sample as coherent or incoherent through binary cross entropy loss. For implementation details, please refer to Appendix A.

Unified Scoring. For a comprehensive evaluation of discourse coherence, our CoUDA further includes a unified scoring strategy, as presented in Figure 2(c). Specifically, our model first assigns a score conditioned on the whole discourse to represent its global coherence level:

$$S_g = f_\theta(D) \tag{1}$$

Then, since global scoring may fail to effectively capture the fine-grained coherence between sentences, we extract consecutive sentence pairs $[s_i; s_{i+1}]$ from the discourse and have our model evaluate the inter-sentential coherence S_l^i of each pairs, where $1 \leq i \leq n-1$:

$$S_l^i = f_\theta([s_i; s_{i+1}]) \tag{2}$$

Notably, although our model is trained for scoring the whole discourse, rather than sentence pairs, the training data includes discourse samples with only two sentences. As a result, our model can generalize to scoring sentence pairs as well. Afterwards, we obtain local coherence score for discourse D by averaging each sentence pair's coherence score:

$$S_l = \text{Average}(\{S_l^1, ..., S_l^{n-1}\}) \tag{3}$$

The global and local scores are then combined via interpolation to form the overall coherence score:

$$\text{Score} = (1-\lambda) \cdot S_g + \lambda \cdot S_l \tag{4}$$

where $\lambda \in [0, 1]$ controls the weight. This unified design also aligns with the coherence rating process of human readers, who consider both discourse organization as a whole, and smooth transitions of focused items between adjacent sentences.

3 Experimental Setup

3.1 Evaluation Tasks

We perform meta-evaluation on the proposed metric model in two task settings, i.e. pointwise scoring and pairwise ranking.

Pointwise Scoring involves assigning coherence scores to text summarization samples and evaluating the correlation between model-assigned scores and human-rated scores. This task closely simulates real-world scenarios. To determine the accuracy of the assigned scores, we compute the correlation coefficients between the model-generated scores and human ratings using *Spearman* (Sedgwick, 2014), *Pearson* (Sedgwick, 2012), and *Kendall's tau* (Abdi, 2007). Following previous work, these correlation scores are reported at both *sample-level* and *dataset-level* (See Appendix A for their definitions).

Pairwise ranking requires the metric models to determine the more coherent option when presented with two candidates. This task serves as an alternative when absolute scores are unavailable, relying solely on relative coherence rankings. For this task, we use accuracy as performance metric.

3.2 Evaluation Datasets

For pointwise scoring, we evaluate model performance on SUMMEVAL (Fabbri et al., 2021), which is a meta-evaluation benchmark for summarization that contains 100 articles with summaries generated by 16 different systems. For each summary, it offers human annotated scores in terms of fluency, coherence, consistency, and relevance.[2]

In pairwise ranking, we evaluate model performance on INSTED (Shen et al., 2021), which is an intruder sentence detection dataset constructed using discourses from CNN and Wikipedia. We denote these two parts as INSTED-CNN and INSTED-WIKI. In this dataset, incoherent discourses are created by randomly substituting a sentence with another one selected using n-gram overlap from different discourses.

[2]In this paper, we focus on discourse coherence, so we neglect coherence evaluation datasets on dialogue.

3.3 Baselines Models

Though more applicable in real scenarios, few work in coherence evaluation has pioneered in pointwise scoring. For a comprehensive performance comparison, we include baselines models from three categories: **1) Pairwise Coherence Evaluators**: UNC (Moon et al., 2019) and MULTINEG (Jwalapuram et al., 2022). UNC captures different levels of coherence via a LSTM-based Siamese architecture; MULTINEG[3] mines hard negative samples constructed via sentence shuffling to train pairwise coherence ranking models. 2) **General Evaluators**: BARTSCORE (Yuan et al., 2021), UNIEVAL (Zhong et al., 2022). BARTSCORE treats text evaluation as a generation task, utilizing BART to assign quality scores for a specific dimension. UNIEVAL reframes text evaluation as a Boolean Question Answering task. Backboned with T5, it is trained with rule-based local augmentation for coherence evaluation. **3) Large Language Models**: G-EVAL (Liu et al., 2023) uses LLMs with chain-of-thoughts to assign quality scores. We experiment with two versions using GPT-3.5-Turbo / GPT-4, respectively denoted as G-EVAL-3.5 / 4. We include more details about using UNIEVAL and G-EVAL in Appendix C and D, respectively.

3.4 Details of Synthetic Data

Data Source. We obtain positive part of data for our framework by sampling from CNN (Nallapati et al., 2016) and Wikipedia (Yang et al., 2015). For CNN, we utilize its source documents rather than summaries, because the latter is constructed by combining bullet points, hence lacks coherence. For each source document, we randomly select 2 to 5 leading sentences, enabling our metric model to generalize to different lengths. The same length constraint is applied on Wikipedia as well. Concretely, we sample 10,000 documents each from CNN and Wikipedia, hence obtaining 20,000 positive samples.

Statistics. For global coherence, we perform permutation on 5,000 positive samples, and acquire 5,000 negative samples for this aspect. For local coherence, we perform gap sentence generation on the remaining 15,000 positive samples using

[3]The original MULTINEG model is backboned with XLNet and trained on the WSJ dataset. For fair comparison, we retrained this model from ALBERT-xxlarge, using the same part of Wikipedia and CNN data. Notably, due to its use of two encoders, MULTINEG has twice the number of parameters compared to COUDA.

Model	#Param. ↓	Sample-Level ↑			Dataset-Level ↑		
		ρ	r	τ	ρ	r	τ
UNC	-	18.8	27.8	14.1	19.8	24.3	14.0
MULTINEG	466M	44.6	48.1	34.0	47.7	47.8	34.3
BARTSCORE	406M	44.8	45.8	34.2	40.8	43.4	29.2
UNIEVAL	770M	56.7	57.8	43.6	58.7	55.6	42.3
G-EVAL-3.5	>10B	47.0	48.4	40.3	43.5	43.8	35.3
G-EVAL-4†	>100B	58.2	-	45.7	-	-	-
COUDA (*ours*)	233M	**60.0**	**62.1**	**46.0**	**65.6**	**64.2**	**47.8**

Table 1: Sample-level and dataset-level Spearman (ρ) / Pearson (r) / Kendall (τ) correlations with human ratings on SUMMEVAL. Best results in each column are denoted **in Bold**. † denotes results reported in the original paper. With only 233M parameters, COUDA largely outperforms previous methods, including GPT-4 based methods.

Model	CNN	Wiki
UNC	96.4	60.5
MULTINEG	94.2	72.1
BARTSCORE	70.7	58.8
UNIEVAL	92.0	77.3
G-EVAL-3.5	82.2	58.5
COUDA	**98.5**	**79.1**

Table 2: Pairwise ranking accuracy on the CNN and Wikipedia split of INSTED.

generative augmentor with context truncation. By setting threshold δ for confidence filtering to 0.5, we obtain 10,889 positive and negative pairs for this aspect. Hence, the final size of our synthetic data (including positive samples) is 31,778. We split it into 30,000 / 1,178 for training and validation.

3.5 Implementation Details.

Our metric model utilizes ALBERT (Lan et al., 2020) as the backbone, benefiting from its sentence order prediction task during pretraining to capture information flow between sentences. Specifically, we use ALBERT-xxlarge with a total of 233M parameters. We set batch size to 32 and learning rate to $1e^{-5}$. Convergence is reached within 3,000 steps. We use the best performing checkpoint on the validation part of synthetic data. Details about generative augmentor are presented in Appendix A. In terms of hyperparameters λ and δ, we simply set both of them to 0.5.

4 Results

In this section, we show that COUDA framework achieves impressive coherence evaluation results on pointwise scoring and pairwise ranking tasks, even when compared with GPT-4 based models. We report average scores across 3 runs with different random seeds.

4.1 Results on SUMMEVAL

Table 1 presents the sample-level and dataset-level correlations of each model with human ratings on SUMMEVAL. Since UNC and MULTINEG are trained through pairwise ranking, their performance on for pointwise scoring is relatively limited. BARTSCORE and UNIEVAL are general evaluators for multiple dimensions such as informativeness and coherence. The former lacks specific training for these dimensions, leading to lower performance, while the latter gain significant improvement through tailored training for coherence. However, UNIEVAL still relies on heuristic rules for augmentation, resulting in limited improvements. The third block presents the results of G-EVAL-3.5 and G-EVAL-4, built upon GPT-3.5-TURBO and GPT-4, respectively. Since there are no exact description of how many parameters GPT-3.5/4 takes, we estimate them as >10B and >100B.

Among baselines models, G-EVAL-4 achieves highest correlation with human ratings, followed by UNIEVAL, which demonstrates strong performance, even surpassing G-EVAL-3.5. Compared with UNIEVAL, COUDA consistently shows its superiority on both sample-level correlation (+3.3/+4.3/+2.4 in ρ, r, τ) and dataset-level correlation (+6.9/+8.7/+5.4 in ρ, r, τ). With only 233M parameters, it also surpasses G-EVAL-4 in both sample-level Spearman and Kendall correlations by 1.8 and 0.3 points, respectively. This remarkable improvement consolidates the efficacy of our designing criteria. Additionally, we notice that performance gain in dataset-level correlation

G	L_G	L_R	Sample-Level			Dataset-Level		
			ρ	r	τ	ρ	r	τ
✓			56.3	57.2	43.1	58.2	56.2	42.1
	✓		57.6	59.4	44.1	62.9	61.6	45.6
		✓	53.8	56.4	41.1	59.4	60.1	43.2
✓		✓	56.6	59.3	42.5	61.5	61.3	44.1
✓	✓		**60.0**	**62.1**	**46.0**	**65.6**	**64.2**	**47.8**

Table 3: Comparison of global augmentation G, our generative local augmentation L_G, previous rule-based local augmentation L_R, and their combinations.

Method	Sample-Level			Dataset-Level		
	ρ	r	τ	ρ	r	τ
Generative Augmentor	30.2	33.5	22.5	26.7	25.4	18.6
+ C.T.	51.4	51.9	38.9	53.8	51.9	38.2
+ C.T. + filter $\delta = 0.2$	53.7	51.3	41.5	55.9	47.8	40.3
+ C.T. + filter $\delta = 0.4$	52.8	54.0	40.4	56.8	53.7	40.1
+ C.T. + filter $\delta = 0.6$	**55.7**	**55.3**	**42.8**	**58.0**	**55.9**	**42.0**
+ C.T. + filter $\delta = 0.8$	46.3	42.7	35.9	49.5	41.5	35.7

Table 4: Analysis of our controlling mechanisms for local augmentation. *C.T.* stands for context truncation. δ is the threshold for confidence filtering.

is much greater than that of sample-level.

4.2 Results on INSTED

Table 2 presents each model's pairwise ranking accuracy on INSTED-WIKI and INSTED-CNN. Both MULTINEG and UNC achieves impressive accuracy. We suppose it is because they are exactly trained using the pairwise ranking setup. UNIEVAL also achieves competitive results, which means that specialized training for coherence greatly enhances model performance. Surprisingly, G-EVAL-3.5 obtains merely above chance accuracy on INSTED-WIKI, indicating that current LLMs are unreliable in pairwise ranking tasks, necessitating further investigation and attention from researchers. Our COUDA, though not directly trained under pairwise ranking settings, achieves best results on both INSTED-CNN and INSTED-WIKI, with a performance gain of 2.2 and 1.8 points, respectively.

5 Comparison of Augmentation Methods

In this section, we validate the advantage of our unified data augmentation strategy for coherence scoring over previous data augmentation strategies.

Compared Data Augmentation Methods. Coherence evaluation emphasizes the sentence structure and organization of a discourse. Due to this special focus, data augmentation strategies designed for other tasks, e.g. EDA (Wei and Zou, 2019), are not directly applicable. Instead, we compare following data augmentation strategies for generating negative samples: **1) G:** Global augmentation via sentence shuffling (Barzilay and Lapata, 2008), which is also adopted in our framework. **2) $\mathbf{L_R}$:** Rule-based local augmentation through sentence intrusion, which employs n-gram overlap to select locally incoherent samples (Shen et al., 2021). **3) $\mathbf{L_G}$:** Our generative local augmentation strategy. **4) $\mathbf{G + L_R}$ or $\mathbf{G + L_G}$:** Combination of global and local augmentation methods.

Global vs. Local vs. Unified. In Table 3, we can see that unifying global and local augmentation data yields the best human correlation, better than using global or local augmentation alone. This aligns well with the linguistic theory of discourse structure that the organization of discourse segments (global coherence), and the transition of attention or focused items (local coherence) are two key factors of discourse coherence, from which our unified data augmentation framework are inspired.

Generative vs. Rule-based. Further, we compare the result of generative augmentation vs. rule-based augmentation for modeling local coherence. First, metric model trained with L_G outperforms that of L_R by a large margin on both sample-level correlation (+3.8/+3.0/+3.0 in ρ, r, τ) and dataset-level correlation (+3.5/+1.5/+2.4 in ρ, r, τ). Second, when combined with global augmentation, $G + L_G$ yields significantly superior performance than $G + L_R$. Based on these two aspects, we can conclude that our generative strategy is more effective than rule-based methods.

6 Analysis

Unified Scoring. First, we study the effectiveness of our unified scoring strategy. Experiment results are demonstrated in Figure 3. First, both global and local scores are beneficial in improving human correlation. Additionally, global scores correlate better with human ratings than local scores.

Controlling Mechanisms. We then analyze the effect of our difficulty controlling mechanisms in local augmentation. Specifically, we train our metric model separately on local augmentation data constructed under different settings to compare

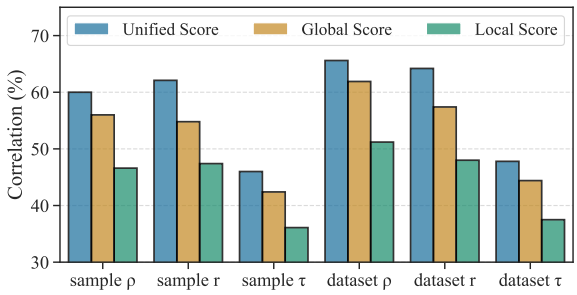

Figure 3: Ablation study of global and local scores in our unified scoring strategy.

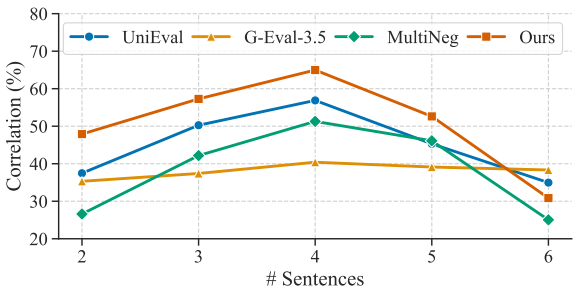

Figure 4: Average of dataset-level Spearman / Pearson / Kendall correlation on SUMMEVAL w.r.t. discourses containing different numbers of sentences.

their impacts. Table 4 presents the results. First, we can see that context truncation contributes a significant portion of performance, without which our generative augmentor suffers a severe performance drop of more than 20 points. This demonstrates the effectiveness of constructing partially coherent samples. Second, we find that our confidence filtering mechanism, through which we filter out easy negative samples, also helps model performance. We found that 0.6 is an optimal threshold that can filter out easy examples while ensuring enough amount of training data. We have also provided a case study in Appendix B.

Discourse Length. We compare our model's performance with strong baselines (MULTINEG, MULTINEG, G-EVAL-3.5) w.r.t. different discourse length. Concretely, we categorize all 1,600 system summaries of SUMMEVAL into different groups according to the sentence numbers they have. We calculate the average of dataset-level Spearman / Pearson / Kendall correlation as defined in Equation 6 for each group. Figure 4 presents the results. On average, our model achieves best results when target discourse contains no more than 5 sentences. As the discourse length increases, all models suffer from performance drop, with G-EVAL-3.5 being the only exception, which renders very steady correlation against length variance. Since each training sample we construct contains no more than 5 sentences (see Appendix A), we assume COUDA's performance drop can be alleviated by training on samples with more sentences.

7 Related Work

7.1 Coherence Evaluation

Coherence evaluation measures the organization and structure of a discourse. Due to the paucity of human-annotated training data, previous work has mainly focused on two synthetic tasks: permutation detection and sentence intrusion detection. Permutation detection task (Barzilay and Lapata, 2005; Elsner et al., 2007; Barzilay and Lapata, 2008; Li and Jurafsky, 2017) requires the model to distinguish original discourse from its sentence shuffled version. Sentence intrusion detection task (Shen et al., 2021) determines whether a discourse contains an intruder sentence from another discourse.

A series of methods have been proposed for these synthetic tasks. Barzilay and Lapata (2005, 2008) introduced the popular entity-based model using Centering Theory (Grosz et al., 1995). It was further extended to combine with entity-specific features (Elsner and Charniak, 2011), convolutional neural networks (Tien Nguyen and Joty, 2017), and graph neural networks (Mesgar et al., 2021). Jwalapuram et al. (2022) attempted to improve model generalization by training their model purely through self-supervision, with negative samples mined from permutation space. Instead, we propose to improve evaluation performance by unifying different aspects of discourse coherence, as inspired by linguistic theory of discourse structure (Grosz and Sidner, 1986). UNC (Moon et al., 2019) captured different levels of coherence via a Siamese architecture that involved bi-linear projection and lightweight convolution-pooling. By contrast, we address this from the perspective of data augmentation rather than model architecture.

7.2 General Evaluators

We denote evaluators capable of assessing multiple quality dimensions by altering input and output contents (Yuan et al., 2021), or adopting different formulas (Scialom et al., 2021; Zhong et al., 2022) as general evaluators. A leading trend is to utilize generation model for quality assess-

ment, such as BARTSCORE (Yuan et al., 2021), UNIEVAL (Zhong et al., 2022). Apart from that, DISCOSCORE (Zhao et al., 2022) compared the focus matrix between the candidate and the reference to calculate the overall quality score.

With the rise of large language models (LLMs), there has been a growing tendency to use LLMs for evaluation purpose (Wang et al., 2023a; Fu et al., 2023; Wang et al., 2023b; Liu et al., 2023). Wang et al. (2023a) adopted ChatGPT for NLG evaluation and achieved competitive results in terms of correlation with human judgments. Liu et al. (2020) used LLMs with chain-of-thought and a form-filling paradigm to assess the quality of text.

8 Conclusion

We propose a unified data augmentation framework called COUDA, with the designing criteria to unify both global and local aspects of coherence, as inspired by linguistic theory of discourse structure. This data framework includes global and local augmentation, a classification paradigm for training and a unified scoring strategy for inference. We specifically propose a novel generative augmentation strategy, which involves post-pretraining a generative model, and applying two controlling mechanisms to control the difficulty of generated samples. With only 233M parameters, our framework achieves remarkable improvement over previous methods, including GPT-4 based metrics.

Limitations

Our work is still limited in some aspects, particularly in handling extra long discourses. Note that our framework requires assigning coherence scores to all adjacent sentence pairs. While this approach allows for detailed modeling of local coherence between sentences, it may be slow when dealing with documents that contain a large number of sentences.

Ethics Statement

Our work complies with the ACL Ethics Policy. Since all datasets we used are publicly available, we have not identified any significant ethical considerations associated with our work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Key R&D Program of China (No. 2022YFC3600402).

References

Hervé Abdi. 2007. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510.

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A Unified Local and Global Model for Discourse Coherence. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 436–443, Rochester, New York.

Micha Elsner and Eugene Charniak. 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*, pages 125–129, Portland, Oregon, USA.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *arXiv:2007.12626*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. *arXiv:2302.04166*.

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

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

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman, London.

Prathyusha Jwalapuram, Shafiq Joty, and Xiang Lin. 2022. Rethinking Self-Supervision Objectives for Generalizable Coherence Modeling. In *Proceedings*

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6044–6059, Dublin, Ireland.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942*.

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.

Sennan Liu, Shuang Zeng, and Sujian Li. 2020. Evaluating Text Coherence at Sentence and Paragraph Levels. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1695–1703, Marseille, France.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv:2303.16634*.

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

Tasnim Mohiuddin, Prathyusha Jwalapuram, Xiang Lin, and Shafiq Joty. 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*, pages 3528–3539, Online.

Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Chi Xu. 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)*, pages 2262–2272, Hong Kong, China.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization Asks for Fact-based Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic.

Philip Sedgwick. 2012. Pearson's correlation coefficient. *Bmj*, 345.

Philip Sedgwick. 2014. Spearman's rank correlation coefficient. *Bmj*, 349.

Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. 2021. Evaluating Document Coherence Modeling. *Transactions of the Association for Computational Linguistics*, 9:621–640.

Julius Steen and Katja Markert. 2022. How to Find Strong Summary Coherence Measures? A Toolbox and a Comparative Study for Summary Coherence Measure Evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6035–6049, Gyeongju, Republic of Korea.

Dat Tien Nguyen and Shafiq Joty. 2017. A Neural Local Coherence Model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. *arXiv preprint arXiv:2303.04048*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large Language Models are not Fair Evaluators. *arXiv preprint arXiv:2305.17926*.

Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv:1901.11196 [cs]*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. *arXiv:2106.11520 [cs]*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339.

Wei Zhao, Michael Strube, and Steffen Eger. 2022. DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence. *arXiv preprint arXiv:2201.11176*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates.

A Details of Data, Generative Augmentor, and Correlation Calculation

Details of Generative Augmentor. Our generative augmentor is initialized with PEGASUS-Large using the checkpoint in Huggingface. We train it on the positive samples mentioned above, with batch size set to 32. Convergence is reached within 5,000 steps. To avoid data leakage in training and prediction, we split our positive samples into part A and part B, each with 20,000 samples. We first train our model solely on part A, and use it to construct negative samples for part B. Then we train a new model solely on part B, and use it to construct negative samples for part A.

Sample-Level and Dataset-Level Correlation. Suppose we have n documents in a dataset, for each document $d_i, i \in \{1, ..., n\}$, we have J system outputs. Let $o_{ij}, j \in \{1, ..., J\}$ be the output of the j^{th} system for the i^{th} document, K be a correlation measure, f_θ and f_h be metric model and human evaluation, respectively, sample-level correlation and dataset-level correlations can be calculated as follows:

(1) *Sample-level* correlation.

$$K^{sample} = \frac{1}{n}\sum_{i=1}^{n} K([f_\theta(o_{i1}), ..., f_\theta(o_{iJ})], [f_h(o_{i1}), ..., f_h(o_{iJ})]) \quad (5)$$

(2) *Dataset-level* correlation.

$$K^{dataset} = K([f_\theta(o_{11}), ..., f_\theta(o_{nJ})], [f_h(o_{11}), ..., f_h(o_{nJ})]) \quad (6)$$

B Case Study

We demonstrate an example of substitute sentences selected or generated using different methods in Table 5. RULE selects the substitute sentence through n-gram overlap, resulting in a relatively easy samples, as the selected sentence is very incoherent with the context. PEGASUS generates a sentence that summarizes the remainder, rather than being coherent with the context. The prediction of our generative augmentor is highly coherent with the context, making it difficult to be distinguished as negative. Through context truncation, we obtain a partially coherent prediction, which is only coherent with proceeding sentences.

Context
The cities of Annecy, Munich and Pyeongchang will battle it out to host the 2018 Winter Olympics. **[mask]** The International Olympic Committee have confirmed they have received applications from France, Germany and South Korea ahead of this week's deadline.
Predictions
RULE: Thousands of South Koreans gathered at the foot of a ski jump well past midnight in a passionate display of excitement that included fireworks, singing, dancing, picnicking and kimchi – the traditional Korean side dish.
PEGASUS: The cities of Annecy, Munich and Pyeongchang will battle it out to host the 2018 Winter Olympics.
Generative Augmentor (GA): The French resort of Annecy, the German city of Munich and the South Korean city of Pyeongchang have all submitted bids to host.
GA w/ Context Truncation: The International Olympic Committee's Executive Board will meet on Wednesday in Copenhagen to pick the host.

Table 5: Comparison of different local augmentation strategies.

Src Doc.	Sample-Level			Dataset-Level		
	ρ	r	τ	ρ	r	τ
UniEval						
Empty ("")	56.7	57.8	43.6	58.7	55.6	42.3
Original Src	57.5	55.4	44.2	59.2	53.3	42.5
CoUDA						
None	60.0	62.1	46.0	65.6	64.2	47.8

Table 6: Performance Comparison of UniEval w/ or w/o Source Document.

C Performance Comparison of UniEval w/ or w/o Source Document

Recall that UNIEVAL requires a source document as input when assessing coherence. Since our framework solely takes the discourse as input, we set its source document to empty string for fair comparison. In this section, we conduct additional experiments to explore how the source document influences coherence evaluation for UNIEVAL. Results are presented in Table 6. It can be observed that whether the source is provided does not have a significant impact on the performance of UNIEVAL. This further consolidates our assumption that coherence is an intrinsic quality of discourse that its evaluation does not require other inputs. Furthermore, even with original source provided to UNIEVAL, COUDA's performance remains substantially superior, verifying the effectiveness of our proposed

[Task Introduction]

You will be given two discourse. Your task is to rate both discourses on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

[Evaluation Criteria]

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby 'the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic'.

[Evaluation Steps]

{Evaluation steps generate through CoT}

[Discourse A]

{discourse A}

[Discourse B]

{discourse B}

Please avoid assigning same coherence score to A and B, because we want to pick the more coherent one.

Output with the following format:

The score of A: <score>

The score of B: <score>

Figure 5: Skewed Template for G-Eval-3.5 in pairwise ranking. We adopt the Balanced Position Calibration strategy proposed by Wang et al. (2023b) to alleviate positional bias of LLMs

method.

D Skewed Template to use G-Eval for Pairwise Ranking

Skewed template to use G-Eval for pairwise ranking is presented in Figure 5. We adopt the Balanced Position Calibration strategy proposed by Wang et al. (2023b) to alleviate positional bias of LLMs.

E The choice of Weight Parameter λ

Figure 6 shows the results of varying weight parameter λ for global and local coherence score. We see that the best weight for Spearman correlation and Kendall correlation is around 0.4, while the best weight for Pearson correlation is around 0.6.

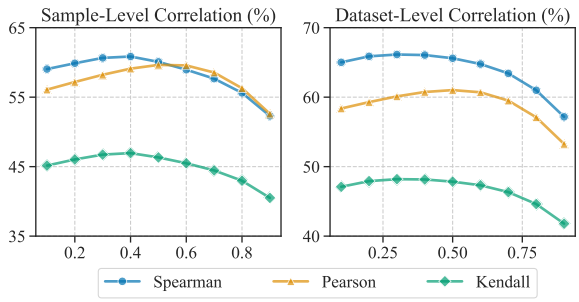

Figure 6: Sample-level correlation and dataset-level correlation on SUMMEVAL with different weight parameter $\lambda \in [0, 1]$ for global and local coherence score.