A Novel Computational Modeling Foundation for Automatic Coherence Assessment

Anonymous ACL submission

Abstract

Coherence is an essential property of well-written texts, that refers to the way textual units relate to one another. In the era of generative AI, coherence assessment is essential for many NLP tasks; summarization, long-form question-answering, etc. Current NLP approaches for modeling coherence often rely on a proxy task, specifically sentence reordering. However, such an approach may not capture the full range of factors contributing to coherence. To bridge this gap, in this work we employ the formal linguistic definition of Reinhart (1980) of what makes a discourse coherent, consisting of three conditions — *cohesion, consistency* and *relevance* – and formalize these conditions as respective computational tasks. We hypothesize that (i) a model trained on all of these tasks will learn the features required for coherence detection, and that (ii) a joint model for all tasks will exceed the performance of models trained on each task individually. We evaluate this modeling approach on two human-rated coherence benchmarks: one of automatically-generated stories and one of real-world texts. Our experiments confirm that jointly training on the proposed tasks leads to better performance on each task compared with task-specific models, and to better performance on assessing coherence overall, compared with strong baselines. Our formal coherence framework paves the way for advanced, broad-coverage automatic assessment.

1 Introduction

The term *coherence* refers to the quality of texts where sentences and paragraphs flow smoothly and are logically connected, creating a clear and understandable progression of ideas. Coherence detection is crucial for NLP tasks involving text quality measurements such as essay scoring or text quality assessment (Somasundaran et al., 2014; Feng et al., 2014; Lai and Tetreault, 2018). Its importance is further amplified nowadays in the era of large language models (LLMs). Ensuring coherent LLM outputs is essential for producing meaningful and understandable text for generative tasks as summarization and question answering (Guan et al., 2021; Xu et al., 2018; Yi et al., 2019), to name a few.

The elusive nature of coherence makes it challenging for NLP systems to assess it effectively. While various linguistic theories of coherence exist (e.g., Halliday and Hasan (1976); Joshi and Weinstein (1981); Givon (1995); Hobbs (1979); Dijk (1979); Mann and Thompson (1988)), current approaches often rely on proxy tasks like *sentence reordering* (Lapata, 2003; Miltsakaki et al., 2000). However, this approach oversimplifies coherence, potentially leading to models that struggle with real-world texts (Laban et al., 2021). Furthermore, coherence is multifaceted, varying across genres, contexts, and styles. Proxy tasks often fail to capture this complexity, leading to models unable to generalize across different domains and real-world settings. On top of that, existing NLP models for coherence, such as Barzilay and Lapata (2008)'s sentence reordering, often lack direct evaluation and are instead assessed on downstream tasks like readability (Guinaudeau and Strube, 2013; Mesgar and Strube, 2016, 2018) or essay scoring (Mesgar and Strube, 2018; Somasundaran et al., 2014; Tay et al., 2017). This approach can be expensive, biased towards the downstream task, and may overlook core aspects of coherence.

This work aims to provide a computationally workable definition of coherence, leveraging Reinhart (1980)'s theory which defines three conditions: *cohesion, consistency*, and *relevance*. Concretely, we propose to instantiate these conditions as computational tasks, and train them jointly to create a model that captures these properties. We hypothesize that (i) such a model will effectively assess coherence, and (ii) shared information between tasks will improve task-specific performance.

To test our hypotheses, we implement a model trained jointly on tasks capturing Reinhart's conditions of coherence. The unified model incorporates five tasks: sentence reordering (Lapata, 2003), discourse relation detection (Miltsakaki et al., 2004), natural language inference (Dagan et al., 2005), NP enrichment (Elazar et al., 2022), and irrelevant-sentence detection. Despite its relatively simple architecture, our model achieves SOTA results on most of these tasks. We then evaluate the model's coherence assessment capability on two human-annotated benchmarks: the *Grammarly Corpus of Discourse Coherence (GCDC)* (Lai and Tetreault, 2018) for real-world texts across four domains, and *CoheSentia* (Maimon and Tsarfaty, 2023) for artificially-generated stories. These benchmarks cover both natural and artificially-generated texts, human-rated for their levels of coherence.

Our empirical findings confirm our hypotheses, showing significant accuracy improvements and producing new SOTA results on both benchmarks. The joint model outperforms standalone models for individual tasks. Our significant performance improvements demonstrate the model's efficacy in automatic coherence assessment. This framework paves the way for future models to not only identify incoherence, but also analyze its causes. Furthermore, integrating this methodology into text generation may lead to higher-quality outputs.

2 The Proposal: Coherence à la Reinhart

This work aims to provide a computationally workable definition of coherence by adopting Reinhart (1980)'s formalization, which identifies coherence through three criteria: *Cohesion*, *Consistency*, and *Relevance*. According to Reinhart, a text is coherent only if it meets all three criteria. Recently, Maimon and Tsarfaty (2023) used this framework to create a benchmark for coherence scoring of GPT-generated text, with human scores for these criteria. Here we take a different approach, using these conditions as the basis for computational modeling, aiming to predict the ingredients of these three properties via jointly trained tasks.

Cohesion The cohesion condition focuses on the formal elements that link sentences together.[1] Reinhart states that a text is cohesive if, for every sentence pair, at least one of two conditions is met:

(1) Referentially linked: A pair of sentences $\langle S_1, S_2 \rangle$ is referentially linked when S_2 references an entity mentioned in S_1. A simple example is using a pronominal anaphor:

> "Dan is nice. Even Su likes him."

Here, the underlined entities co-refer. Other types of referential links are prepositional links (Elazar et al., 2022) or bridging anaphora (Hou, 2021).

(2) Linked by a semantic sentence connector. A pair of sentences $\langle S_1, S_2 \rangle$ is connected if a discourse relation links them. These connectors indicate semantic relations like cause and effect, comparison, contrast, and more (Prasad et al., 2008). An example of linking by a semantic connector is:

> "It was raining. So, we stayed inside."

The sentences are cohesive due to the existence of the "So" connector. These connectors may be explicit or implicit (Pitler et al., 2009).

Consistency The consistency condition pertains to the formal semantic aspects of a text, ensuring *logical* coherence, which is crucial for interpreting and deriving meaning. Formally, this condition requires that for a set of sentences $\{S_i\}_{i=0}^{n-1}$, the meaning of each sentence S_i must be consistent with all previous sentences $\{S_j\}_{i=0}^{i-1}$. This means all sentences can be true within a single world, not violating this worlds assumptions and restrictions. An example of a violation is shown below:

> "My father is dead now. That's why he has decided to smoke a pipe" (Freeman and Gathercole, 1966)

Despite being cohesive (anaphora & connectors), the passage lacks coherence due to world knowledge violations (a deceased cannot decide).[2]

Relevance The relevance condition involves *pragmatic* aspects, imposing constraints on the relationships of all sentences $\{S_i\}_{i=0}^{N-1}$ to the discourse topic and other contextual elements. An example of a violation of this condition is as follows:

> "I poured some chemical into a beaker. The chemical fell on my hand. The professor immediately took me to the emergency bath. He is a great musician."

[1] The terms "coherence" and "cohesion" may be confusing to non-linguists. Cohesion relates to the surface forms used (e.g., connectors, pronouns), while coherence pertains to the overall semantics and flow of ideas.

[2] The consistency condition was further explored by Honovich et al. (2021) to enhance the reliability of automatically generated texts.

The last sentence is cohesive and consistent with the previous sentences but is irrelevant to the overall context and topic of the story.

All in all, Reinhart's theory outlines conditions encapsulating the fundamental aspects of coherence to determine text coherence. We propose designing NLP tasks to detect these properties.

3 Research Hypotheses and Tasks

At the core of our approach is the implementation of Reinhart's coherence conditions as computational tasks, using a minimal set of NLP tasks designed to capture the features of cohesion, consistency, and relevance. We hypothesize that a model trained jointly on *all* tasks will detect coherence effectively, and will outperform models trained on each task individually.

To verify this, we define five tasks reflecting these coherence conditions.

The Sentence Reordering (SRO) Task This self-supervised task, proposed by Lapata (2003), involves reordering shuffled sentences to restore their original coherent form. For example, given the following input: "(1) Finally, the parser is evaluated. (2) We develop a useful parser. (3) Then we present our parser. (4) We first describe the older one." the correct order is $(2) \rightarrow (4) \rightarrow (3) \rightarrow (1)$.

Extending prior work on natural sentence order and coherence (Lin et al., 2011), a model excelling at paragraph reconstruction should capture syntactic and semantic relationships between sentences, reflecting both *cohesion* and *consistency*.

The Discourse-Relation Recognition (DRR) Task Given a pair of sentences (discourse units - DUs), we aim to predict their discourse relation, reflecting notions such as cause and effect, comparison, and contrast. For example, with the following input: "John worked all night. He slept all day today." the model is expected to detect a relation marker reflecting *contingency* (e.g., 'so', 'hence').

The discourse relation identification task enhances the model's ability to connect sentences, addressing the second sub-condition of *cohesion*.

The NP Enrichment (NPE) Task Introduced by Elazar et al. (2022), the NPE task identifies prepositional links between noun phrase (NP) entities. Given NP pairs , it determines the existence of a prepositional relation and identifies the best preposition describing it $p(NP_1 NP_2)$. For a paragraph with k NP entities, the model outputs the prepositional links for all NP pairs where such a relation exists (or NONE otherwise).

For example, in the paragraph: "Crown Princess Mary of Denmark gives birth to a male child." there are 4 NPs and thus 12 NP pairs. Sample outputs for these NP pairs are: (1) in(birth, Denmark) and (2) of(birth, male child).

A model trained on this task captures referential links between different parts of the discourse, serving as a proxy for the referential linking sub-condition of *cohesion*.

The Natural Language Inference (NLI) Task The NLI task (Bowman et al., 2015) aims to determine the truth value of a hypothesis based on a given premise. For example, given the premise: "John inspects the uniform of a figure in some East Asian country." and the hypothesis: "John is sleeping." the output will be a *contradiction*.

NLI evaluates NLP models' ability to capture logical relationships between sentences, serving as a proxy for the *consistency* condition.

The Irrelevant Sentence Recognition (ISR) Task We propose a self-supervised task where the model identifies irrelevant sentences in a coherent paragraph. Given a paragraph with N sentences, including one irrelevant sentence s, the model detects and outputs the irrelevant sentence.

For example, given the following input: "(1) Rick is a helpful kid. (2) He does the dishes. (3) He avoids doing his homework. (4) He helps older people." The irrelevant sentence is (3).

The model is trained to assess sentence relevance to the overall topic and context, acting as a proxy for the *relevance* condition.

Putting It All Together: We propose a Multi-Task Learning (MTL) approach, where a model is jointly trained on these tasks to capture all coherence conditions outlined by Reinhart. This method leverages shared information during training, with the goal of enhancing both overall coherence detection and individual task performance. To assess coherence, we define two types of tasks:

- **The Coherence Scoring Task** To confirm our hypothesis that the proposed model captures coherence, we evaluate its performance on the coherence scoring task, where given a paragraph P the model predicts the coherence score C as a human reader would.

- **The Coherence Reasoning Task** To examine conditions contributing to incoherence (cohesion, consistency, relevance) beyond a final score, we use the coherence reasoning task proposed by Maimon and Tsarfaty (2023). Given a paragraph P and a new sentence s, the model predicts whether s is *cohesive*, *consistent*, or *relevant* to P using distinct classifiers.

We hypothesize that utilizing the MTL-powered architecture will improve the results on both coherence scoring and coherence reasoning.

4 Coherence Assessment Setup

Here, we detail the models and experimental setup for the coherence assessment tasks we define.

4.1 The Coherence Scoring Task

Models: We use two architectures for coherence scoring: Classification-Based (BERT (Devlin et al., 2019)) and Generation-Based (T5 (Raffel et al., 2020)). The model predicts for a given text 3-way or 5-way scores, depending on the dataset.

In the Classification-Based models, the coherence score C is predicted given the text P using a classification head.

In Generation-Based models, the input includes the text with dataset-specific prompts and an output. Example prompts and outputs are in Appendix F.

Further details on experimental settings are in Appendix B.

Datasets and Evaluation: We evaluate our model on two datasets: GCDC (Lai and Tetreault, 2018) and CoheSentia (Maimon and Tsarfaty, 2023). GCDC includes real-world text from various domains (Clinton emails, Enron emails, Yahoo Answers, Yelp reviews) with coherence scores from 1 (not coherent) to 3 (highly coherent). CoheSentia features GPT-3 generated stories (fiction and non-fiction) with scores ranging from 1 to 5. We use the "incremental final score" for CoheSentia stories. Dataset sizes and splits are detailed in Table 1.

To remain compatible with Lai and Tetreault (2018), we use accuracy as the metric for evaluating the final coherence score of the text.

Baselines: To assess the effectiveness of our proposed model, we compare its performance on each dataset to the current SOTA models. For GCDC Lai and Tetreault (2018) introduced the ParSeq model with three stacked LSTMs for sentence, paragraph, and document embeddings, followed by a coherence classification layer. The SOTA model for GCDC by Liu et al. (2023) uses a multi-step approach: identifying document graph structures, converting subgraphs, constructing corpus-level graphs based on shared subgraphs, and encoding connections with a GCN.

For CoheSentia Maimon and Tsarfaty (2023) created the SOTA model using a prompt-based approach with Flan-T5-large to assess coherence by adding a question at the beginning of each text.

4.2 The Coherence Reasoning Task

Models: In Classification-Based models, the reasoning for each coherence attribute given the paragraph P and the new sentence s is predicted using a classification head.

In Generation-Based models, the input includes the text with prompts and an output. Prompts and outputs for both datasets are in Appendix F.

Datasets and Evaluation: We evaluate our model on the CoheSentia corpus (Maimon and Tsarfaty, 2023), which contains automatically generated stories with human annotations for cohesion, consistency, and relevance. We use precision, recall, and F1 scores for each property.

Baselines: We evaluate our model's effectiveness on the CoheSentia dataset by comparing it to the current SOTA model by (Maimon and Tsarfaty, 2023), which uses a prompt-based approach with Flan-T5-large, adding a question at the beginning of each text to assess coherence.

5 Task-Specific Experimental Setup

In this section, we elaborate on the modeling of the coherence proxy tasks. For each task, we evaluate two model variants: Classification-Based and Generation-Based, as detailed below. Table 2 summarizes the datasets, evaluation metrics, and key statistics for each task (see further elaboration in Appendix A). For the Generation-Based models, prompts and outputs are detailed in Appendix F.

The Sentence Reordering Models For the Classification-Based models, we adopt the topological sort architecture from Shrimai Prabhumoye (2020). Each paragraph's sentence pairs are represented as triplets $\langle S_i, C_k, S_j \rangle$, indicating whether S_i precedes or follows S_j. The model has two stages: a binary classification head that predicts the pairwise relations and a second stage that pro-

Dataset	Split			Per Instance			
	Train	Validation	Test	Max #tokens	Avg #tokens	Max #sent.	Avg #sent.
GCDC	3.6k	800	800	333	156	10	32
CoheSentia	350	75	75	226	150	15	6.5

Table 1: Main Statistics on the Datasets for Coherence Scoring

duces the predicted order using the topological sort algorithm (Tarjan, 1976).

The Generation-Based models use prompts and predict the outputs with the final order.

The Discourse-Relation Recognition Models In the Classification-Based models, the input consists of a pair of DUs: $\langle DU_1, DU_2 \rangle$. A classification head predicts the discourse relation between them.

In the Generation-Based models, the input is an argument pair, and the model employs a chain-of-thought (CoT) method (Wei et al., 2023) to predict the discourse relation.[3] The CoT structure is $\langle$connector$\rangle \rightarrow \langle l_1$ relation$\rangle \rightarrow \langle l_2$ relation$\rangle$. That is, the model adopts a three-stage approach to predicting the L_2 relation type. The model first infers the implicit connective, then generalizes it to a broader relation category.

The NP Enrichment Models In the Classification-Based models, we extend the Bi-Affine architecture from Dozat and Manning (2017) to predict preposition relations between NP pairs instead of syntactic dependency labels. NP embeddings are created by pooling tokens representing each NP, and the model head predicts the preposition using the NP's anchor and complement representations (Figure 4 in Appendix).

The Generation-Based models predict prepositional relations for each NP pair independently. The input consists of the document text and a prompt specifying the NP pair.

The NL Inference Models Classification-Based models predict relations between the premise and hypothesis.

Generation-Based models use prompts and predict outputs for premise-hypothesis pairs.

Irrelevant Sentence Recognition Models The Classification-Based models have two stages. Sentence pairs form triplets $\langle S_i, C_k, S_j \rangle$, where $C_k = 0|1$ indicates relevance. A binary classification head determines the relation. In the second stage, the sentence with the lowest combined relations score is deemed irrelevant.

Generation-Based models use prompts and predict the irrelevant sentence as the output.

[3]CoT detection of discourse relations outperformed simpler prompts in our preliminary experiments.

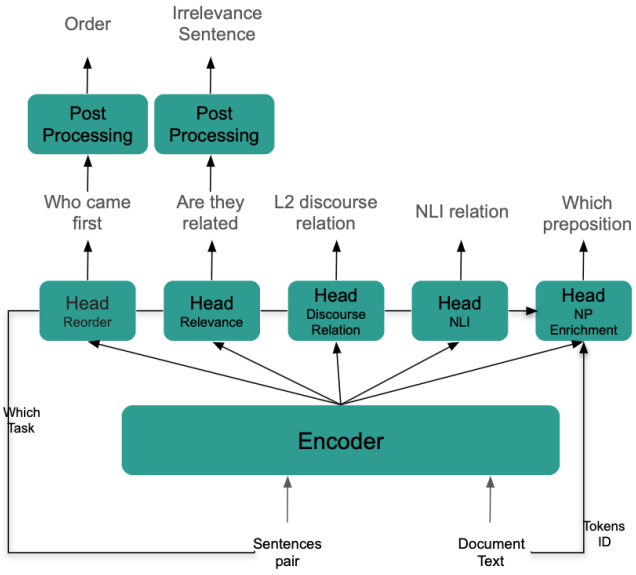


Figure 1: Illustration of the encoder-only model where the input is a pair of sentences (most tasks) or for NPE task the input is a document and the token IDs of different NPs

The Overall Joint Architecture To test our hypotheses, we implemented both Classification-Based and Generation-Based models, trained to solve all tasks jointly.

For the Classification-Based variant, we use MTL (Caruana, 1997) with a BERT encoder shared across tasks, each with a unique classification head. Each head predicts task-specific outputs (see Fig. 1). To address forgetting in MTL (Goodfellow et al., 2014), we implement an interleaved training strategy, alternating tasks in batches, ensuring each batch contains samples from a distinct task, effectively mitigating forgetting during MTL training.

For the Generation-Based model, we use the T5 encoder-decoder model, which allows concurrent fine-tuning of multiple tasks using distinct prompts. The prompt structure remains the same as individual task fine-tuning, but batches contain samples from specific tasks, distinct from the previous batch.

Task	Dataset	Metrics	Split			Per Instance			
			Train	Dev	Test	Max #toks	Avg #toks	Max #sent.	Avg #sent.
SRO	RocStories (Mostafazadeh et al., 2016)	PMR (Chen et al., 2016) Acc (Logeswaran et al., 2017)	68k	14k	14k	135	57	5	5
ISR	RocStories (Mostafazadeh et al., 2016)	Accuracy	68k	14k	14k	152	77	6	6
DRR	PDTB3 (Prasad et al., 2019)	Accuracy	17.5k	1.7k	1.5k	556	30	2	2
NPE	TNE (Elazar et al., 2022)	F1, Precision & Recall	3.5k	500	500	284	163	15	6.9
NLI	MNLI (Williams et al., 2018)	Accuracy	393k	7.5k	2.5k	(194,70)	(20,10)	(8,8)	(2,2)

Table 2: The datasets and metrics used for each task and the train/dev/test split size with the max and average number of tokens and sentences. For the NLI task (x,y) refer to the numbers of (premise, hypothesis) respectively

Model	GCDC	CoheSentia
Lai and Tetreault (2018)	57.5	—
SOTA	61.2	35.3
Ours-None (bert-large)	50.2	34.3
Ours-ALL (bert-large)	72.5	55.7
Ours-None (t5-large)	56.3	34.8
Ours-ALL (t5-large)	**76.4**	**62.3**
Controlled-nonCoherence (t5-large)	52.8	36.8

Table 3: Accuracy on Coherence Scoring The SOTA for GCDC is by Liu et al. (2023) and for CoheSentia is Maimon and Tsarfaty (2023)

6 Results

We first aim to test the hypothesis that a model jointly trained on tasks reflecting the different coherence conditions will effectively assess coherence. Table 3 shows the coherence scores of our jointly fine-tuned model (Ours-ALL) on the GCDC and CoheSentia datasets, compared to current SOTA models on either dataset. Compared to these models, our jointly fine-tuned model shows significant improvements in coherence scoring, especially on CoheSentia. We observe a 15% and 27% accuracy gain for GCDC and CoheSentia respectively, demonstrating that our proposed approach and selected tasks effectively contribute to capturing fundamental aspects of coherence.

We further analyze the contribution of the proxy tasks (Ours-All) by comparing it to a model without such fine-tuning (Ours-None) to isolate performance gains. As evident in Table 3 these tasks dramatically enhance performances. These results are supported by further qualitative analysis in Appendix D.

Next we analyze the MTL model's success on assessing the coherence conditions (cohesion, consistency, relevance), by fine-tuning on the coherence reasoning task. We compare the results to SOTA from Maimon and Tsarfaty (2023), who fine-tuned the Flan-T5 model with a simple prompt, and to our model without initial fine-tuning on the coherence proxy tasks (Ours-None). Table 4 summarizes the coherence reasoning task results for all attributes and metrics. Our model achieves SOTA performance across all coherence conditions, demonstrating the efficacy of our approach.

Finally, We evaluate the task-specific performance of models trained with either individual or joint fine-tuning on the proxy tasks, using both the Classification-Based and Generation-Based variations. Results are in Table 5, alongside comparisons to current SOTA on these benchmarks. Our findings consistently show that Generation-Based models outperform Classification-Based ones. More importantly, joint fine-tuning across all tasks consistently surpasses individual fine-tuning, particularly in the SRO, ISR, and DRR tasks, where it leads to significant performance improvements and even surpasses SOTA benchmarks. For the NPE task, joint fine-tuning achieves substantial recall gains, though precision falls short of SOTA results, offering a more balanced performance. The exception is the NLI task, where our model's performance is lower than SOTA.

In summary, our MTL model outperforms single-task models on all tasks, achieving SOTA results except for NLI, in line with our hypothesis on the benefits of the joint architecture. Moreover, our MTL model, jointly trained on coherence proxy tasks, significantly improves performance, enhancing coherence scoring for both datasets and excelling in coherence reasoning, in line with the second part of our hypothesis.

7 Analysis

7.1 The Effect of Different Tasks on Coherence Scoring

This section examines how fine-tuning on diverse subsets of coherence proxy tasks affects coherence scoring. We fine-tune models on various combinations of these tasks, then perform final fine-tuning and evaluation on the coherence scoring task.

Model	Cohesion			Consistency			Relevance		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
SOTA	72.4	72.1	72.2	59.6	67.5	63.3	56.4	74.6	59.5
Ours-None (bert-large)	66.4	59.4	62.7	60.4	56.5	59.6	49.2	49.9	49.5
Ours-ALL (bert-large)	74.7	70.5	72.5	70.6	68.2	69.3	59.8	61.1	60.4
Ours-None (t5-large)	81.1	80.3	80.7	60.4	62.6	61.5	48.1	49.6	48.8
Ours-ALL (t5-large)	**83.1**	**83.2**	**83.1**	**78.5**	**80.3**	**79.4**	**70.8**	**76.9**	**73.7**

Table 4: Results for Coherence Reasoning Task. The SOTA is by Maimon and Tsarfaty (2023)

Model	SRO		ISR	DRR	NPE			NLI
	PMR	ACC	Accuracy	Accuracy	F1	P	R	Accuracy
SOTA	81.9	90.8	-	64.7	64.0	**80.5**	53.1	**92.0**
Ours-Individual (bert-large)	51.8	69.5	60.4	60.0	53.1	67.1	44.0	87.4
Ours-ALL (bert-large)	67.1	83.2	78.6	65.7	64.4	79.8	54.2	90.2
Ours-Individual (t5-large)	75.7	87.8	80.4	64.8	59.8	68.5	53.1	89.9
Ours-ALL (t5-large)	**83.8**	**92.1**	**82.2**	**67.3**	**76.7**	76.7	**76.7**	91.5

Table 5: Results for all proxy tasks compared to SOTA performances. The SOTA model for SRO is ReBART (Basu Roy Chowdhury et al., 2021), for DRR is Contrastive Learning (Long and Webber, 2023), for NPE is TNE (Elazar et al., 2022) and for NLI T5-11B (Raffel et al., 2020)

Figure 2 shows the impact of fine-tuning proxy coherence tasks on coherence scoring performance. Models fine-tuned on any coherence proxy task outperform those without fine-tuning (Ours-None), highlighting their effectiveness. Performance generally improves with more tasks, especially beyond three, indicating cumulative benefits.

Interestingly, NLI fine-tuning significantly enhances performance, likely due to its role in improving the model's ability to capture consistency, crucial for coherence assessment. Additionally, ISR fine-tuning is more impactful when combined with other tasks. These findings underscore the importance of task selection and task interaction during fine-tuning for optimal coherence scoring.

7.2 Impact of Non-Coherence Tasks Fine-Tuning

In this Section we aim to empirically refute a possible hypothesis that the joint ALL model outperforms the NONE model simply due to its complexity, regardless of the nature of the tasks used (i.e., tasks reflecting coherence conditions).

To this end, we compare the performance of our fine-tuned on coherence-tasks model (Ours-ALL) with a model fine-tuned on three tasks orthogonal to coherence, followed by fine-tuning on the coherence scoring task: (i) Machine Translation (MT): We sample 15k instances from the WMT14 dataset (Bojar et al., 2014). (ii) Named-Entity Recognition (NER): We use the Conll2003 dataset (Tjong Kim Sang and De Meulder, 2003) containing annotations for 14k instances. (iii) Part-of-Speech (POS), using the same Conll2003 dataset containing the POS tags as well.

Model In these experiments, we used the T5-large model as the basis, employing specific prompt and output designs for each task. For the NER and POS tasks, we adapted the "Sentinel + Tag" architecture by Raman et al. (2022). Detailed prompts and sample outputs are provided in Appendix F.

Results Following the same procedure as in our main experiments, fine-tuned models were assessed for coherence scoring using the GCDC and CoheSentia benchmarks (detailed in Table 3).

Fine-tuning on tasks orthogonal to coherence yielded minimal to no improvements over the baseline (Ours-None) and significantly underperformed compared to our final MTL model (Ours-ALL). This highlights the importance of coherence-specific proxy tasks for effective coherence detection, as unrelated tasks can hinder performance.

7.3 Cross-Domain Generalization

Since GCDC and CoheSentia present different domains and writing styles, we evaluate model generalizability by fine-tuning on one dataset and assessing coherence on the other. Table 6 presents the results for our MTL model (Ours-ALL) and the non-coherence fine-tuned model (Ours-None) under three settings: fine-tuning on CoheSentia only, GCDC only, and both combined.

Results demonstrate performance gains across domains, highlighting the generalizability of our method. Combining data improves performance, with Ours-ALL showing a 12% and 14% error reduction on CoheSentia and GCDC, respectively, compared to in-domain scenarios, underscoring the utility and transferability of the learned features.

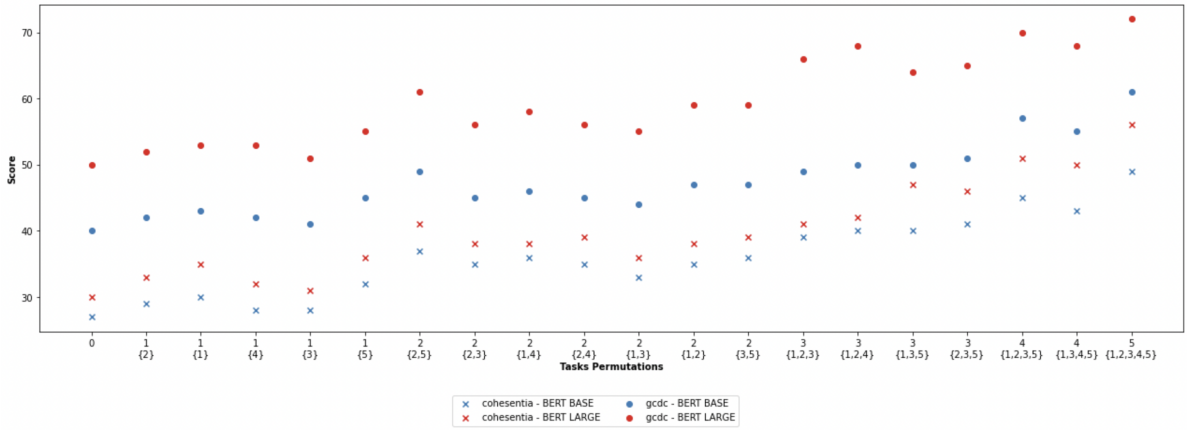


Figure 2: Accuracy for Coherence Scoring Task for both GCDC and CoheSentia with different proxy coherence task-subsets. The labels are tasks IDs (1-SRO, 2-ISR, 3-DRR, 4-NPE, 5-NLI)

Model	GCDC	CoheSentia
Ours-None-CoheSentia	52.8	34.8
Ours-None-GCDC	56.3	28.5
Ours-None-Both	57.5	35.4
Ours-ALL-CoheSentia	71.8	62.3
Ours-ALL-GCDC	76.4	59.5
Ours-ALL-Both	**79.8**	**66.7**

Table 6: Accuracy on coherence scoring on both datasets when fine-tuned based on T5-model on only one dataset

7.4 Effects of Different Tasks on One Another

We investigate the impact of fine-tuning on various coherence task subsets on individual task performance. The model was trained on different task combinations with increasing numbers of tasks and evaluated on each task separately.

Figure 3 shows consistent performance gains in the Sentence Reordering (SRO) task for BERT models as more tasks are jointly fine-tuned (see Appendix E for other tasks). This supports our hypothesis that shared information among coherence tasks enhances individual task performance.

The impact of specific tasks varies; for instance, DRR minimally affects SRO, likely due to limited training data. Notably, NLI significantly influences the performance of various tasks. The ISR task notably improves performance on other tasks, suggesting its effectiveness in capturing relevance errors, crucial for coherence assessment. We thus emphasize the introduction of this self-supervised ISR task and advocate for its exploration in future research to enhance coherence assessment.

The overall performance trends are similar for both BERT-base and BERT-large models, indicating that the impact of specific tasks is consistent regardless of model size.

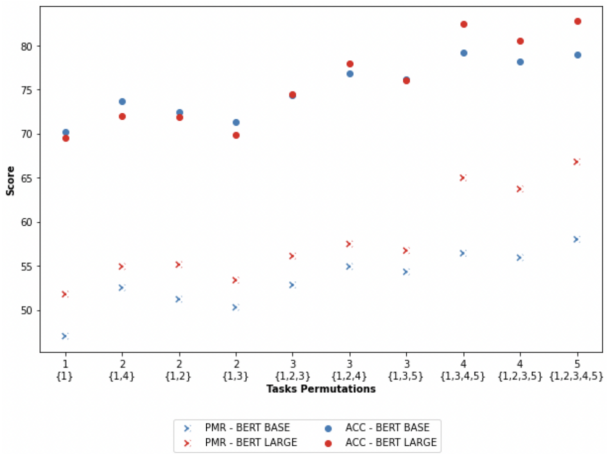


Figure 3: Results for SRO task, for different subsets of coherence tasks fine-tuned upon. The labels are the number of tasks and in curly brackets which tasks (1 - SRO, 2 - ISR, 3 - DRR, 4 - NPE, 5 - NLI)

8 Conclusion and Future Work

In this paper we propose a new coherence modeling method, based on Reinhart (1980)'s theory which defines the conditions needed for coherence: cohesion, consistency, and relevance. We use five key NLP tasks as proxies for these conditions, and train an MTL model on them jointly. Our unified coherence model achieves SOTA results on these individual tasks, and moreover it excels in coherence scoring for both real-world and generated texts. We conjecture that this framework will enhance NLP systems' ability to quantify and evaluate text quality automatically. Future follow-up research will focus on using these conditions for improved coherent-text generation, and for detecting particular causes of incoherence automatically. Our code and models are publicly available, to encourage further research on *coherence scoring* and *coherence reasoning*.

Limitations

While this work advances the modeling and automatic evaluation of coherence, limitations exist that suggest promising avenues for future research.

Dataset Limitations Existing coherence evaluation datasets like GCDC and CoheSentia, along with datasets for our proxy tasks, primarily focus on relatively short texts. To address this, we analyzed the performance of our MTL models (Ours-ALL) and the non-coherence version (Ours-None) on GCDC and CoheSentia across various text lengths after fine-tuning for coherence scoring (see Figure 7 in the Appendix). As expected, for both models and datasets, accuracy decreased with longer texts, highlighting the increased difficulty of assigning coherence scores for complex passages. This observation aligns with recent work suggesting that while LLMs can handle longer texts, their reasoning abilities might decline with increasing text length (Levy et al., 2024).

Focus on Short Texts Our current study focused on short texts ($\leq$ 512 tokens). The effectiveness of our approach on longer documents remains an open question for future exploration. We hypothesize that incorporating coherence proxy tasks could benefit the model's performance on longer texts, but further investigation is necessary.

Acknowledgements

References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. *Preprint*, arXiv:1807.05154.

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

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. Is everything in order? a simple way to order sentences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10769–10779, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ale s Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Preprint*, arXiv:1508.05326.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Neural sentence ordering. *Preprint*, arXiv:1607.06952.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Teun A Van Dijk. 1979. Pragmatic connectives. *Discourse processes*, 3:447–456.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. *Preprint*, arXiv:1611.01734.

Yanai Elazar, Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2022. Text-based np enrichment. *Preprint*, arXiv:2109.12085.

Vanessa Wei Feng, Ziheng Lin, , and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1:940–949.

Thomas Freeman and CE Gathercole. 1966. Perseveration—the clinical symptoms—in chronic schizophrenia and organic dementia. *The British Journal of Psychiatry*, 112(482):27–32.

T Givon. 1995. Coherence in text vs. coherence in mind. *Coherence in spontaneous text*, pages 31–59.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2014. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. *Preprint*, arXiv:2105.08963.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, Dallas, Texas.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive science*, 3(1):67–90.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *Preprint*, arXiv:2104.08202.

Yufang Hou. 2021. End-to-end neural information status classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1377–1388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Aravind K Joshi and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure-centering. *IJCAI*, pages 385–387.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can transformer models measure coherence in text? re-thinking the shuffle test. *Preprint*, arXiv:2107.03448.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. *In Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. Association for Computational Linguistics*, 1:214–223.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. *EMNLPIJCNLP*, pages 2273–2283.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *Preprint*, arXiv:2402.14848.

Li Liang, Zheng Zhao, and Bonnie Webber. 2020. Extending implicit discourse relation recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147, Online. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, , and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics*, 1:997–1006.

Wei Liu, Xiyan Fu, and Michael Strube. 2023. Modeling structural similarities between documents for coherence assessment with graph convolutional networks. *Preprint*, arXiv:2306.06472.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. *Preprint*, arXiv:2004.12617.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2017. Sentence ordering and coherence modeling using recurrent neural networks. *Preprint*, arXiv:1611.02654.

Wanqiu Long and Bonnie Webber. 2023. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations. *Preprint*, arXiv:2301.02724.

Aviya Maimon and Reut Tsarfaty. 2023. Cohesentia: A novel benchmark of incremental versus holistic assessment of coherence in generated texts. *CoRR*, abs/2310.16329.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text Interdisciplinary Journal for the Study of Discourse*, 1:243–281.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423, San Diego, California. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. Association for Computational Linguistics.

Miltsakaki, Eleni, Kukich, and Karen. 2000. Automated evaluation of coherence in student essays.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *Preprint*, arXiv:1604.01696.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th*

International Joint Conference on Natural Language Processing of the AFNLP, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow discourse parsing using convolutional neural network. In *Proceedings of the CoNLL-16 shared task*, pages 70–77, Berlin, Germany. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Karthik Raman, Iftekhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi, and Krishna Srinivasan. 2022. Transforming sequence tagging into a seq2seq task. *Preprint*, arXiv:2203.08378.

Tanya Reinhart. 1980. *Conditions for text coherence. Poetics Today 1(4): 16t-180*, volume 1.

Alan W Black Shrimai Prabhumoye, Ruslan Salakhutdinov. 2020. Topological sort for sentence ordering.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1:950–961.

Robert Endre Tarjan. 1976. Edge-disjoint spanning trees and depth-first search. *Acta Informatica*, 6(2):171–185.

Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. 2017. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. *Preprint*, arXiv:1711.04981.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Wei Xiang, Bang Wang, Lu Dai, and Yijun Mo. 2022. Encoding and fusing semantic connection and linguistic evidence for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3247–3257, Dublin, Ireland. Association for Computational Linguistics.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. *Preprint*, arXiv:1808.06945.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. *Preprint*, arXiv:1904.13015.

A Tasks Specific Experimental Settings

In this section, we further elaborate on the datasets and evaluation metrics used for each one of the coherence proxy tasks.

A.1 The Sentence Reordering Task Setup

Topological Sort: A topological sort (Tarjan, 1976) linearly orders vertices in a DAG. The algorithm is presented in Algo A.1.

[t] Input: A digraph G with n vertices

Output: A topological ordering v1,v2...vn of G. L ← Empty list that will contain the sorted nodes S ← Set of all nodes with no incoming edge S is not empty remove a node n from S add n to L each node m with an edge e from n to m remove edge e from the graph m has no other incoming edges insert m into S graph has edges error (graph has at least one cycle) L (a topologically sorted order)

Dataset: We use the ROCStories (Mostafazadeh et al., 2016) dataset (Licence ID is CC-BY 4.0.) which contains 5-sentence stories. We use the standard 85:15 train/test split and randomly select a subset of the train for validation.

Evaluation: We use two common evaluation metrics for the reordering task:[4]

- Perfect Match Ratio (PMR): Chen et al. (2016) calculate the percentage of samples for which the entire sequence was correctly predicted.

$$PMR = \frac{1}{N}\sum_{i=1}^{N} 1\{\hat{O}^i = O^i\}$$

- Sentence Accuracy (Acc): Logeswaran et al. (2017) calculate the percentage of sentences for which their absolute position was correctly predicted.

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_i}\sum_{j=1}^{v_i} 1\{\hat{O}^i_j = O^i_j\}$$

A.2 The Discourse-Relation Recognition Task Setup

Dataset: We use the Penn Discourse TreeBank3 (PDTB3) Level 2 dataset (Miltsakaki et al., 2004; Prasad et al., 2008; Liang et al., 2020). We only used labels with more than 100 instances, which leaves us with 14 senses from L_2. The variability of data splits used in the literature is substantial, therefore, we follow earlier work by Ji and Eisenstein (2015); Bai and Zhao (2018); Liu et al. (2020); Xiang et al. (2022) using Sections 2-20, 0-1 and 21-22 for training, validation and testing respectively. When multiple annotated labels are present, we adopt the approach described by Qin et al. (2016) and consider them as distinct instances during the training phase. During testing, if a prediction matches any of the reference labels, it is considered correct.

Evaluation: We use the accuracy metric on the number of sentence pairs the model correctly predicted the L_2 discourse relation:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} 1\{\hat{R}^i = R^i\}$$

A.3 The NP Enrichment Task Setup

Token Classification Head: Figure 4 is an illustration of the token classification head for the NPE task.

[4]There are 5 metrics, we used the most common 2.

Dataset: We use the TNE dataset (Elazar et al., 2022) (Licence Free) which contains documents and relations between every noun pair in it (with a total number of nouns of 190k and a total number of NP relations of 1M). There are 28 possible relations (including 'no relation'). This dataset's advantage is that it contains real-world long paragraphs. As in the original publication split the data at the document level.

The distribution of the possible preposition between pair of nouns in TNE dataset is in Figure 5

Evaluation: We report precision, recall & F1 on NP pairs with prepositional links between them.

A.4 The NL Inference (NLI) Task Setup

Dataset and Evaluation: We use the MNLI dataset (Williams et al., 2018) (Licence ID CC-BY-3.0). with the accuracy metric on the amount of hypothesis-premise pairs that the model correctly predicts their relation R:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} 1\{\hat{R}^i = R^i\}$$

A.5 The Irrelevant Sentence Recognition Task Setup

Dataset: We again use ROCStories as in sentence reordering. Each story within the ROCStories dataset was augmented with a single, randomly inserted sentence. The irrelevant sentence for each story was randomly selected from the entire ROCStories dataset, with the sole constraint that it contained entities present in the target story. Both this and the sentence reordering task leverage the same benchmark, retaining the same train/dev/test splits.

Evaluation: We use the accuracy metric on the percentage of paragraphs where the model correctly detected the irrelevant sentence S:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} 1\{\hat{S}^i = S^i\}$$

A.6 Overall Experimental Settings

We trained each model three times, reporting the mean performance. Training utilized multiple Tesla V100 GPUs (up to 4) with 32GB memory each. For each architecture, the settings are:

1. Classification-Based: BERT (base and large) served as the encoder with fine-tuning across

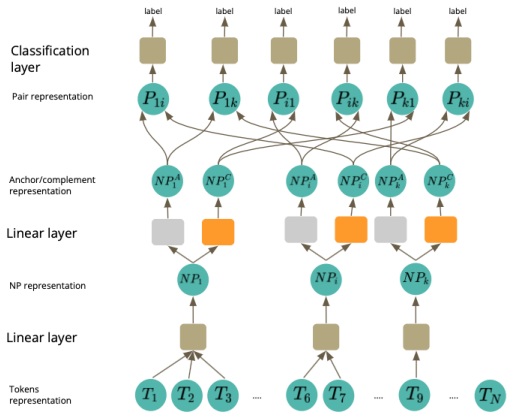


Figure 4: Illustration of the token head which contains several stages: starting with (1) embedding for each token in the text, (2) creating an embedding for each NP when it acts as the complement and the anchor separately, (3) a representation for each NP pair and finally (4) a classification layer

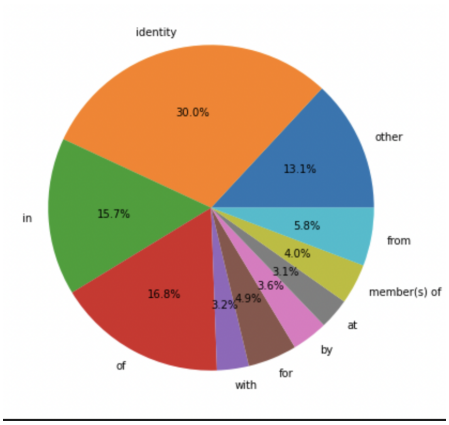


Figure 5: Distribution of the main prepositions in the NP Enrichment test set

all layers. We used Adam optimizer with a learning rate of 5e-5 and a dropout of 0.5. For tasks requiring classification (SRO, ISR, DRR, NLI), we employed a linear classification head with 512 hidden dimensions and 0.3 dropouts. The NPE utilized a different head structure (details omitted for brevity). Cross-Entropy loss was used for all datasets.

2. Generation-Based: T5 (base and large) models were used as the backbone. Training employed Adam optimizer with a learning rate of 5e-5. Models were trained with task-specific prompts and corresponding ground truth labels for supervised learning.

Both architectures shared the following hyper-parameters: fine-tuning for 3 epochs with early stopping, batch size of 4, and gradient accumulation steps of 2. The hyper-parameters were chosen using parameters-grid. Our code is based on the Huggingface library (Wolf et al., 2020).

B Coherence Assessment Experimental Settings

For each architecture, the settings are:

1. Classification-Based (BERT base and large): Encoder with fine-tuning across all layers, Adam optimizer (learning rate 5e-4), dropout (0.3). Each dataset used a linear classification head (512 hidden dimensions, 0.1 dropout). Cross-Entropy loss was used.

2. Generation-Based (T5 base and large): Encoder-decoder architecture, Adam optimizer (learning rate 2e-5). Inputs included prompts specific to each dataset (GCDC or CoheSentia) and the paragraph text.

The models share hyperparameters: 50 epochs with early stopping (accuracy), batch size of 4, and gradient accumulation steps of 2. We employed 10-fold cross-validation on both datasets (following Lai and Tetreault (2018)) using a single Tesla V100 GPU with 32GB memory. The hyper-parameters were chosen using parameters-grid. Our code is based on the Huggingface library (Wolf et al., 2020).

C Text Length vs. Coherence Score

The accuracy of the models on both coherence datasets based on different lengths is in Figure 7.

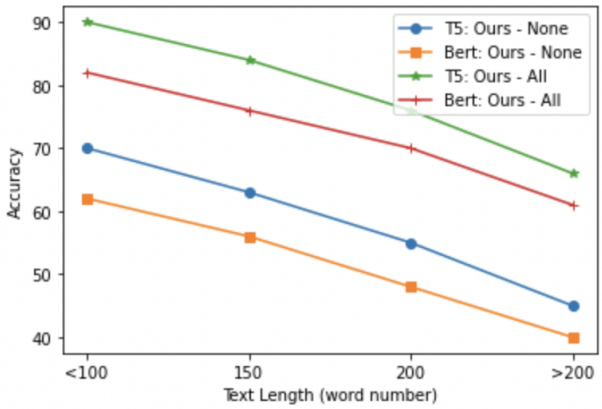


Figure 6: Accuracy For GCDC based on number of words

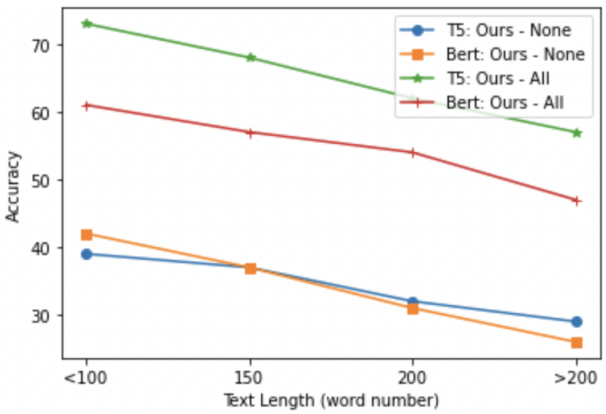


Figure 7: Accuracy For CoheSentia based on number of words

D Qualitative Analysis

D.1 Qualitative Analysis

To gain qualitative insights, we sampled 50 misclassified examples by SOTA models, from CoheSentia and GCDC. We then assessed these examples on various models, including our MTL model (Ours-ALL) and the non-coherence fine-tuning version (Ours-None).

For CoheSentia, the previous SOTA models favor extreme scores, likely due to training data imbalance. Our model exhibits greater robustness, predicting a more balanced distribution of scores. Figure 8a and Table 7 present an example of a text from the CoheSentia dataset and the predictions of the models. In this example, the base model (Ours-None) failed on coherence prediction, while our final model (Ours-ALL) succeeded. Figure 8b presents an example of text from GCDC dataset and Table 8 the predictions of different models on the coherence scoring task. This example highlights a complex case with cohesion and relevance violations. Both the baseline and ISR-trained models missed this issue, while our MTL model achieved accurate prediction.

Model	Prediction
Ground Truth	Medium
SOTA	High
Ours-None (BERT-large)	High
Ours-None (T5-large)	High
Ours-ALL (BERT-large)	Medium
Ours-ALL (T5-large)	Medium

Table 7: Predicted Coherence scores for the text in Figure 8a

Model	Prediction
Ground Truth	Low
SOTA	Medium
Ours-None (BERT-large)	Medium
Ours-None (T5-large)	Medium
Ours-ALL (BERT-large)	Low
Ours-ALL (T5-large)	Low

Table 8: Predicted Coherence scores for the text in Figure 8b

E Results for Subsets of Tasks

Figures 9a, 9b, 9c, 10a and 10b visualize the performance of coherence proxy tasks across fine-tuning settings for BERT-base and BERT-large models. It highlights how subsets of tasks impacts target task performance.

F T5 Prompts and Outputs for Different Tasks

In Table 9 we detail the various prompts used for fine-tuning T5 models on all explored tasks in this work.

In Table 10 we detail the various outputs used for fine-tuning T5 models on all explored tasks in this work.

"Shed been a widow for over two years and was starting to lose hope of ever seeing her husband again. One day, she received an email from him asking if he could come out for fun at her funeral. She skeptically agreed but soon found herself enjoying his company more than she could have imagined. As they went around the Neapolitan town where she belonged, it quickly became clear that their bond was even stronger then before– they laughed and danced together like teenagers on celebrated days like this one. It seemed equitable that he should be there too– as long as he didnt mind being the man in attendance at her burial pyre."

(a) CoheSentia

"Guy from Mexico is in NY and is cooperating. Discussions with him continue this am. Since he is cooperating, no move to court or to presentment scheduled yet.\n\nMexican support has been excellent throughout. Alice has call sheet for Espinosa — call can take place whenever its convenient for you later this morning (Espinosa is apparently out on West Coast, but Ops could confirm time difference).

Holding off for now on other calls that rest of us would make (Saudis, et al), pending further developments in NY.

Will let you know as soon as we have more."

(b) GCDC

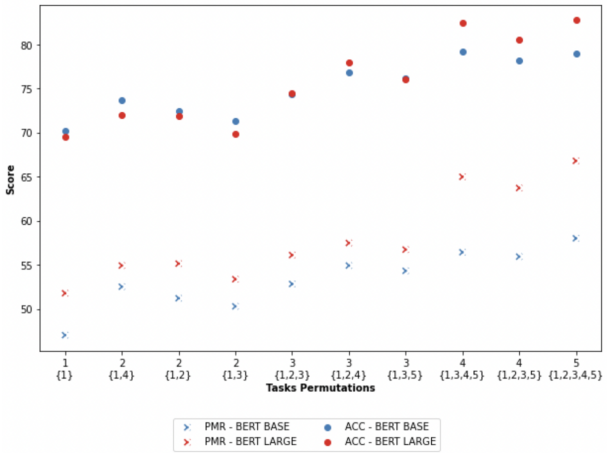
Figure 8: Sample Texts for coherence scoring tasks: GCDC & CoheSentia benchmarks

Task Name	Dataset Name	Prompt
SRO	ROCStories	"reorder: what is the order of the sentences so that the paragraph is coherent? sentence 1: $\langle S_1 \rangle$ sentence 2: $\langle S_2 \rangle$ ... $\langle S_N \rangle$"
ISR	ROCStories	"relevance: what is the irrelevant sentence in the text? sentence1: $\langle S_1 \rangle$ sentence2: $\langle S_2 \rangle$ sentence3: ...$\langle S_N \rangle$"
DRR	PDTB3	"discourse relation: what is the discourse relation between $\langle DU_1 \rangle \langle DU_2 \rangle$"
NPE	TNE	"coreference text: what are the preposition relations between <NP_i> and <NP_j>? text: <P>"
NLI	MNLI	"mnli: does this hypothesis contradict, entail, or neutral with the premise? hypothesis: $\langle H \rangle$ premise: $\langle P \rangle$"
Coherence Scoring	GCDC	"GCDC coherence: what is the coherence score of the text (3 - high, 1 - low)? text: $\langle P \rangle$"
Coherence Scoring	CoheSentia	"CoheSentia coherence: what is the coherence score of the text (5 - high, 1 - low)? title: $\langle T \rangle$ text: $\langle P \rangle$"
MT	WMT14	"Machine Translation: what is the translation of the next text from language $< source\_language >$ to $< target\_language >$?: text in source language"
NER	Conll2003	"NER task: what is the entity recognition tagging of each token in the next text? $< extra\_id\_0 >$ token1 $< extra\_id\_1 >$ token2 ..."
POS	Conll2003	"POS task: What is the part of speech tagging of each token in the next text? $< extra\_id\_0 >$ token1 $< extra\_id\_1 >$ token2 ..."
Cohesion Reasoning	CoheSentia	"Cohesion reasoning: previous data: <d_i> new sentence: <si>. Task: is the new sentence cohesive in regard to the previous data? give a yes or no answer to each item "
Consistency Reasoning	CoheSentia	"Consistency reasoning: previous data: <d_i> new sentence: <si>. Task: is the new sentence consistent in regard to the previous data? give a yes or no answer to each item "
Relevance Reasoning	CoheSentia	"Relevance reasoning: previous data: <d_i> new sentence: <si>. Task: is the new sentence relevant in regard to the previous data? give a yes or no answer to each item "

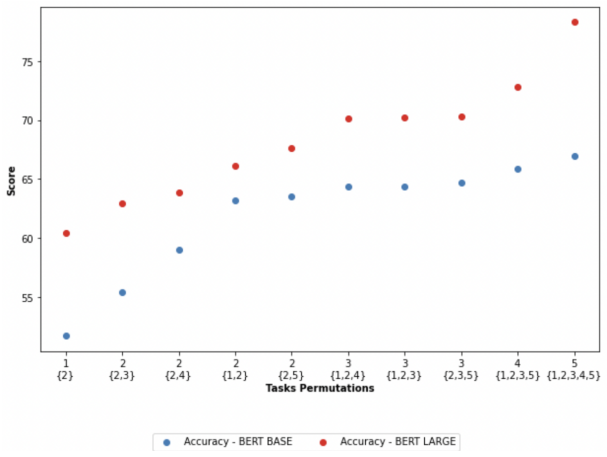
Table 9: Prompts for all tasks in this paper when using T5 model as the backbone model

Task	Dataset	Outputs
SRO	ROCStories	list of position markers $[Y_1, Y_2, ..., Y_N]$ (Y_i-position of the i_{th} sentence of the corresponding ordered sequence S_i in the shuffled input)
ISR	ROCStories	the index of the irrelevant sentence in the paragraph
DRR	PDTB3	"$\langle$connector$\rangle \rightarrow \langle l_1$ relation$\rangle \rightarrow \langle l_2 \rangle$"
NPE	TNE	the preposition
NLI	MNLI	Contradict / Entails / Neutral
Coherence scoring	GCDC	the score
Coherence scoring	CoheSentia	the score
MT	WMT14	the translated text
NER	Conll2003	"$< extra\_id\_0 > ner\_tag\_token1 < extra\_id\_2 > ner\_tag\_token2$ ..."
POS	Conll2003	"$< extra\_id\_0 > pos\_tag\_token1 < extra\_id\_2 > pos\_tag\_token2$ ..."
Cohesion reasoning	CoheSentia	Yes / No
Consistency reasoning	CoheSentia	Yes / No
Relevance reasoning	CoheSentia	Yes / No

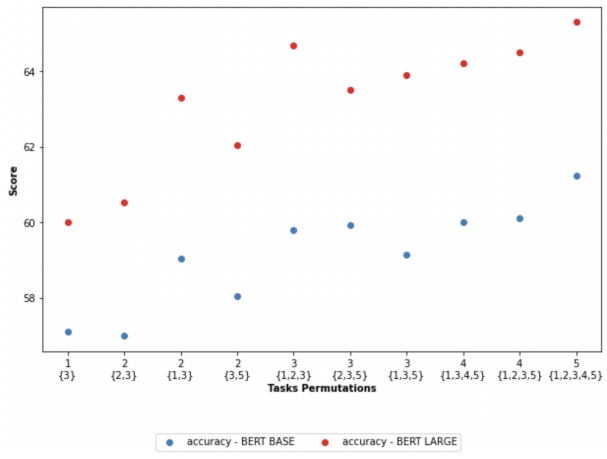
Table 10: Outputs for all tasks in this paper when using T5 model as the backbone model



(a) SRO

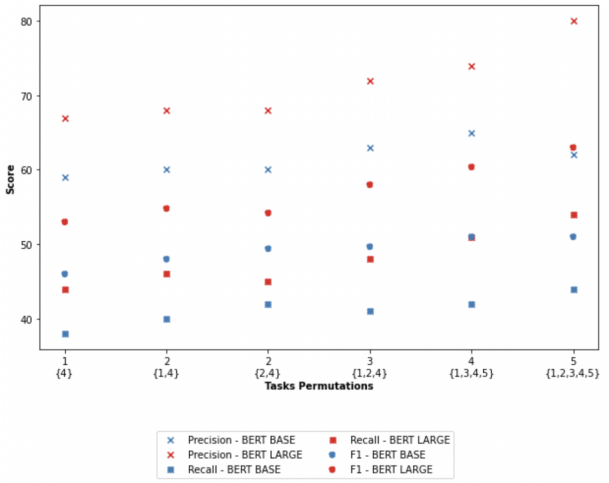


(b) ISR

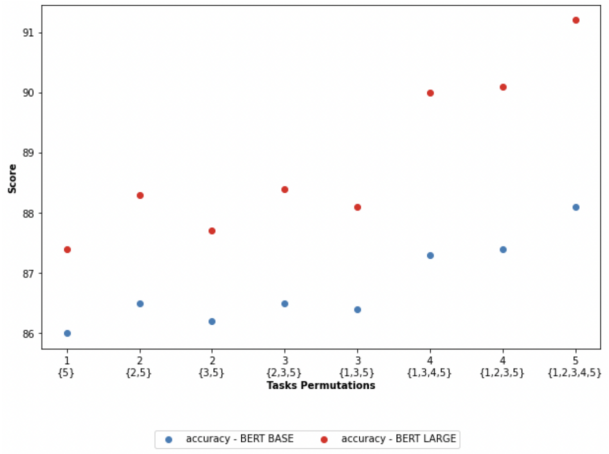


(c) DRR

Figure 9: Results for all tasks, for different permutations of tasks fine-tuned upon. The labels are the number of tasks and in curly brackets which tasks (1 - SRO, 2 - ISR, 3 - DRR, 4 - NPE, 5 - NLI)



(a) NPE



(b) NLI

Figure 10: Results for all tasks, for different permutations of tasks fine-tuned upon. The labels are the number of tasks and in curly brackets which tasks (1 - SRO, 2 - ISR, 3 - DRR, 4 - NPE, 5 - NLI)