A Novel Computational Modeling Foundation for Automatic Coherence Assessment

Aviya Maimon and **Reut Tsarfaty**
aviyamn@gmail.com
reut.tsarfaty@biu.ac.il
Department of Computer Science, Bar Ilan University

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 such as summarization, long-form question-answering, and more. 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 remedy this, in this work we employ the formal linguistic definition by Reinhart of what makes a discourse coherent, consisting of three conditions, *cohesion, consistency* and *relevance*, and formalize these conditions as respective computational tasks, which are in turn jointly trained. We evaluate this modeling approach on two human-rated coherence benchmarks: one of automatically-generated stories and one of real-world texts. Our experiments show 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. Our proposed computational framework thus paves the way for a more advanced, broad-coverage coherence assessment.

1 Introduction

The concept of *coherence* refers to the quality of textual flow, where sentences and paragraphs are logically connected, enabling a clear and understandable progression of ideas. Coherence is critical for many NLP tasks like summarization and question answering, directly impacting output accuracy and naturalness (Wu and Hu, 2018; Wang et al., 2021). As generative large language models (LLMs) become more prominent, coherence assessment has become essential for evaluating and improving text quality (Guan et al., 2021; Xu et al., 2018; Yi et al., 2019). This study introduces a novel approach to coherence assessment by leveraging linguistic theory alongside established NLP tasks.

Developing a robust model capable of accurately evaluating text coherence has been challenging due to several reasons. First, the scarcity of large-scale datasets specifically designed for training coherence assessment models, which are very difficult to produce (Lai and Tetreault, 2018; Maimon and Tsarfaty, 2023). Additionally, existing methods for coherence detection rely on a proxy task, and have not been able to accurately capture the multifaceted nature of coherence (Laban et al., 2021).

Our goal is to address these challenges by developing a model that assesses coherence effectively. Our vantage point is the theoretical foundation provided by Reinhart's linguistic framework, which defines coherence through three essential conditions: *cohesion, consistency*, and *relevance* (§2.2). Guided by these principles, we designed our model to integrate five carefully chosen NLP tasks that reflect these conditions. Specifically, the model incorporates sentence reordering (Lapata, 2003), implicit discourse relation detection (Miltsakaki et al., 2004), natural language inference (Dagan et al., 2005), NP enrichment (Elazar et al., 2022), and irrelevant-sentence detection (§3). Each task is supported by a large-scale dataset, and an associated multi-task model thereof enables capturing the multifaceted nature of coherence.

We hypothesize that jointly fine-tuning a model on these tasks will produce a model that effectively captures the core properties of coherence as delineated by the theoretical framework (§3). To test this, we develop a unified model jointly fine-tuned on the selected tasks that act proxies for the coherence conditions (§4). We then evaluate the model's coherence assessment capabilities through an additional fine-tuning stage on coherence-specific tasks (§5). We use two human-annotated benchmarks: the *Grammarly Corpus of Discourse Coherence (GCDC)* (Lai and Tetreault, 2018) for real-world texts across domains, and *CoheSentia* (Maimon and Tsarfaty, 2023) for synthetic texts.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5359–5377

Our results (§6) demonstrate that models jointly trained on our selected coherence proxy tasks outperform models without such training, achieving new SOTA performances on various coherence assessment benchmarks. We also show that training on tasks unrelated to coherence offers little to no value for coherence evaluation, validating our task selection. Finally, given shared knowledge across tasks, we expect the unified model to enhance individual proxy-task performances, and our experiments indeed confirm this for most of the tasks.

All in all, our approach presents a promising direction for future research to enable the development of tools that not only detect (in)coherence but also offer actionable insights into its causes. By integrating coherence assessment into text generation, we envision significant improvements in the quality and coherence of machine-generated texts.

2 Essential Preliminaries

2.1 Coherence in NLP

A key challenge in assessing coherence automatically lies in the elusive nature of coherence. While linguistic theories provide insights into the concept of coherence (Halliday and Hasan (1976); Joshi and Weinstein (1981); Givon (1995)), contemporary methods for automatically assessing coherence often rely on a proxy task, such as sentence reordering (Lapata, 2003), assuming this task will effectively capture the coherence mechanisms.

However, a single proxy task limits the models' ability to handle real-world texts effectively (Laban et al., 2021). Coherence varies across genres, contexts, and styles (Jurafsky, 2000), making it difficult for proxy tasks to generalize well across domains. Additionally, such 'coherence assessment' models are often evaluated through a downstream task (e.g., essay scoring) (Guinaudeau and Strube, 2013; Mesgar and Strube, 2016), which may introduce profound task-specific biases.

Finally, while coherence-scoring datasets do exist (Lai and Tetreault, 2018; Maimon and Tsarfaty, 2023), their limited size makes them insufficient for training a model solely on these resources.

2.2 Coherence in Linguistics

Our proposal is based on Reinhart's theory, which defines a text as *coherent* if and only if it meets three formal conditions: *Cohesion*, *Consistency*, and *Relevance*.

Cohesion The cohesion condition refers to the formal elements that link sentences.[1] According to Reinhart, a text is cohesive if each two sentences meet at least one of two conditions:

(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. For example:

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

Here, the underlined entities co-refer. Other referential links include prepositional links (Elazar et al., 2022) and bridging anaphora (Hou, 2021).

(2) Linked by a semantic 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 or contrast (Prasad et al., 2008). For example:

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

The sentences are cohesive due to the presence of the connector "So". These connectors can 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 meanings. 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 preceding 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 as folllow:

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

Despite being cohesive, 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 is:

> "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] To avoid confusion, 'cohesion' refers to surface elements (e.g., connectors, pronouns), while 'coherence' concerns the overall meaning and flow of ideas.

[2] Consistency has been explored by Honovich et al. (2021) to assess the reliability of automatically generated texts.

While the last sentence is cohesive and consistent, it is irrelevant to the context and topic of the story.

Reinhart's theory offers a comprehensive framework for studying text coherence, encompassing its fundamental aspects. Maimon and Tsarfaty (2023) applied Reinhart's framework for coherence benchmarking on GPT-generated text using human raters. In contrast, here we aim to directly model Reinhart's conditions through a set of tasks, aiming to predict these properties and develop effective models for overall coherence assessment.

3 Research Hypotheses and Tasks

To design the coherence assessment model, we start by mapping Reinhart's coherence conditions to computational tasks, employing a minimal set of NLP tasks designed to capture the key features of *cohesion, consistency*, and *relevance*. Our hypothesis is that a model jointly fine-tuned on *all* of these tasks will effectively learn to capture coherence features. Additionally, we anticipate that this unified architecture, when further fine-tuned on coherence assessment tasks, will outperform models fine-tuned on those coherence assessment tasks directly. To validate these hypotheses, we define five tasks that reflect the coherence conditions:

The Sentence Reordering (SRO) Task This 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 $(2) \rightarrow (4) \rightarrow (3) \rightarrow (1)$.

A model excelling at paragraph reconstruction should capture syntactic and semantic relationships between sentences, reflecting both *cohesion* and *consistency* (cf. Lin et al. (2011)).

The Implicit Discourse-Relation Recognition (IDRR) Task Given a pair of sentences (discourse units; DUs), the aim is to predict the discourse relation between them, reflecting notions such as cause and effect, comparison, and contrast (Pitler et al., 2009). For example, with the following input: "John worked all night. He slept all day today." the model is expected to detect a discourse relation 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 implicit prepositional links between noun phrase (NP) entities. Given two NP mentions in a text, it determines the existence of a prepositional relation and identifies the best preposition describing it $p(NP_1 NP_2)$ (or NONE is no relation exists).

For example, in the paragraph: "[Crown Princess Mary] of [Denmark] gives [birth] to a [male child]." (Elazar et al., 2022) there are 4 NPs (marked by [.]) and thus 12 potential NP pairs. Two examples of valid prepositional pairs and their relations are (1) *in*(birth, Denmark) and (2) *of*(birth, male child).

A model trained on this task captures referential links between sentences, 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 semantic relation between a premise-hypothesis pair as entailment/contradiction/neutral. 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 aims to identify irrelevant sentences in an otherwise coherent paragraph. Given a paragraph with N sentences, the model aims to find the irrelevant sentence. For example, given the following input: "(1) Rick is helpful. (2) He does the dishes. (3) He kicked his brother. (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.

Unlocking Coherence Modeling We employ a Multi-Task Learning (MTL) approach, jointly fine-tuning the model on those five coherence proxy tasks to capture all aspects of coherence. Our experiments explore two architectures: Classification-Based (BERT (Devlin et al., 2019)) and Generation-Based (T5 (Raffel et al., 2020)). The Classification-Based model utilizes a shared encoder across all tasks, with a dedicated classifier head for each task. In the Generation-Based model, unique prompts are crafted for each task and can be used either

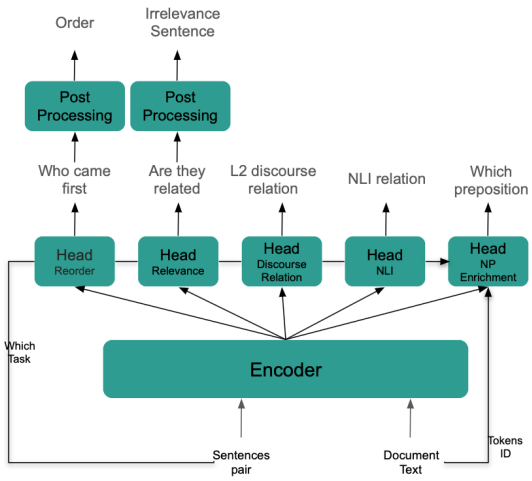


Figure 1: Illustration of the encoder-only model, which takes a pair of sentences as input (for most tasks) or a document with token IDs for the NPE task

independently or interleaved within the joint setup. Both methods utilize shared information during training to enhance overall coherence detection and improve performance on individual tasks.

4 Joint Models of Coherence

This section presents the proposed coherence models, which are based on the proxy tasks described in Section 3. We introduce both classification-based (BERT) and generation-based (T5) variants, along with the datasets and evaluation metrics used for assessment.

4.1 Classification-Based Modeling

The Classification-Based model is an MTL (Caruana, 1997) model where a shared encoder is used across all proxy tasks, with each task having its own distinct classification head which we detail shortly. Each classification head is responsible for predicting task-specific outputs, and the tasks are trained jointly (see Fig. 1). To mitigate catastrophic forgetting in MTL (Goodfellow et al., 2014), we implement an interleaved training strategy, where each batch contains samples from a single task that differ from those in the preceding batch. To address the diverse requirements of the proxy tasks, we design distinct classification-heads tailored to the nature and complexity of each task.

The NLI and IDRR tasks are well-suited for a simple classification head. In both cases, the model encodes a pair of sentences (or discourse units (DUs)) and predicts the relationship between them, an entailment label or a discourse relation label, respectively.

The reordering (SRO) and relevance (ISR) tasks require additional post-processing and we adopt a two-stage architecture. In both tasks, the input examples consist of either five sentences (SRO) or six sentences (ISR). For each pair of sentences, the first stage processes the pair, while the second stage aggregates the results into a unified prediction. In the first stage, each sentence pair $\langle S_i, S_j \rangle$ is encoded. For SRO, we adopt the architecture proposed by Shrimai Prabhumoye (2020), where a binary classification head predicts whether sentence S_i precedes or follows S_j. The overall predicted order is then generated in the second stage using the topological sort algorithm (Tarjan, 1976). For the ISR task, the binary classification head identifies relevant and irrelevant pairwise relationships, and in the second stage, the sentence with the lowest cumulative relevance score across all pairwise relationships is deemed irrelevant.

Finally, for NPE, an intricate task, a more complex model head is necessary. For this task, we extend the Bi-Affine architecture proposed by Dozat and Manning (2017) to predict prepositional relations between candidate NP pairs. To construct NP embeddings, we pool the token representations corresponding to each NP within the text. Given that the order of NPs within a pair is crucial for preposition detection, we assign distinct representations to each NP based on its position as either the first one (i.e. the anchor) or the second one (i.e. the complement). The model then predicts the appropriate preposition using these position-sensitive representations (See also Figure 4 in Appendix A).

4.2 Generation-Based Modeling

The Generation-Based model is a generative model fine-tuned on all the proxy tasks using task-specific prompts we design (Appendix G detail the specific prompts). We utilize a T5 encoder-decoder architecture, which allows simultaneous fine-tuning across multiple tasks. We employ an interleaved training strategy, where each batch contains samples from a specific task, different from the previous batch, to enhance multitask learning.

Here too, we describe the tasks based on the complexity of formulating prompts from the input and the extent to which post processing the model's output is required. For the SRO, ISR, and NLI tasks, prompt construction is straightforward, with output predictions directly corresponding to the input. The model processes the text to generate the correct sentence order for SRO, the premise-hypothesis re-

lationship for NLI, or the identification of the irrelevant sentence for ISR. For the IDRR and NPE tasks, the model requires more complex prompts and produces less straightforward outputs. In the IDRR task, the input is an argument pair, and the model uses a chain-of-thought method (Wei et al., 2023) to predict the discourse relation, following a three-stage structure: $\langle$connector$\rangle \rightarrow \langle l_1$ relation$\rangle \rightarrow \langle l_2$ relation$\rangle$.[3] In the NPE task, the model independently predicts prepositional relations for each NP pair, using the document text and a prompt that specifies the NP pair.

4.3 Datasets and Evaluation

Datasets For both the SRO and ISR tasks, we use the RocStories dataset (Mostafazadeh et al., 2016), which consists of five-sentence stories. For the SRO task, we randomly shuffle the sentences within each story. For the ISR task, each story is augmented with a single, randomly inserted sentence. The irrelevant sentence is selected from the entire RocStories dataset, with the constraint that it contains entities present in the target story.

For the IDRR task, we use the Penn Discourse TreeBank 3 (PDTB3) dataset (Miltsakaki et al., 2004; Prasad et al., 2008), specifically utilizing L_2 discourse senses.

For the NPE task, we employ the TNE dataset (Elazar et al., 2022), which consists of documents annotated with relations between every NP pair. The dataset includes approximately 190K nouns and 1M NP relations, covering 28 possible relations (including the 'no relation' class). A key advantage of this dataset is that it provides real-world, long-form paragraphs for evaluation.

Finally, for the NLI task, we use the MNLI dataset (Williams et al., 2018).

Evaluation Metrics For all tasks, we employ their standard evaluation metrics. Specifically, for the SRO task, we use both Perfect Match Ratio (**PMR**) (Chen et al., 2016) and **sentence accuracy** (Logeswaran et al., 2017). PMR measures the proportion of samples for which the predicted sequence exactly matches the ground truth:

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

Sentence accuracy, on the other hand, evaluates the proportion of sentences correctly placed in their absolute positions:

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

For the ISR, IDRR, and NLI tasks, we use **accuracy** as the primary evaluation metric. For the NPE task, we assess performance using **F1 score, precision, and recall**.

Table 1 summarizes datasets, metrics, and key statistics we use for each task. (See extended details in Appendix A).

5 Overall Coherence Assessment

5.1 Coherence Assessment Tasks

To confirm that the proposed joint models effectively capture coherence, we define two types of coherence assessment tasks:

The Coherence Scoring Task: Here we aim to evaluate the model's ability to predict the coherence score C for a given paragraph P, simulating the judgment of a human reader.

The Coherence Reasoning Task: To analyze specific coherence attributes, we leverage the coherence reasoning task, as proposed by Maimon and Tsarfaty (2023). In this task, the model is provided with a paragraph P and a new sentence s, and aims to make 3 binary decisions, whether s is *cohesive*, *consistent*, or *relevant* with respect to P.

5.2 The Coherence Scoring Task Setup

Models: A coherence scoring model aims to predict a 3-way or 5-way score for the text input. We develop two architectures for coherence scoring, Classification-based (BERT (Devlin et al., 2019)) and Generation-based (T5 (Raffel et al., 2020)), as described in the previous section. For both architectures, we begin with our joint model fine-tuned on the proxy tasks, and conduct an additional *second* fine-tuning for the coherence scoring task. In the Classification-based model, the second fine-tuning stage involves predicting the coherence score C for a given text P. In contrast, the Generation-based model generates the coherence score C by processing the input text combined with dataset-specific prompts. Example prompts and outputs can be found in Appendix G, while detailed hyperparameters settings are given in Appendix B.

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

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
IDRR	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 1: 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

Datasets and Evaluation: We evaluate our models 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). Inn contrast, CoheSentia features GPT-3 generated stories, spanning both fiction and non-fiction, with scores ranging from 1 to 5. Accordingly, the model performs 3-way classification for GCDC and 5-way classification for CoheSentia.

Dataset sizes and splits are detailed in Table 7.

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

Baselines: We evaluate our model's effectiveness by comparing its performance against current SOTA models on each dataset. For GCDC, Lai and Tetreault (2018) introduced the ParSeq model, using stacked LSTMs for sentence, paragraph, and document embeddings, followed by a coherence classifier. The latest SOTA by Liu et al. (2023) uses a multi-step approach involving graph structures, subgraphs, and GCN encoding. For CoheSentia, Maimon and Tsarfaty (2023) achieved SOTA with a prompt-based method using Flan-T5-large (the prompt adds a question at the beginning of the text).

We compare two models fine-tuned on the coherence scoring task: the joint models previously trained on coherence proxies (Ours-ALL), and the respective base models with no such fine-tuning (Ours-None). The comparison quantifies the effect of the proxy tasks on coherence assessment.

5.3 The Coherence Reasoning Task Setup

Models: Similarly to coherence scoring, we employ Classification- and Generation-based architectures, both built on our unified joint models fine-tuned on the coherence proxy tasks. During the second fine-tuning stage, the Classification-based model employs separate binary classification heads to predict each one of the coherence attribute for a given paragraph P and sentence s. In contrast, the Generation-based model produces outputs directly from prompted inputs (see Appendix G for the prompts).

Datasets and Evaluation: We evaluate our model on the CoheSentia corpus (Maimon and Tsarfaty, 2023), which contains automatically GPT3-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 (Ours-ALL) 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.

We also compare our model, which was fine-tuned on coherence proxy tasks, with one that was not (Ours-None), to quantify the impact of the proposed joint fine-tuning on coherence reasoning.

6 Results

6.1 Coherence Scoring Results

We first test the hypothesis that fine-tuning a model previously fine-tuned jointly on coherence proxy tasks improves coherence scoring over fine-tuning the base model. Table 3 shows that our jointly fine-tuned model (Ours-ALL) outperforms SOTA models, with 15% and 27% accuracy gains on GCDC and CoheSentia, respectively, underscoring the value of our approach and task selection, in capturing core aspects of coherence. To isolate the effect of proxy tasks, we compare fine-tuning our MTL model (Ours-ALL) with fine-tuning the base model without prior exposure to proxy tasks (Ours-None). As shown, the proposed MTL fine-tuning on proxy tasks significantly boosts performance.

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 2: Main Statistics on the Datasets for Coherence Scoring

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)

6.2 Coherence Reasoning Results

We analyze our joint models' success in assessing the coherence conditions (cohesion, consistency, relevance), by fine-tuning them on the coherence reasoning task. We compare our results (Ours-ALL) to SOTA from Maimon and Tsarfaty (2023), who fine-tuned the Flan-T5 model with a simple prompt, as well as to a model that was fine-tuned on coherence reasoning tasks without prior MTL fine-tuning on coherence proxy tasks (Ours-None).

Table 4 shows 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. Maimon and Tsarfaty (2023) noted that detecting relevance is more challenging than consistency, which is harder than cohesion. While our model still shows a disparity in the model's capabilities for identifying these different attributes, we have significantly narrowed this gap.

6.3 Cross-Domain Generalization

To assess generalizability across different domains and writing styles, we fine-tune the model on one dataset (GCDC or CoheSentia) and evaluate coherence on the other. Table 5 presents results for our MTL model (Ours-ALL) in Generation-Based settings 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.

6.4 Task-Specific Results

We evaluate the task-specific performance of models trained with either individual (Ours-Individual) or joint fine-tuning (Ours-ALL) on proxy tasks, using both the Classification- and Generation-Based variations. Results are in Table 6, alongside comparisons to current SOTA on these benchmarks.

Our findings show that joint fine-tuning across all tasks consistently surpasses individual fine-tuning, particularly in the SRO, ISR, and IDRR 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. An exception is the NLI task, where our model performs below SOTA. We conjecture that this discrepancy is due to SOTA results being achieved with T5-XXL (11B parameters), significantly larger than our backbone model, T5-large.

Additionally, our findings consistently show that Generation-Based models outperform the Classification-Based ones.

7 Analysis

7.1 The Impact of Task Selection

A hypothesis may be raised, that the joint ALL model outperforms the NONE model merely due to its added complexity, regardless of the nature of the tasks used (i.e., tasks reflecting coherence conditions). To refute this, we compare our model fine-tuned on the coherence proxy tasks with a model that has been fine-tuned on tasks unrelated to coherence. The tasks that are orthogonal to coherence are: (i) Part-of-Speech (POS) tagging: 14k instances from the CoNLL2003 dataset (Sang and Meulder, 2003); (ii) Named Entity Recognition (NER): also using the CoNLL2003 dataset; and (iii) Machine Translation (MT): 15k instances from the WMT14 dataset (Bojar et al., 2014).

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

For these experiments, we employed a Generation-Based model, specifically T5-large, as the backbone model, and used distinct prompt and output designs for each task. For NER and POS, we adapted the "Sentinel + Tag" architecture by Raman et al. (2022). Detailed prompts and sample outputs are in Appendix G.

Using our primary experimental protocol, we evaluated the fine-tuned models on the GCDC and CoheSentia benchmarks through coherence scoring fine-tuning. As shown in Table 3, fine-tuning on unrelated tasks yielded minimal improvements over the baseline (Controlled-nonCoherence) and is significantly underperformed compared to our final joint model (Ours-ALL). This highlights the critical importance of selecting coherence-specific proxy tasks for effective coherence detection, as unrelated tasks can hinder performance.

7.2 The Effect of Different Tasks on the Overall Coherence Scoring

To examine how fine-tuning on diverse subsets of coherence proxy tasks affects coherence scoring, we fine-tune models on various combinations of these tasks, and then perform final fine-tuning and evaluation on the coherence scoring task.

Figure 2 shows the impact of fine-tuning proxy coherence tasks on coherence scoring performance. Models fine-tuned on any one of the coherence proxy task outperform those without fine-tuning (Ours-None), highlighting their effectiveness. Performance generally improves with the addition of more tasks, particularly after three, suggesting cumulative benefits. Notably, fine-tuning with NLI significantly boosts performance, likely enhancing the model's ability to capture consistency, which is essential for coherence assessment. Furthermore, ISR fine-tuning is more impactful when combined with other tasks. These findings emphasize the importance of task selection and interaction during fine-tuning for optimal coherence scoring.

7.3 The Effects of Different Tasks on One Another

We also investigate how fine-tuning on diverse subsets of coherence proxy tasks influence the performance of individual proxy tasks. The model was trained on different task combinations with increasing numbers of tasks and evaluated on each task separately. Figure 3 shows the empirical results.

Figure 3 shows consistent performance gains in the Sentence Reordering (SRO) task for BERT models as more tasks are jointly fine-tuned (see Appendix F for other tasks). This supports our hypothesis that fine-tuning on coherence proxy tasks facilitates knowledge transfer and promotes learning of shared, generalizable representations.

The impact of specific tasks varies; for example, IDRR has a minimal impact on SRO, likely due to limited training data, whereas NPE substantially improve SRO performance. The ISR task notably improves performance on other tasks. We thus emphasize the introduction of this self-supervised ISR task and advocate for its exploration in future research to enhance coherence assessment.

Regarding the model size, overall performance trends are similar for both BERT-base and BERT-large models, suggesting that the influence of specific tasks remains consistent regardless of the size.

7.4 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

Model	SRO		ISR	IDRR	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 6: Results for all proxy tasks compared to SOTA performances. The SOTA model for SRO is ReBART (Basu Roy Chowdhury et al., 2021), for IDRR is Contrastive Learning (Long and Webber, 2023), for NPE is TNE (Elazar et al., 2022) and for NLI T5-11B (Raffel et al., 2020)

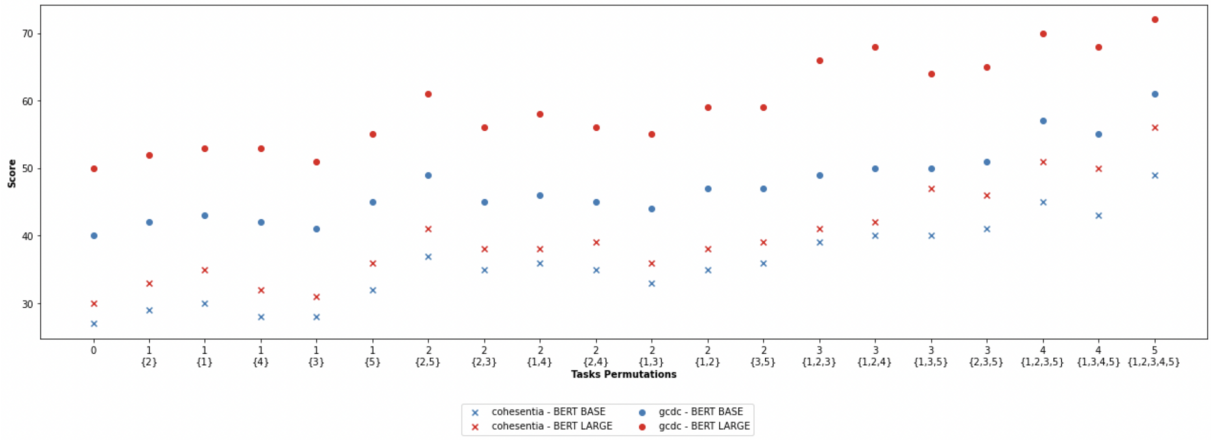


Figure 2: Accuracy for Coherence Scoring Task for both GCDC and CoheSentia with different proxy coherence task-subsets. The labels are tasks IDs (1-SRO, 2-ISR, 3-DRR, 4-NPE, 5-NLI)

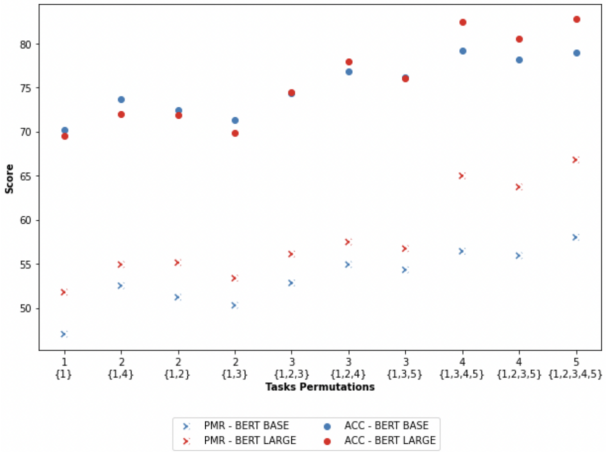


Figure 3: Results for SRO task, for different subsets of coherence tasks fine-tuned upon. The labels are the number of tasks and in curly brackets which tasks (1 - SRO, 2 - ISR, 3 - IDRR, 4 - NPE, 5 - NLI)

(Ours-None).

For CoheSentia, previous SOTA models tend to favor extreme scores, likely due to imbalances in the training data. In contrast, our model demonstrates greater robustness by predicting a more balanced distribution of scores. Appendix E provides qualitative examples of this behavior for both CoheSentia and GCDC, further highlighting our model's ability to achieve a more evenly distributed scoring pattern compared to existing approaches.

8 Conclusion and Future Work

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

Limitations

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

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 joint 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; Maimon and Tsarfaty, 2023). Additionally, Goldman et al. (2024) argue that long-context evaluation has not been properly addressed, as it involves two distinct axes: scope and dispersion. Based on their definitions, coherence evaluation can be categorized as a task with both high scope and high dispersion, making it particularly challenging to evaluate for long contexts. This dual nature highlights not only the difficulty but also the importance of coherence evaluation as a critical component of assessing LLM capabilities.

Our current study focused on short texts (≤ 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.

Acknowledgments

The authors would like to thank Valentina Pyatkin for fruitful discussions and sound advice. This work was funded by the European Research Council (ERC-StG grant number 677352), the Israeli Science Foundarion (ISF grant 670/23), the Ministry of Science and Technology (MOST grant number 3-17992) and the Israeli Innovation Authority (IIA, KAMIN grant), for which we are grateful.

References

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

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.

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.

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.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. *Preprint*, arXiv:2407.00402.

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.

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.

Daniel Jurafsky. 2000. Speech and language processing.

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.

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.

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.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *Preprint*, arXiv:cs/0306050.

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

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

Ziao Wang, Xiaofeng Zhang, and Hongwei Du. 2021. Building the directed semantic graph for coherent long text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2563–2572.

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.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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 1.

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.

Algorithm 1 Topological Sort Algorithm

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
while S is not empty do
    remove a node n from S
    add n to L
    for each node m with an edge e from n to m
do
        remove edge e from the graph
        if m has no other incoming edges then
            insert m into S
        end if
    end for
end while
if graph has edges then
    return error (graph has at least one cycle)
else
    return L (a topologically sorted order)
end if
```

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.

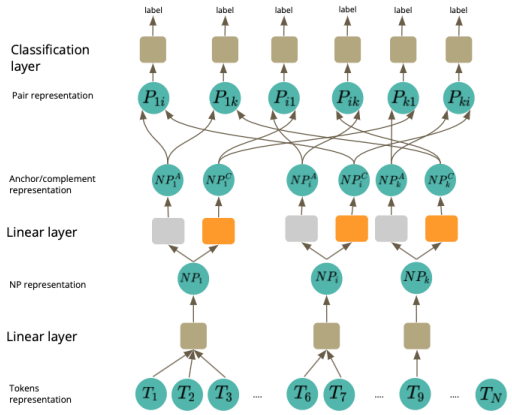


Figure 4: Illustration of the token head which contains several stages: starting with (1) embedding for each token in the text, (2) creating an embedding for each NP when it acts as the complement and the anchor separately, (3) a representation for each NP pair and finally (4) a classification layer

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

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

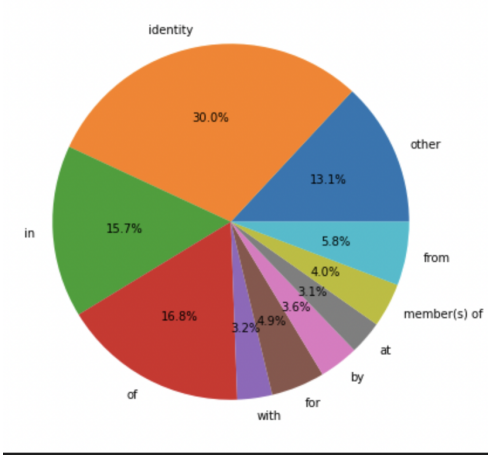


Figure 5: Distribution of the main prepositions in the NP Enrichment test set

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 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, IDRR, 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).

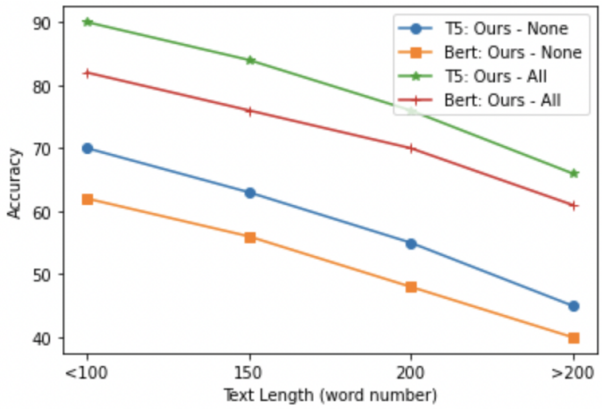


Figure 6: Accuracy For GCDC based on number of words

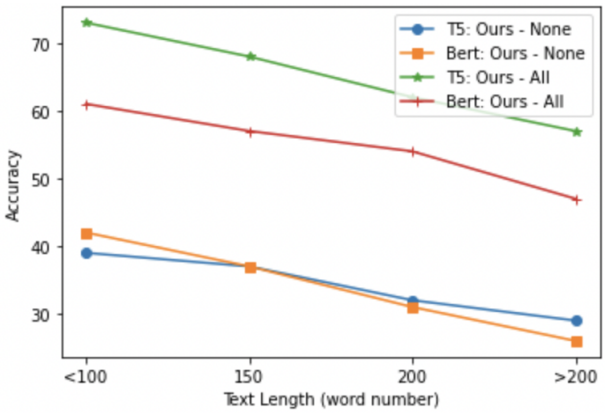


Figure 7: Accuracy For CoheSentia based on number of words

C Coherence Assessment Datasets Statistics

Dataset sizes and splits are detailed in Table 7.

D Text Length vs. Coherence Score

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

E Qualitative Analysis

Figure 8a and Table 8 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 9 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.

F 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.

G T5 Prompts and Outputs for Different Tasks

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

In Table 11 we detail the various outputs used for fine-tuning T5 models on all explored tasks in this work.

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 7: Main Statistics on the Datasets for Coherence Scoring

"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

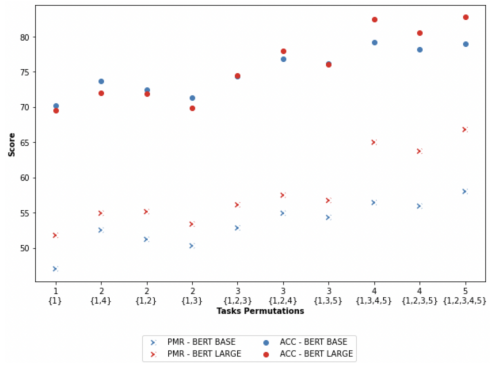
Figure 8: Sample Texts for coherence scoring tasks: GCDC & CoheSentia benchmarks

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

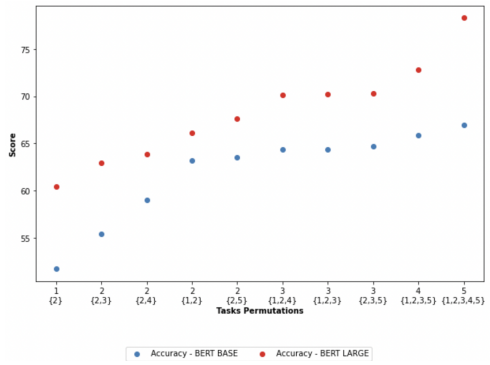
Table 8: Predicted Coherence scores for the text in Figure 8a

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

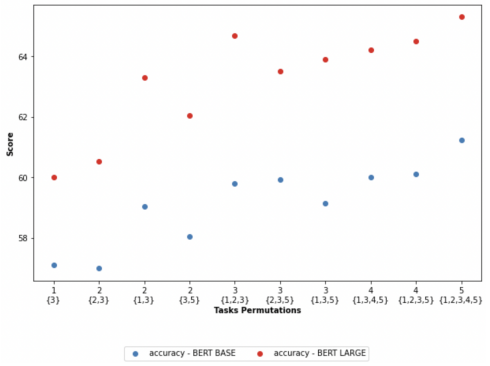
Table 9: Predicted Coherence scores for the text in Figure 8b



(a) SRO

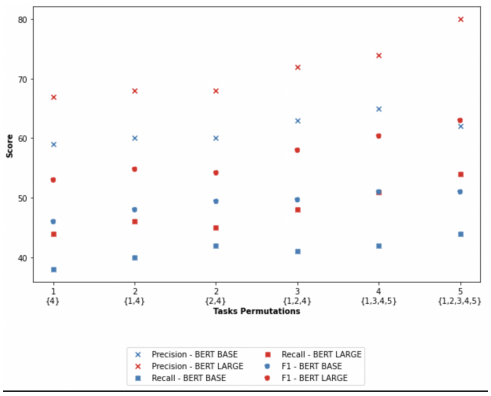


(b) ISR

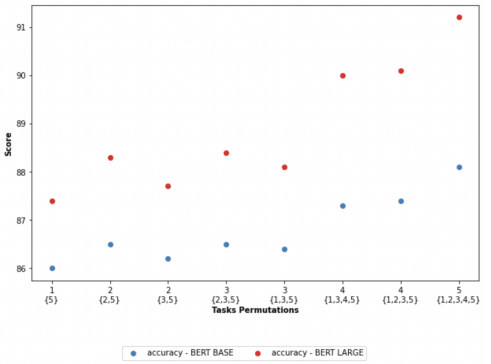


(c) DRR

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 - IDRR, 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 - IDRR, 4 - NPE, 5 - NLI)

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: ...$\langle S_N \rangle$"
IDRR	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 10: 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
IDRR	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 11: Outputs for all tasks in this paper when using T5 model as the backbone model