Incorporating Distributions of Discourse Structure for Long Document Abstractive Summarization

Dongqi Pu **Yifan Wang** **Vera Demberg**
Department of Computer Science
Department of Language Science and Technology
Saarland Informatics Campus, Saarland University, Germany
{dongqipu,yifwang,vera}@lst.uni-saarland.de

Abstract

For text summarization, the role of discourse structure is pivotal in discerning the core content of a text. Regrettably, prior studies on incorporating Rhetorical Structure Theory (RST) into transformer-based summarization models only consider the nuclearity annotation, thereby overlooking the variety of discourse relation types. This paper introduces the 'RSTformer', a novel summarization model that comprehensively incorporates both the types and uncertainty of rhetorical relations. Our RST-attention mechanism, rooted in document-level rhetorical structure, is an extension of the recently devised Longformer framework. Through rigorous evaluation, the model proposed herein exhibits significant superiority over state-of-the-art models, as evidenced by its notable performance on several automatic metrics and human evaluation.[1]

1 Introduction

For writing a good summary of a long document, it is of paramount importance to discern the salient information within the text and to comprehend the intricate interconnections among its various components. Contemporary leading-edge systems for abstractive (long) text summarization employ Transformer (Vaswani et al., 2017) encoder-decoder architecture (Zaheer et al., 2020; Guo et al., 2022). These sequence-to-sequence (seq2seq) models first transform the source document into a high-dimensional content representation and then decode the predicted summary conditioned on the representation (Belinkov and Bisk, 2018; Xu and Durrett, 2019; Cao and Wang, 2022; Balachandran et al., 2021). It has been demonstrated in the past that such an architecture does a poor job of digging high-level discourse structure during the encoding phase (Lin et al., 2019; Zhang et al., 2020; Koto et al., 2021a; de Wynter et al., 2023). However, discourse structure is very important for deciding what to include vs. not to include in the summary (Marcu, 1997, 1999, 1998; Zhong et al., 2020). Given that previous work has indicated that the performance of neural language models can be enhanced through the incorporation of latent structure information (Ettinger, 2020; Miaschi et al., 2020; Qian et al., 2021; Pu and Sima'an, 2022), we will here explore the integration of discourse relation structure into the Longformer model (Beltagy et al., 2020); this architecture has been shown to be particularly suitable for encoding long input texts.

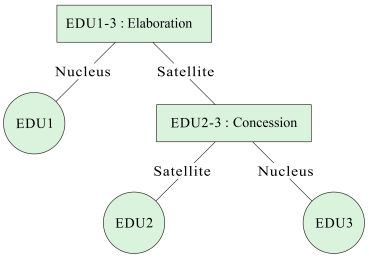

Figure 1: An example of RST tree: [*Rhetorical structure theory (RST) is a theory of text organization.*]$^{\text{EDU1}}$ [*Although the RST structure is difficult to annotate,*]$^{\text{EDU2}}$ [*there are still many scholars who have studied it.*]$^{\text{EDU3}}$

Rhetorical Structure Theory (RST) serves as a discourse framework designed to articulate the interrelationships among sentences at the document level. This framework distinguishes a plethora of coherence relations delineating the manner in which two text segments are interconnected (e.g., one segment might give a reason for a claim made in another segment, or alternatively, two segments may contrast with one another). Moreover, RST distinguishes between paratactic relations, where both segments carry equivalent discourse importance, and hypotactic relations, which classify the segment of greater centrality to the overarching discourse structure as the 'nucleus' and the less central one as the 'satellite'. Figure 1 shows a simple example of an RST tree. In this instance, EDU1 serves

[1]The project information can be accessed by visiting: https://dongqi.me/projects/RSTformer.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5574–5590

as the nucleus of the elaboration relation, whereas the combination of EDUs 2 and 3 constitutes the satellite of said relation. Furthermore, we can see that EDU3 assumes a more central role within the concession relation, hence it is marked as its nucleus, while EDU2 holds less important: if EDU2 was left out, the elaboration relation between EDUs 1 and 3 would still hold, but if EDU3 was removed, an elaboration relation between EDU1 and EDU2 would not hold, and the coherence would be lost. As has been recognized early on (Marcu, 1997, 1999), this discourse information can be effectively used in summarization tasks.

While there have been some previous attempts at integrating discourse structure into neural text summarization models, as seen in Gabriel et al. (2021); Dong et al. (2021); Xiao et al. (2020); Xu et al. (2020); Cohan et al. (2018), these approaches do not utilize relation labels and solely consider the 1-best RST tree obtained from preprocessing of a discourse parser. We argue that this leads to two significant issues: Firstly, information pertaining to relation type is overlooked, despite its known relevance to the summarization task. Secondly, there may be benefits in considering distributions over coherence relation labels, rather than limiting analysis to the 1-best results (Pu and Sima'an, 2022). One reason is that external discourse parsers are known to perform poorly on out-of-domain data (Atwell et al., 2022; Liu et al., 2021b; Gessler et al., 2021; Koto et al., 2021b; Liu et al., 2020; Nguyen et al., 2021), and may hence propagate errors into the summarization model. There is a subsequent risk that these errors will be incrementally amplified during back-propagation, thus potentially impairing the model's performance. A second reason is that there might inherently be several coherence relations holding at the same time (Yung et al., 2022), which might be beneficial to represent through the distributions of the discourse structure. Hence, we posit that the output of the RST parser holds greater significance when it not only provides the model with the n-best results but also conveys the remaining uncertainty associated with them.

In the remainder of the paper, we explore whether incorporating the labeled discourse relation structure with uncertainty, which can be understood as the distributions of discourse structure, into the attention mechanism can effectively augment the performance of neural summarization models. Our main contributions are as follows:

- We represent a generic approach for infusing labeled discourse relations with uncertainty into the encoder's self-attention layer of Longformer, wherein the self-attention heads are made to specialize in specific discourse categories. Additionally, our modules are orthogonal to the choice of the underlying encoder-decoder Transformer-based architecture, thereby enabling them to be seamlessly incorporated into other advanced models.
- We provide empirical evidence supporting the notion that conveying uncertainty and introducing labeled discourse relations to the Transformer are complementary actions, both significantly contributing to the enhancement of the final performance. Our model also surpasses current state-of-the-art models across multiple evaluation metrics.
- Quantitative and qualitative analyses show that our model exceeds the baseline model in both novel word generation and factual consistency checking. Furthermore, our model comes closer to human answers in terms of sentence alignment and overall generation quality.

2 Related Work

2.1 Text Summarization with RST

Rhetorical Structure Theory offers a structured paradigm for describing how various discourse units relate to one another in a text. The RST tree structure, as illustrated in Marcu (1997) and Louis et al. (2010), can serve as a valuable tool for content selection in the process of summarization.

For instance, Kikuchi et al. (2014) characterize the dependencies between sentences by constructing RST trees and pruning the parts that are marked as 'satellites' while preserving the important content ('nucleus') of the document as predicted summaries. Although RNN-based models are sometimes argued to be sufficient in implicitly learning discourse and semantic relations, Liu et al. (2019)'s work underscores the value of explicitly integrating RST trees into the summarization model, thereby highlighting the significance of discourse relation for the neural summarization network. It is also worth noting that while the attention mechanism can more effectively uncover discourse relations without explicit training, it tends to unearth only superficial discourse structure and is often prone to mistakes (Vig and Belinkov, 2019; Sachan et al.,

2021; Xiao et al., 2021; Huber and Carenini, 2022; Davis and van Schijndel, 2020).

Although attention-based models excel in executing downstream tasks such as summarization, the explicit incorporation of discourse relations can yield additional benefits. Work highly related to ours includes the model of Xiao et al. (2020), which improves the performance of an extractive summarization model by transmuting the RST structure into a dependency tree and explicitly integrating it into the computation of the attention mechanism. Follow-up works Xu et al. (2020) and Dong et al. (2021) further confirm the influence of RST structure on improving attention mechanism by incorporating discourse structure into a transformer-based model and a graph neural network model for the summarization task, respectively. However, all of these neural strategies apply the one-best structure derived from an external discourse parser.

2.2 Text Summarization with Longformer

The Longformer model (Beltagy et al., 2020), based on a sparse attention mechanism, is considered to be an effective means for processing long documents. Its essence is to make each token only pay attention to a window of a certain size, so that the time complexity of the model is reduced from a quadratic correlation with the text length to a linear correlation. Longformer-related models have since been employed in several summarization tasks (e.g., Zhang et al., 2022; Otmakhova et al., 2022; Elaraby and Litman, 2022; Xie et al., 2022; Pu et al., 2022).

At the same time, there have also been recent attempts at integrating text structure information with the Longformer model in summarization tasks. Huang and Kurohashi (2021) first employ the Longformer to encode input documents and propose an extractive summarization model based on a heterogeneous graph of discourse and coreference relations. Liu et al. (2021a) extend the Longformer to model different types of semantic nodes in the original text as heterogeneous graphs and directly learn relations between nodes. Specifically, they treated tokens, entities, and sentences as different types of nodes, and the multiple sparse masks as different types of edges to represent relations (e.g., token-to-token, token-to-sentence). Elaraby and Litman (2022) improve the performance of the strong baseline Longformer by integrating argument role labeling into the summarization process to capture the argumentative structure of legal documents. Ruan et al. (2022) and Cao and Wang (2022) enhance extractive and abstractive summarization tasks, respectively, by introducing the text's hierarchical structure (e.g., section title) into the Longformer model.

3 Proposed Approach

In the realm of document discourse parsing, the performance of the RST parser leaves much to be desired (Yu et al., 2022; Nguyen et al., 2021; Liu et al., 2021b), with parsing performance deteriorating in conjunction with escalating document complexity. Merely passing the 1-best RST tree risks imparting misleading information to the summarization model.

Inspired by Pu and Sima'an (2022), the approach to alleviating the aforementioned problems is that we retain uncertainty inside the parser, which can convey the parser's confidence in each discourse relation. Furthermore, we contend that discourse relation labels (types) can provide more fine-grained labeled probability distributions that can assist attention heads of the Transformer-based model to capture the importance of different discourse units. This in turn would contribute to a more precise estimation of the context vector and can enhance the quality of source document encoding. Discourse parsers tend to be more precise (and have more peaked probability distributions) for local coherence relations, which span only a short amount of text, compared to global relations spanning large portions of a text. This aligns well with the dilated (yet still limited) sliding window attention mechanism of the Longformer (Beltagy et al., 2020). We, therefore, integrate the probability distributions over local coherence relations into the attention window w of the Longformer.

3.1 RST Tensor with Labeled Distributions

The discourse-driven neural seq2seq summarization task can be modeled as follows:

$$P(t|s,d) \approx \prod_{i=1}^{T} P(t_i|t_{<i}, \mathrm{encode(s,}d)) \quad (1)$$

In the above equation, s, t, and d denote the source, target sequence, and discourse representation, respectively. T signifies the target sequence length and $\mathrm{encode}(\cdot)$ represents the encoder of the summarization model. Previous research (Xu et al., 2020; Cohan et al., 2018; Dong et al., 2021; Li

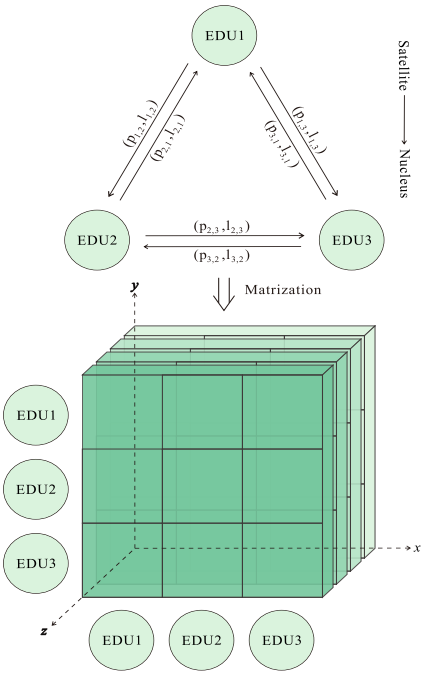

Figure 2: Labeled discourse distributions

et al., 2020; Chen and Yang, 2021) has confirmed that the probability of generating appropriate summaries by incorporating d into the model's encoder is significantly greater than the probability of generating proper summaries without the incorporation of d.

Our main idea is to find a better method to incorporate discourse structure d. To inject discourse structure, we first apply a 'matrixization' approach to represent the discourse structure and produce a compact tensor representation appropriate for the Longformer model (Pu and Sima'an, 2022).

Figure 2 illustrates by an example how we convert the graph of all potential RST relations (including the n-best RST trees present within the graph) into a three-dimensional labeled discourse distribution (LDD) tensor. The x-axis and y-axis of the tensor represent the elementary discourse units (EDUs) in the source document, while the z-axis represents the type of discourse relation. Each point represents a confidence value $p(edu_i, edu_j, l) \in [0, 1] \subseteq \mathbb{R}$ of an elementary discourse unit edu_i connecting to another elementary discourse unit edu_j from source text via the relation l. It should be noted that the generation of the LDD tensor should meet the conditions: 1) $p(edu_i, edu_i) = 0$, as no unit is dependent on itself; 2) we only extract the relation probability of nucleus units, since nucleus EDUs are more central to the text and should be given more attention. In the example shown in Figure 1, we only extract the discourse relation probabilities of EDU1 and EDU3.

3.2 RST Sparse Attention

We propose a novel Longformer-Encoder-Decoder Summarization model: RSTformer, which incorporates LDD into each layer of the Longformer encoder in a discourse-aware manner. Given that each encoder layer shares an identical configuration, Figure 3 displays one layer architecture of our proposed model.

The standard dilated sliding window attention layer of Longformer employs a multi-head fixed-size window attention mechanism. For a pre-specified window size w, each token attends to $\frac{1}{2}w$ tokens on either side. For an input sequence of length T, the input of dilated sliding window attention heads in the RSTformer layer comprises the hidden representation tensor $\mathrm{X} \in \mathbb{R}^{T \times d_{model} \times h}$ and labeled discourse distribution tensor $\mathrm{LDD} \in \mathbb{R}^{T \times d_{model} \times h}$, where d_{model} represents the size of the hidden representation and h denotes the number of attention heads.

As usual in multi-head self-attention, we multiply the text feature representation tensor with $\mathrm{q}, \mathrm{k}, \mathrm{v} \in \mathbb{R}^{d_{model} \times d \times h}$ to obtain the corresponding $\mathrm{Q} \in \mathbb{R}^{T \times d \times h}$, $\mathrm{K} \in \mathbb{R}^{T \times d \times h}$, and $\mathrm{V} \in \mathbb{R}^{T \times d \times h}$ matrices, where $d = d_{model}/h$. Subsequently, the attention weight matrix is obtained by:

$$\mathrm{S} = \frac{\mathrm{Q} \cdot \mathrm{K}^{\top}}{\sqrt{d}} \tag{2}$$

Longformer utilizes two sets of projections, $\mathrm{Q_s}$, $\mathrm{K_s}$, $\mathrm{V_s}$ to compute the attention scores of sliding window attention, and $\mathrm{Q_g}$, $\mathrm{K_g}$, $\mathrm{V_g}$ to compute attention scores for global attention. Notably, $\mathrm{Q_g}$, $\mathrm{K_g}$, $\mathrm{V_g}$ are all initialized with values that match $\mathrm{Q_s}$, $\mathrm{K_s}$, $\mathrm{V_s}$ respectively. The dilated sliding window attention operates by calculating a fixed number of the diagonals of $\mathrm{QK}^{\top}$ through sliding chunks query-key multiplication. This process yields a resulting tensor $\mathrm{S} \in \mathbb{R}^{T \times w+1 \times h}$. Similarly, LDD and V adopt the same *chunk* method as employed by Longformer to acquire the sliding window attention matrix.

It should be noted here that we inject the sliding window attention tensor S obtained from the preceding computation by element-wise multiplication with the LDD tensor:

$$\mathrm{S} \odot \mathrm{LDD} \tag{3}$$

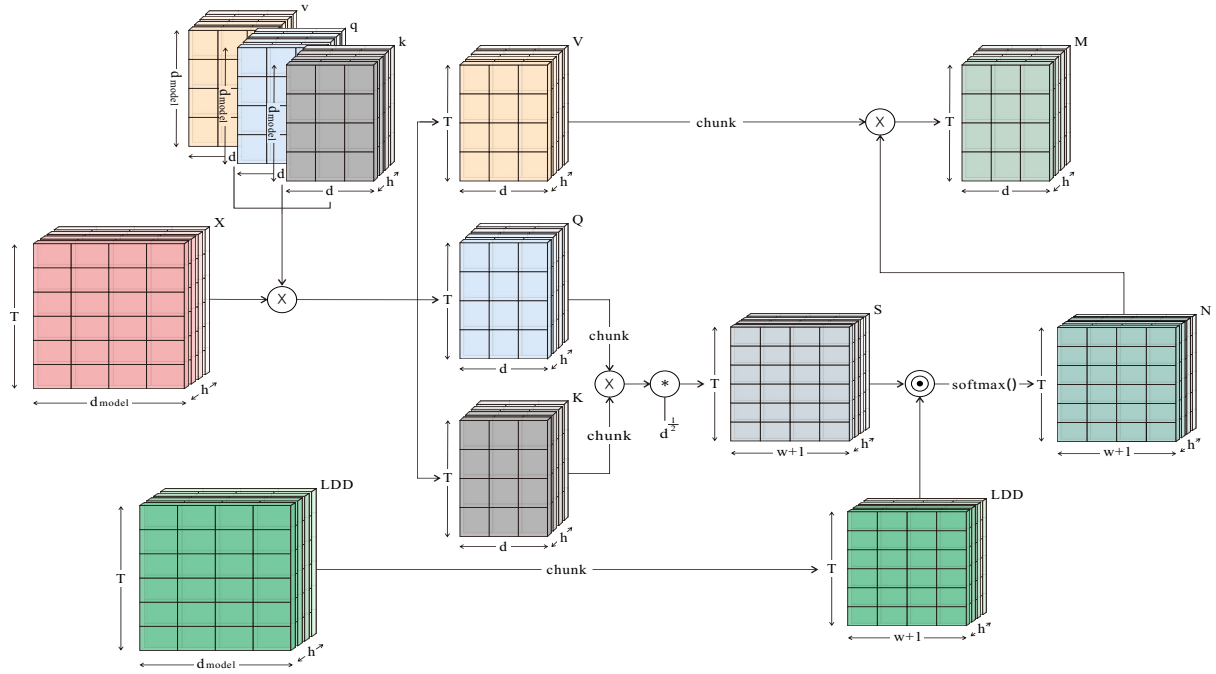

Figure 3: Model architecture: we show a schematic diagram of incorporating LDD tensor into the attention layer of the model. Specifically, X is text embedding matrix, and LDD is incorporated with attention matrix S in the form of element-wise multiplication. In order to ensure the consistency of matrix shape, we also apply an identical *chunk* method as Longformer in LDD.

The motivation behind employing element-wise multiplication is to allow the learning parameters of the attention mechanism 'dynamically' to optimize the summarization objective but also diverge the least from the parser probabilities in discourse distribution (Pu and Sima'an, 2022). The estimation of attention weights is adjusted to align with the utility of discourse relations for the ultimate summarization task.

Following, the obtained weights are further processed using the softmax function to derive the final tensor representing the discourse-infused distribution:

$$\mathrm{N} = \mathrm{softmax}(\mathrm{S} \odot \mathrm{LDD}) \tag{4}$$

It should be emphasized that each attention head is assigned a different discourse matrix LDD_l for a specific relation l, This allocation enables heads to concentrate on and learn different discourse labels (Pu and Sima'an, 2022). In doing so, attention heads can be specialized and acquire a deeper understanding of the impact of discourse labels.[2]

Finally, the discourse-injected weights N are multiplied with the value matrix V to obtain the attention weights M for this layer and then transfer M to the next Longformer encoder layer for further computation.

$$\mathrm{M} = \mathrm{N} \cdot \mathrm{V} \tag{5}$$

4 Experiments and Analysis

4.1 Experimental Setup

Parser We employ an external RST parser called *DMRST* (Liu et al., 2021b, 2020) to automatically parse the source documents. The probability or uncertainty of discourse relations is extracted from the logits layer of the DMRST[3] model. In cases where DMRST fails to parse the source document, we simply skip the LDD generation process and proceed with the normal Longformer procedure.

Datasets We conduct our experiments on three recent long document summarization datasets: BookSum Chapter (Kryscinski et al., 2022), eLife (Goldsack et al., 2022), and Multi-LexSum (Shen et al., 2022). We choose these datasets because of their high heterogeneity and we want to investigate whether our approach can maintain adequate generalization performance across different data domains. Table 1 shows the statistics of the datasets.

[2]Appendix A details the grouping of discourse relations.

[3]https://github.com/seq-to-mind/DMRST_Parser

Dataset	Training	Validation	Test	Avg. Doc Words	Avg. Summary Words	Coverage	Density	Compression Ratio
BookSum Chapter	9600	1431	1484	3834.40	363.81	0.764	1.504	15.198
eLife	4346	241	241	10133.07	382.69	0.819	1.761	27.650
Multi-LexSum	3177	454	908	58210.99	547.04	0.926	3.394	95.390

Table 1: Datasets statistics

Coverage refers to the percentage of words in the summary that are from the source document. A higher coverage ratio indicates that a greater proportion of summary words are derived directly from the source text. It is mainly used to measure the degree of derivation of the summary from the text. *Density* is defined as the average length of the extracted segments to which each summary word belongs (Segarra Soriano et al., 2022). *Compression ratio* is defined as the ratio between the length of the source document and summary (Scialom et al., 2020).

Evaluation Metrics We evaluate the quality of different summarization systems using Rouge-{1, 2, L} score (Lin, 2004), BERTscore (Zhang et al., 2019), Meteor score (Banerjee and Lavie, 2005), {1, 2, 3, 4}-gram novelty (Kryściński et al., 2018), SummaC (Laban et al., 2022) and sentence alignment (Liu and Liu, 2021) as criteria for the model's effectiveness.

In detail, Rouge-{1,2} is mainly evaluated based on the co-occurrence of {1,2}-gram in summary, while the calculation of Rouge-L uses the longest common subsequence. BERTScore is used to compute the semantic similarity score of candidate sentences to reference sentences through contextual embedding. Meteor is an improvement based on BLEU (Papineni et al., 2002), which also considers the impact of sentence fluency and synonyms on semantics. {1, 2, 3, 4}-gram novelty indicates the capacity of the model to generate new words, rather than merely extracting words from the original text. SummaC detects semantic inconsistency by segmenting documents into sentence units and aggregating scores between sentence pairs.

Training and Inference Hyper-parameters for the baseline, proposal models, and ablation models are all kept identical. We adopt the same configuration as Longformer (Beltagy et al., 2020): All experiments are optimized using Adam (Kingma and Ba, 2014) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$, and weight decay = 0.1) with Adafactor (Shazeer and Stern, 2018), the number of warm-up steps is 1500, and the initial learning rate is set to $3e^{-9}$ with cosine learning rate schedule. We also apply NoisyTune (Noise lambda = 0.2) (Wu et al., 2022) for efficient fine-tuning. The size of the local attention window is $w = 1024$, and we choose cross-entropy as loss function.

During the training phase, we save the checkpoint with the highest Rouge-2 F1 score on the validation set as the final model. The experiments are all run for 30 epochs using a batch size of 1 with early stopping implemented. In order to prevent over-fitting, we set the dropout rate to 0.1 in all layers of the model. For model inference, we adopt a beam size of 4 with a length penalty of 2.0, and we set the no-repeat n-gram size to 3.

4.2 Results

The experimental results for each model are presented in Table 2. To estimate a lower bound in performance, we simply use the original document as the summary. Further trivial models include the Lead-3 model which simply picks the first three sentences of the document as the summary. Lead-K similarly extracts the first K sentences of the document, until a similar length as the reference summary is reached. Longformer and state-of-the-art (SOTA) models serve as our baseline and comparison models, respectively. The remaining two models are the models we proposed. RSTformer (w/o relations) refers to the model that preserves whether there are relations between EDUs and ignores the type of relations by summing the third dimension of LDD tensors. RSTformer (w relations) is the final model we propose, with the only difference being the inclusion of the impact of RST types.

Both RSTformer versions are found to outperform the baseline model on various measures. The higher scores reflect an improved choice of words (Rouge & Meteor scores), and also the semantics of the text (BERTscore).[4] The proposed model,

[4]The version of BERTscore we use comes from the original paper version (Zhang et al., 2019) with HuggingFace default API (https://huggingface.co/spaces/evaluate-metric/bertscore).

RSTformer, demonstrates robust generalization capabilities across different datasets, highlighting its promising potential in various summarization domains.

In most of our summarization experiments, we furthermore find that incorporating discourse structure with types provides better experimental results than the discourse distributions without types, even beating the SOTA model on our experimental datasets. This observation suggests that providing more discourse information, especially type distribution probabilities, is a promising approach.

Ablation Study We also define two additional control conditions to examine the impact of RST attention (LDD) on model performance:

- **Without Attention Calculation (WAC)**: We skip the previous calculation of attention weights, and directly replace attention weights with LDD tensor.
- **Random Identical Attention (RIA)**: We assign fixed random values to LDD tensor, regardless of the probability of discourse relations.

Table 3 shows that the RST attention cannot fully replace the calculation of the attention mechanism. Although the performance is significantly lower than the baseline model, its main noteworthy advantage is that it saves considerable computations and parameters. Experiments by introducing random noise demonstrate that random values do indeed negatively impact the model's performance. Furthermore, it also confirms the effectiveness of incorporating the probability distributions of discourse structure.

Human Evaluation To better analyze the effectiveness of our model, we randomly select 10 samples from the BookSum dataset and hire human annotators to conduct the human evaluation. The recruited annotators are all master's students or doctoral students with computer science-related or computational linguistics-related backgrounds. All annotators were compensated with the standard hourly salary set by the university. At the time of evaluation, we provide 3 candidate summaries for each source document, namely outputs from our final proposed model and baseline model, along with the ground truth summary. Each instance is assigned to 3 participants who are instructed to rate the faithfulness, informativeness, readability, and conciseness of the candidate summaries on a scale of 1 to 5. They are also supposed to give an overall rank of three summaries and identify which one is generated by humans. Detailed information regarding the human evaluation process can be found in Appendix B. Table 4 reports the human evaluation results.

For each human evaluation indicator, we compute the average value to represent whether the candidate system has good performance in that indicator. Best and Worst indicate the proportion of times a summary by a particular model is judged to be best or worst among the three options. While neural summarization models still exhibit a notable performance gap when compared to human-generated summaries, our proposed model consistently outperforms the baseline model across all metrics.

4.3 Analysis

Sentence Alignment We examine the alignment distributions of generated summaries to explore whether the improved model can be closer to human-summarized text (Liu and Liu, 2021). Our results are depicted in Figure 4 and Appendix C.

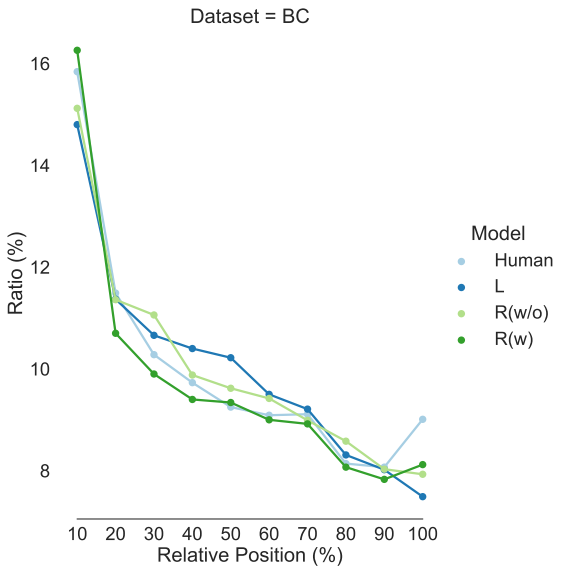

Figure 4: Sentence alignment distribution. L = Longformer, R(w/o) = RSTformer(w/o relations), R(w) = RSTformer(w relations), BC = BookSum Chapter.

From a broader perspective, the sentence alignment distribution of our proposed models is more closely aligned with that of human summarizers. In addition, the generated summaries produced by our models demonstrate a greater emphasis on the content of the second half of the document, resulting in summaries that are more comprehensive and

Dataset	Model	Rouge-1 F1	Rouge-2 F1	Rouge-L F1	BERTscore	Meteor
BookSum Chapter	Full article (lower bound)	13.742	4.019	13.421	0.805	21.299
	Lead-3	17.683	2.747	16.708	0.812	9.815
	Lead-K	29.149	4.641	28.034	0.805	24.091
	Longformer(baseline)	33.636	9.626	32.611	0.846	27.160
	RSTformer(w/o relations)	33.604	10.149	32.631	0.850	26.811
	RSTformer(w/ relations)	34.019	**10.275**†‡	**32.870**	**0.853**†‡	**27.473**‡
	SOTA model (Kryscinski et al., 2022)	**37.510**	8.490	17.050	0.156	-
	Our compared to baseline	+Δ0.383	+Δ0.649	+Δ0.259	+Δ0.007	+Δ0.313
	Our compared to SOTA	−Φ3.491	+Φ1.785	+Φ15.820	+Φ0.697	Φ-
eLife	Full article (lower bound)	6.893	2.327	6.675	0.831	13.864
	Lead-3	16.266	3.634	15.088	0.832	7.163
	Lead-K	37.188	7.971	35.151	0.832	25.331
	Longformer(baseline)	46.778	13.318	44.317	**0.855**	27.921
	RSTformer(w/o relations)	46.862	14.008	44.458	**0.855**	27.685
	RSTformer(w/ relations)	**48.696**†‡	**14.843**†‡	**46.129**†‡	0.847	**29.526**†‡
	SOTA model (Goldsack et al., 2022)	46.570	11.650	43.700	-	-
	Our compared to baseline	+Δ1.918	+Δ1.525	+Δ1.812	−Δ0.008	+Δ1.605
	Our compared to SOTA	+Φ2.126	+Φ3.193	+Φ2.429	Φ-	Φ-
Multi-LexSum	Full article (lower bound)	3.862	2.198	3.786	0.784	8.825
	Lead-3	16.135	6.387	15.421	0.770	9.538
	Lead-k	29.145	9.276	27.734	0.784	24.266
	Longformer(baseline)	45.751	21.272	43.131	0.865	33.282
	RSTformer(w/o relations)	46.424	22.730	43.978	0.867	33.808
	RSTformer(w/ relations)	46.421	22.888†‡	**43.979**	**0.867**‡	**33.941**
	SOTA model (Shen et al., 2022)	**53.730**	**27.320**	30.890	0.420	-
	Our compared to baseline	+Δ0.670	+Δ1.616	+Δ0.848	+Δ0.002	+Δ0.659
	Our compared to SOTA	−Φ7.309	−Φ4.432	+Φ13.089	+Φ0.447	Φ-

Table 2: Model performance. The bold numbers represent the best results with respect to the given test set. Δ and Φ represent the improvement of our model compared to the baseline and SOTA models, respectively. † and ‡ indicate statistical significance ($p<0.05$) against the baseline model via T-test and Kolmogorov-Smirnov test. Each result of the three distinct SOTA models is directly replicated from their original papers.

Dataset	Model	Rouge-1	Rouge-2	Rouge-L
BookSum Chapter	Longformer	33.636	9.626	32.611
	RSTformer(WAC)	31.956	8.772	31.049
	RSTformer(RIA)	32.881	9.067	31.899
eLife	Longformer	46.778	13.318	44.317
	RSTformer(WAC)	39.076	8.461	37.114
	RSTformer(RIA)	41.761	10.901	40.062
Multi-LexSum	Longformer	45.751	21.272	43.131
	RSTformer(WAC)	42.903	18.440	40.773
	RSTformer(RIA)	42.213	20.785	31.219

Table 3: F1 scores for ablation study

Candidate	Faithful	Informative	Readable	Concise	Best | Worst
Human	4.40	4.83	4.83	4.33	83.3% | 0.0%
Longformer	2.50	2.57	3.43	2.70	6.7% | 56.7%
RSTformer(w relations)	2.97	2.90	3.73	3.00	10.0% | 43.7%

Table 4: Human evaluation results

coherent in nature.

N-gram Novelty & Inconsistency Detection We also study the level of abstractiveness and factual consistency in the generated summaries. To evaluate the abstractiveness, we employed N-gram novelty as a measure to determine whether the model can generate words that are not present in the original text, rather than solely extracting content from the source document. For inconsistency detection, we utilize the latest SummaC method (Laban et al., 2022) for testing. Our results are shown in Figure 5 and Figure 6 respectively.

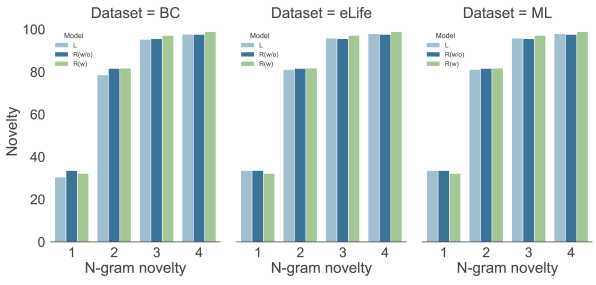

Figure 5: N-gram novelty. L = Longformer, R(w/o) = RSTformer(w/o relations), R(w) = RSTformer(w relations), BC = Booksum Chapter, ML = Multi-LexSum.

Compared with the baseline model, incorporating discourse information into the model does increase the ability of the model to generate novel words, especially evident in the context of 3-gram

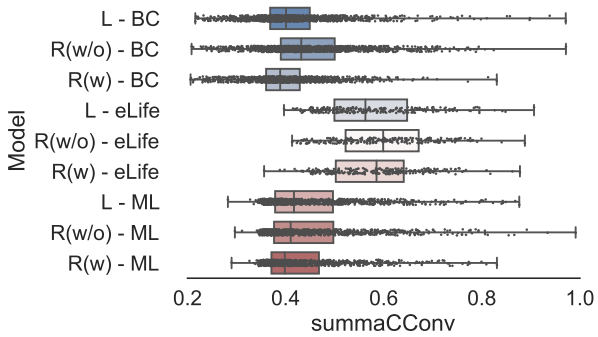

Figure 6: Consistency check. L = Longformer, R(w/o) = RSTformer(w/o relations), R(w) = RSTformer(w relations), BC = Booksum Chapter, ML = Multi-LexSum.

and 4-gram, the gap becomes more prominent. In addition, the proposed model also performs better than the baseline model in terms of model consistency checks. Due to the increased ability to generate creative words, the semantic coherence ability of the models incorporating typed discourse relations is lower than that of models without typed discourse relations.

5 Conclusion

This paper introduces a novel supervised discourse enhanced Longformer model. This strategy mainly improves the local attention mechanism in the Longformer model by leveraging the rhetorical structure as uncertainty distributions. The experimental findings provide strong evidence that the proposed approach is straightforward, and can effectively employ the discourse structure of source documents to improve the summary performance of Longformer. Furthermore, this strategy also has a high potential capability for application in other seq2seq natural language tasks.

6 Limitations

The present study has certain limitations that should be acknowledged. Firstly, the RST parsing task itself is known to be highly complex and challenging, and achieving high accuracy in this task is not guaranteed. Although we have utilized the most high-performing parser, there is still room for further improvement in the RST parsing performance, which could potentially enhance the downstream summarization task.

Another limitation pertains to the size of the data used for human evaluation. Due to the nature of long document summarization and the length of the original texts (often spanning several pages), scaling up the evaluation process, such as through crowd-sourcing, becomes difficult. Consequently, we are only able to evaluate a limited number of 10 documents, which may not be fully representative of the entire dataset.

Furthermore, another potential risk in our study is the limitation in obtaining an unlimited number of training samples. The data samples investigated are often small subsets of real-world data or may exhibit certain biases, which may not accurately reflect the distribution of real-world data. Although we have verified the effectiveness of our model using highly diverse and heterogeneous datasets from different domains, it is important to note that the model's performance on the specific dataset of interest may not be as robust as its performance on unseen real-world data.

Finally, both training and evaluating the models require significant computational resources. Despite our attempts to optimize the computation by replacing the original attention calculation with the RST attention tensor (as demonstrated in the ablation experiment), we have not achieved satisfactory results. The high computational costs pose a challenge, as they result in increased human and material resources required for the model.

7 Ethics Considerations

The datasets we use are all public, and our experiment processes have no privacy disclosure issues. As for human evaluation, all participants are voluntary and paid, and come from master or doctoral students with a background in computer science or computational linguistics, and all of them are proficient in English. They first need to read the instructions and evaluate without revealing which model generates which summary.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). We are grateful to the anonymous reviewers and area chairs for their exceptionally detailed and helpful feedback.

References

Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, and Malihe Alikhani. 2022. The change that matters in discourse parsing: Estimating the impact of domain shift on parser error. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 824–845, Dublin, Ireland. Association for Computational Linguistics.

Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, and Yulia Tsvetkov. 2021. StructSum: Summarization via structured representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2575–2585, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Shuyang Cao and Lu Wang. 2022. HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, Online. Association for Computational Linguistics.

Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. 2023. An evaluation on large language model outputs: Discourse and memorization. *arXiv preprint arXiv:2304.08637*.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021. Discourse understanding and factual consistency in abstractive summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 435–447, Online. Association for Computational Linguistics.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for*

Computational Linguistics: NAACL 2022, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Yin Jou Huang and Sadao Kurohashi. 2021. Extractive summarization considering discourse and coreference relations based on heterogeneous graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3046–3052, Online. Association for Computational Linguistics.

Patrick Huber and Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. *arXiv preprint arXiv:2204.04289*.

Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. Single document summarization based on nested tree structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320, Baltimore, Maryland. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021a. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021b. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726, Online. Association for Computational Linguistics.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Zhenwen Li, Wenhao Wu, and Sujian Li. 2020. Composing elementary discourse units in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019. Single document summarization as tree induction. 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 1745–1755, Minneapolis, Minnesota. Association for Computational Linguistics.

Ye Liu, Jianguo Zhang, Yao Wan, Congying Xia, Lifang He, and Philip Yu. 2021a. HETFORMER: Heterogeneous transformer with sparse attention for long-text extractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 146–154, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. Multilingual neural RST discourse parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021b. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan. Association for Computational Linguistics.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*.

Daniel Marcu. 1998. Improving summarization through rhetorical parsing tuning. In *Sixth Workshop on Very Large Corpora*.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 293:123–136.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. RST parsing from scratch. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625, Online. Association for Computational Linguistics.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dongqi Pu, Xudong Hong, Pin-Jie Lin, Ernie Chang, and Vera Demberg. 2022. Two-stage movie script summarization: An efficient method for low-resource long document summarization. In *Proceedings of The Workshop on Automatic Summarization for Creative Writing*, pages 57–66, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Dongqi Pu and Khalil Sima'an. 2022. Passing parser uncertainty to the transformer: Labeled dependency distributions for neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 41–50, Ghent, Belgium. European Association for Machine Translation.

Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernandez Astudillo. 2021. Structural guidance for transformer language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3735–3745, Online. Association for Computational Linguistics.

Qian Ruan, Malte Ostendorff, and Georg Rehm. 2022. HiStruct+: Improving extractive text summarization with hierarchical structure information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1292–1308, Dublin, Ireland. Association for Computational Linguistics.

Devendra Sachan, Yuhao Zhang, Peng Qi, and William L. Hamilton. 2021. Do syntax trees help pre-trained transformers extract information? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2647–2661, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Encarnación Segarra Soriano, Vicent Ahuir, Lluís-F. Hurtado, and José González. 2022. DACSA: A large-scale dataset for automatic summarization of Catalan and Spanish newspaper articles. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5931–5943, Seattle, United States. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *arXiv preprint arXiv:2206.10883*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. NoisyTune: A little noise can help you finetune pretrained language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2:*

Short Papers), pages 680–685, Dublin, Ireland. Association for Computational Linguistics.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2020. Do we really need that many parameters in transformer for extractive summarization? discourse can help ! In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 124–134, Online. Association for Computational Linguistics.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2021. Predicting discourse trees from transformer-based neural summarizers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152, Online. Association for Computational Linguistics.

Qianqian Xie, Jimin Huang, Tulika Saha, and Sophia Ananiadou. 2022. GRETEL: Graph contrastive topic enhanced language model for long document extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6259–6269, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3292–3303, Hong Kong, China. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST discourse parsing with second-stage EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020. Sg-net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9636–9643.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9709–9716.

A Appendix: RST Relation Category

RST type	RST label
Temporal	Temporal
Contingency	Cause, Condition
Comparison	Comparison, Contrast, Concession, Topic-Change
Expansion	Explanation, Elaboration, Background, Topic-Comment

Table A: RST relation category

B Appendix: Questionnaire of Human Evaluation

Here we provide a more detailed description of the criterion in our human evaluation.

- **Faithfulness**
  1. Completely hallucinated content
  2. A lot of hallucinated content and factual mistakes
  3. Most content is supported by the source document
  4. Only one or two characters or events contradicted or not mentioned in the source
  5. All information in the summary is faithful/supported by the source
- **Informativeness**
  1. No important information in the source is covered in the summary
  2. Only covers a small fraction of the source document information; one cannot learn the main content of the story from only the summary
  3. Covers around half of the important points from the source; one can learn the main content of the story from only the summary
  4. Only a few important points are missing in the summary
  5. All important information is summarized
- **Readability**
  1. Not understandable at all
  2. Hard to understand the content of the summary
  3. The summary is overall readable, with most sentences correct and fluent
  4. Easy to understand, with only occasional grammatical mistakes or incoherent sentences
  5. Fluent, with minor or no grammatical mistakes, coherent sentences, and clear structure
- **Conciseness**
  1. All information in the summary is redundant or unimportant
  2. Most of the information in the summary is redundant or unimportant
  3. Around half of the content in the summary is redundant
  4. Only a few points in the summary are redundant
  5. No information in the summary is redundant

User interface and instructions for rating and ranking can be found in Figure 7 and Figure 8.

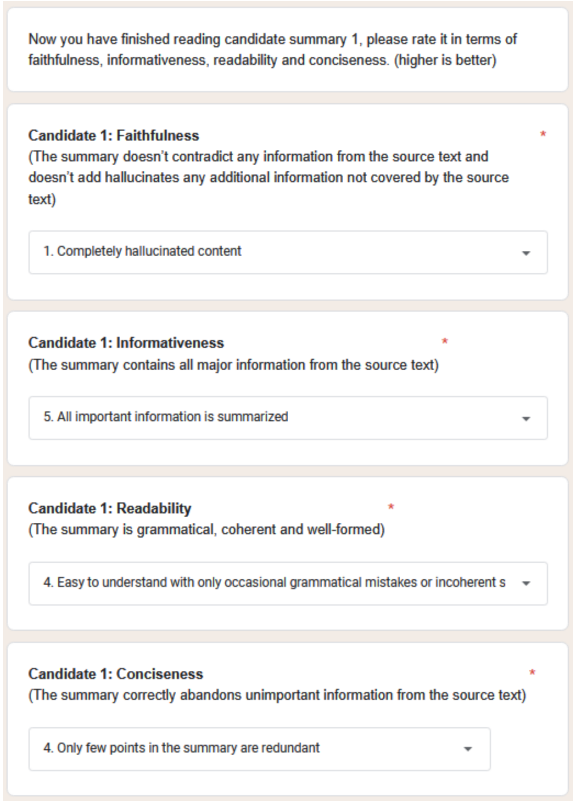

Figure 7: Instructions to rate candidate summaries in terms of each metric in human evaluation.

C Appendix: Sentence Alignment for Other Datasets

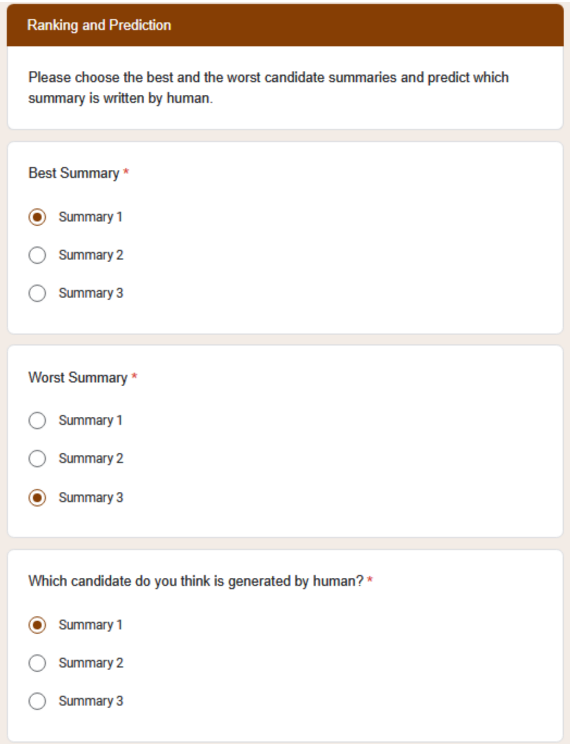

Figure 8: Instructions to rank all three candidates and predict which one is generated by human.

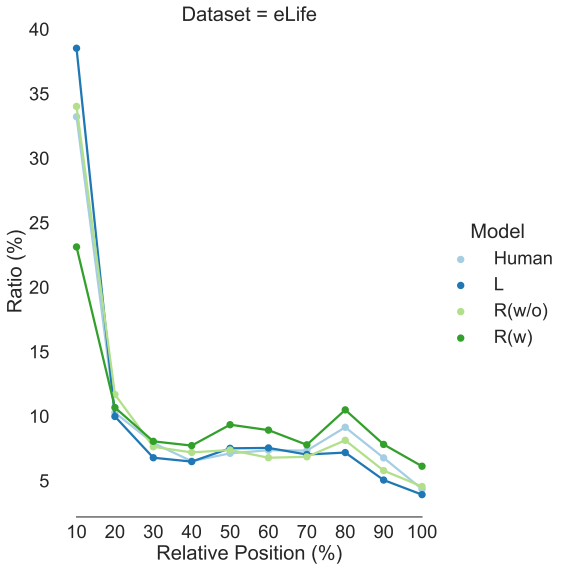

Figure 9: Sentence alignment distribution. L = Longformer, R(w/o) = RSTformer(w/o relations), R(w) = RSTformer(w relations).

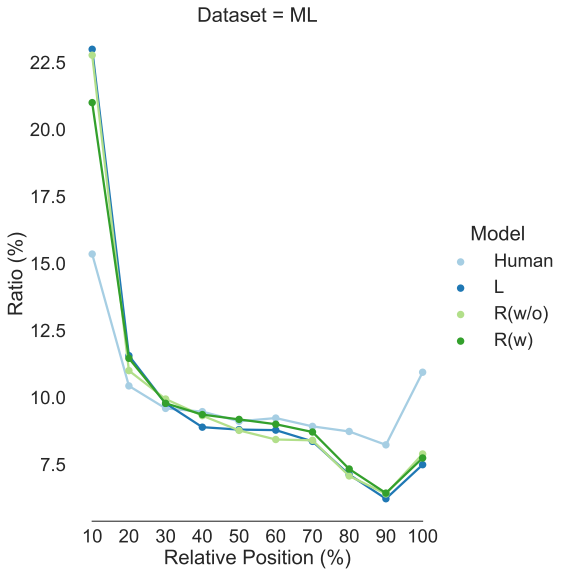

Figure 10: Sentence alignment distribution. L = Longformer, R(w/o) = RSTformer(w/o relations), R(w) = RSTformer(w relations), ML = Multi-LexSum.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 and Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 3 and Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 3 and Section 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3 and Section 4

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4.2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 4.2

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.2

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 4 and Appendix

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.