Structure-Preserving Document-Level Data Augmentation via Discourse Parsing

Anonymous ACL submission

Abstract

Data augmentation (DA) plays a vital role in improving model performance in low-resource or limited-supervision scenarios within Natural Language Processing (NLP). Existing DA methods, such as synonym replacement and back-translation, have demonstrated effectiveness on the word or sentence level; however, they frequently neglect discourse-level coherence and logical flow, which are essential for complex tasks dependent on inter-sentential relationships. In this paper, we propose a structure-preserving document-level data augmentation framework based on large language models (LLMs) and fine-grained discourse structure parsing. Our approach identifies rhetorical relations between sentence pairs and extracts key phrases, which are then replaced with topic-unrelated content while preserving the original discourse structure. Experimental results on text summarization and question answering show that training with data augmented by our method consistently outperforms the baseline, demonstrating the effectiveness of structure-preserving data augmentation for document-level NLP tasks.

1 Introduction

Data augmentation (DA) has emerged as a pivotal strategy in Natural Language Processing (NLP) to address data sparsity, particularly under low-resource or limited supervision scenarios. By synthetically expanding the training data, DA improves both model robustness and generalization. Existing approaches primarily operate at the lexical or sentence level, such as synonym replacement, word insertion and deletion, or back-translation, and have demonstrated effectiveness across a wide range of NLP tasks. (Van Dyk and Meng, 2001; Shorten and Khoshgoftaar, 2019; Shorten et al., 2021; Feng et al., 2021; Chen et al., 2023)

However, lexical- and sentence-level data augmentation methods are primarily effective for sentence-level NLP tasks and perform poorly on document-level tasks involving paragraph processing, such as text summarization and question answering. Recently, several studies have begun to explore data augmentation techniques tailored for document-level tasks. Huang et al. (2023) proposes a transfer learning framework with document-level data augmentation to enhance the transfer of accurate document-level knowledge for aspect-level sentiment classification tasks. Wu et al. (2024) introduces a target-side augmentation method that generates diverse candidate translations to facilitate the training of document-level machine translation models. Despite their success, these methods often overlook higher-level textual structures, particularly document-level coherence and intra-paragraph logical flow. As a result, the augmented paragraphs may contain semantically inconsistent noise, which can undermine the effectiveness of model training.

To address this issue, some researchers have explored incorporating discourse-level information into document-level data augmentation. For example, Feng et al. (2020) proposes a dialogue discourse-aware augmentation strategy to construct a pseudo-summarization corpus from existing meeting transcripts. Li et al. (2025) applies content reordering and segment-level paraphrasing to maintain discourse consistency while enhancing diversity. Binte et al. (2024) generates complex question–answer pairs guided by discourse relations. However, these methods typically rely on limited or coarse-grained discourse representations and lack a unified mechanism for capturing fine-grained rhetorical relations across diverse textual domains. This limitation restricts their applicability to specific NLP tasks and hinders their potential for broader adoption.

Preserving discourse structure in data augmentation offers two key benefits: (1) Relying solely on discourse structure as the augmentation

guide—without task-specific signals—enables the method to generalize across document-level NLP tasks. (2) Maintaining the input's discourse structure in augmented data simplifies obtaining corresponding target labels, thereby minimizing the risk of introducing noise. In this paper, we propose a structure-preserving data augmentation framework for document-level NLP tasks, based on fine-grained discourse structure parsing. Our method first parses discourse to identify rhetorical relations and extract key phrases from sentence pairs. These phrases are then replaced with new content while preserving the original rhetorical structure. Building on this, the framework generates coherent paragraphs that maintain discourse coherence and introduce semantic diversity, enabling effective document-level augmentation. We evaluated our method on two representative downstream tasks: text summarization and question answering. Experimental results demonstrate that our approach consistently improves performance across various settings, highlighting the effectiveness of structure-preserving data augmentation.

Our contributions are summarized as follows:

(1) A fine-grained discourse parsing framework that structurally models rhetorical relations and sentence-level key semantics within documents.

(2) A novel discourse-level data augmentation method that manipulates rhetorical structures and core content to generate coherent and diverse textual instances.

(3) Extensive experiments on summarization and question answering tasks demonstrating significant performance improvements over existing data augmentation techniques.

2 Related Works

Data augmentation (Feng et al., 2021) generates new data by transforming existing data points using operations designed based on prior knowledge of the problem's structure (Wang and Yang, 2015; Wei and Zou, 2019). This augmented data can be derived from labeled examples and directly applied in supervised learning (Wei and Zou, 2019), or utilized in semi-supervised learning on unlabeled data via consistency regularization (Xie et al., 2020). Based on their targets and approaches, current data augmentation methods can be categorized into token-level augmentation (Niu and Bansal, 2018; Kumar et al., 2020; Miao et al., 2020), sentence-level augmentation (Chen et al., 2020; Yang et al., 2020), adversarial data augmentation (Cheng et al., 2019; Morris et al., 2020), and hidden-space augmentation (Malandrakis et al., 2019; Chen et al., 2021). Recently, Large Language Models (LLMs) have been increasingly employed for data augmentation due to their strong generative capabilities and task adaptability, achieving remarkable results (Ye et al., 2022; Yu et al., 2023; Ubani et al., 2023; Chung et al., 2023; Dai et al., 2025). However, token- and sentence-level data augmentation methods are primarily effective for sentence-level NLP tasks and perform poorly on document-level tasks involving paragraph processing, such as summarization and question answering.

Recently, several studies have begun exploring data augmentation techniques tailored for document-level tasks (Huang et al., 2023; Wu et al., 2024; Bao et al., 2023). However, existing document-level data augmentation methods primarily operate at the lexical or syntactic level and often overlook higher-level structures such as inter-sentential coherence and document-level logic. As a result, the augmented data may introduce semantically inconsistent noise in document-level NLP tasks, undermining the effectiveness of model training. To address this limitation, recent efforts have focused on document-level data augmentation that incorporates discourse-level structure (Feng et al., 2020; Pasunuru et al., 2021; Binte et al., 2024; Li et al., 2025). However, despite these advances, most existing discourse-aware data augmentation methods rely on coarse-grained or task-specific discourse representations and lack a unified framework for modeling fine-grained rhetorical relations that generalize across diverse textual domains. Therefore, we propose a structure-preserving data augmentation framework for document-level NLP tasks, grounded in fine-grained discourse structure parsing.

We also summarize relevant work on discourse parsing. Foundational frameworks such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the Penn Discourse Treebank (Prasad et al., 2008) provide structured representations of discourse through rhetorical trees or argument spans. Recently, neural discourse parsers have been developed (Morey et al., 2017; Shi et al., 2020; Hu and Wan, 2023). In this paper, we aim to improve both the efficiency and effectiveness of document-level discourse parsing. To this end, we refine

the set of rhetorical relation types, streamline the discourse parsing process, and propose an LLM-based fine-grained discourse parsing framework that structurally models rhetorical relations and sentence-level key semantics within documents.

3 Methodology

As illustrated in Algorithm 1 and Figure 1, we propose a structure-preserving document-level data augmentation framework consisting of three key stages. First, rhetorical relations between sentence pairs are identified through discourse parsing, and salient words or phrases are extracted from each sentence. Next, the extracted key phrases are replaced with topic-irrelevant content to introduce semantic variability. Finally, a new document is generated by preserving the original rhetorical structure while incorporating the modified content. This approach increases data diversity without compromising discourse coherence, thereby providing reliable support for downstream tasks that require both semantic understanding and structural reasoning.

3.1 Discourse Parsing and Key Phrase Extraction

Given an input document $\mathbf{D}$, we first decompose it into an ordered sequence of sentences $\mathbf{D} = \{s_1, s_2, \ldots, s_n\}$, where s_i denotes the i-th sentence in $\mathbf{D}$ and n is the total number of sentences. Lines 1–9 of Algorithm 1 perform discourse parsing and key phrase extraction within the input document $\mathbf{D}$.

Note that, to simplify the discourse parsing process, we do not adopt Rhetorical Structure Theory (RST) (Mann and Thompson, 1988); instead, we employ a simplified dependency-style approach to discourse parsing. Specifically, the rhetorical relation $R_{i,j}$ between sentences s_i and s_j must satisfy the following three constraints: (a) $R_{i,j}$ indicates that there exists a rhetorical relation from s_j to s_i ($i < j$), but not vice versa. (b) Except for s_1, each sentence s_j (where $j = 2, \ldots, n$) must have exactly one rhetorical relation $R_{i,j}$ with a preceding sentence s_i (where $i < j$). (c) $R_{i,j}$ is selected from one of the six types of rhetorical relation defined in Table 1, which capture the most common and semantically salient inter-sentential relationships observed in natural text.These relations are designed to balance expressiveness, interpretability, and applicability to both generation and analysis tasks.

Algorithm 1: Structure-Preserving Document-Level Data Augmentation

Input: Document $\mathbf{D} = \{s_1, \ldots, s_n\}$
Output: Augmented document $\mathbf{D}' = \{s'_1, \ldots, s'_n\}$

```
/* Discourse parsing and key phrase extraction */
```
1 Initialize OriList $\leftarrow [\,]$;
2 **for** j *in* $(2, \cdots, n)$ **do**
3 $\quad s_i \leftarrow \texttt{FindValid}(s_1, \cdots, s_{j-1})$;
4 $\quad R_{i,j} \leftarrow \alpha(s_i, s_j)$;
5 $\quad \mathbf{K}_i \leftarrow \beta(s_i)$; // $\mathbf{K}_i = (k_i^1, \cdots, k_i^p)$
6 $\quad \mathbf{K}_j \leftarrow \beta(s_j)$; // $\mathbf{K}_j = (k_j^1, \cdots, k_j^q)$
7 $\quad \mathbf{A}_{i,j} \leftarrow (s_i, \mathbf{K}_i, s_j, \mathbf{K}_j, R_{i,j})$;
8 $\quad$ OriList.append($\mathbf{A}_{i,j}$) ;
9 **end**

```
/* Key phrase replacement */
```
10 Initialize AugList $\leftarrow [\,]$;
11 **for** $\mathbf{A}_{i,j}$ *in* OriList **do**
12 $\quad \mathbf{K}'_i \leftarrow \gamma(\mathbf{K}_i, (\mathbf{K}'_1, \cdots, \mathbf{K}'_{i-1}))$ // $\mathbf{K}'_i = (k_i'^1, \cdots, k_i'^p)$
13 $\quad \mathbf{K}'_j \leftarrow \gamma(\mathbf{K}_j, (\mathbf{K}'_1, \cdots, \mathbf{K}'_{j-1}))$ // $\mathbf{K}'_j = (k_j'^1, \cdots, k_j'^q)$
14 $\quad \mathbf{A}'_{i,j} \leftarrow (s'_i, \mathbf{K}'_i, s'_j, \mathbf{K}'_j, R_{i,j})$;
15 $\quad$ AugList.append($\mathbf{A}'_{i,j}$) ;
16 **end**

```
/* Paragraph generation */
```
17 **for** i *in* $(1, \cdots, n)$ **do**
18 $\quad s'_i = \delta(\mathbf{A}'_{i,*}, \mathbf{A}'_{*,i})$;
19 $\quad \mathbf{D}'$.append(s'_i) ;
20 **end**

Therefore, for a given sentence s_j, the function $\mathrm{FindValid}(s_i, \cdots, s_{j-1})$ in Line 3 of Algorithm 1 identifies a sentence s_i that holds a rhetorical relation with s_j according to the following procedure:

(1) The algorithm first checks whether a valid rhetorical relation exists between s_j and s_{j-1}. If so, s_i is set to s_{j-1}.

(2) If no rhetorical relation is found between s_{j-1} and s_j, the algorithm sequentially examines each sentence from s_1 to s_{j-2} to determine whether a valid rhetorical relation exists with s_j. If such a sentence is found, s_i is set to the first one that satisfies the condition.

(3) If no valid rhetorical relation is detected between s_j and any preceding sentence, s_i is set to s_{j-1}, and $R_{i,j}$ is assigned a default "Con-

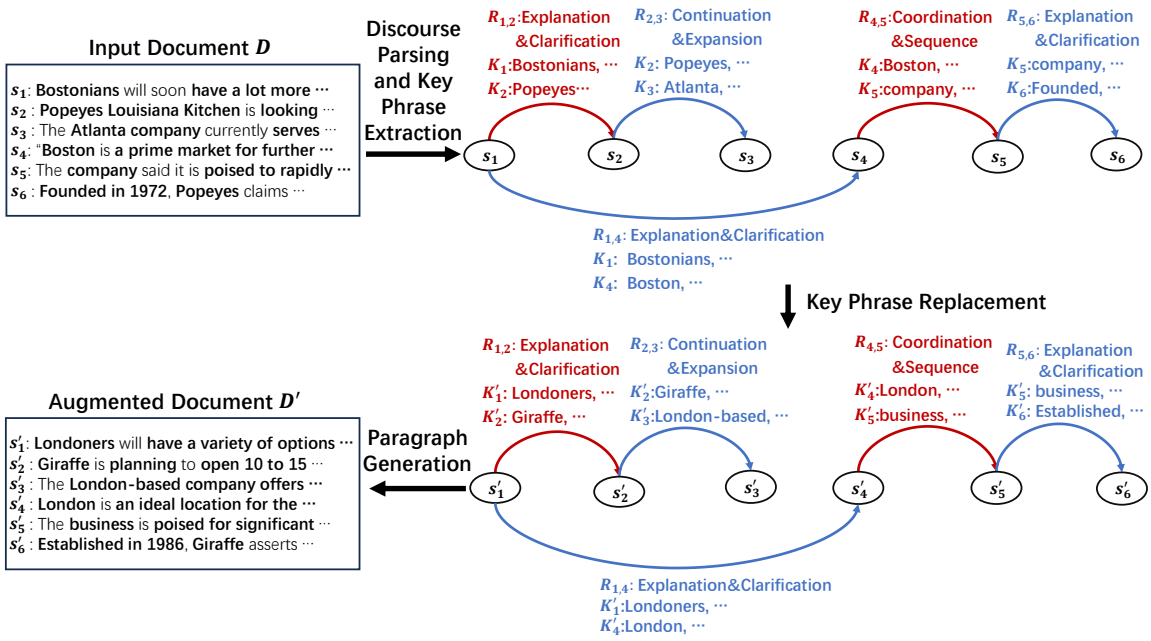

Figure 1: Structure-preserving document-level data augmentation

tinuation & Expansion" relation.

Since most sentences s_j typically have rhetorical relations with either s_{j-1} or s_1, the above process reduces the discourse parsing time complexity from $O(n^2)$ to approximately $O(n)$.

As shown in Lines 4–8 of Algorithm 1, $\alpha(s_i, s_j)$ denotes the function that identifies the rhetorical relation between s_i and s_j, while $\beta(s_i)$ denotes the function that extracts key phrases from sentence s_i. If key phrases have already been extracted from s_i, $\beta(s_i)$ directly returns the previously obtained result to ensure consistency in the extraction process. The extracted information is stored as $\mathbf{A}_{i,j} = (s_i, \mathbf{K}_i, s_j, \mathbf{K}_j, R_{i,j})$, where $\mathbf{K}_i = (k_i^1, \ldots, k_i^p)$ and $\mathbf{K}_j = (k_j^1, \ldots, k_j^q)$ represent the key phrases of s_i and s_j, respectively. Each $\mathbf{A}_{i,j}$ is added to OriList, which is then used to replace the key phrases $\mathbf{K}_i$ and $\mathbf{K}_j$.

3.2 Key Phrase Replacement

Lines 10–16 of Algorithm 1 focus on replacing the extracted key phrases in each entry of OriList while preserving the corresponding rhetorical relations within the document $\mathbf{D}$.

For each parsed input data $\mathbf{A}_{i,j}$, where the rhetorical relation $R_{i,j}$ and key phrases $\mathbf{K}_i$ and $\mathbf{K}_j$ have been extracted as described in Section 3.1, we proceed to construct the corresponding augmented key phrases $\mathbf{K}'_i$ and $\mathbf{K}'_j$. For $s_i \in \mathbf{A}_{i,j}$, the function $\gamma(\mathbf{K}_i, (\mathbf{K}'_1, \cdots, \mathbf{K}'_{i-1}))$ generates replacement key phrases for $\mathbf{K}_i$ by referencing both the original key phrases $\mathbf{K}_i$ and the previously replaced key phrases $(\mathbf{K}'_1, \cdots, \mathbf{K}'_{i-1})$. This process ensures thematic consistency and discourse-level coherence throughout the augmented document.

Specifically, the function γ replaces each key phrase with a semantically consistent but topically distinct alternative, guided by a shared augmentation topic inferred from previous replacements. The resulting structure-preserved augmented data is formalized as: $\mathbf{A}'_{i,j} = (s'_i, \mathbf{K}'_i, s'_j, \mathbf{K}'_j, R_{i,j})$, where s'_i and s'_j are placeholders for sentences in the augmented data to be generated in Section 3.3. Each $\mathbf{A}'_{i,j}$ is appended to AugList, which is subsequently used to generate the final augmented document $\mathbf{D}'$.

3.3 Paragraph Generation

Lines 17–20 of Algorithm 1 describe the final step of generating the augmented document $\mathbf{D}'$ using the modified key phrases and the original rhetorical relations stored in AugList.

For each sentence placeholder s'_i, the function $\delta(A'_{i,*}, A'_{*,i})$ generates a new sentence based on the structure-preserved related data $A'_{i,*}$ and $A'_{*,i}$. Each generated s'_i must consist of exactly one sentence, ensuring that the total number of sentences in the augmented document remains identical to that of the original input document. The final augmented document $\mathbf{D}'$ is then constructed by assembling the generated sentences $\{s'_1, \ldots, s'_n\}$.

Rhetorical Relation	Explanation
Causality & Condition	Expresses a cause, consequence, or conditional premise between two sentences.
Contrast & Transition	Highlights semantic contrast, opposition, or a transitional shift in topic or stance.
Coordination & Sequence	Presents parallel content or arranges ideas in a temporal or logical sequence.
Explanation & Clarification	Provides elaboration, clarification, examples, or further explanation of the preceding content.
Summary & Generalization	Summarizes, generalizes, or abstracts previously stated content.
Continuation & Expansion	Naturally continues or extends a prior idea, often without explicit connectives.

Table 1: Six rhetorical relation types and their corresponding descriptions.

It is worth noting that, as described in Section 3.2, we only perform key phrase replacement within individual sentences. Moreover, the rhetorical relations $R_{i,j}$ remain unchanged during the generation process described in this section. As a result, the augmented document $\mathbf{D}'$ preserves not only intra-sentence semantic similarity with the original input document $\mathbf{D}$, but also maintains global structural consistency.

Furthermore, since both sentence-level and paragraph-level discourse structures are preserved, the corresponding target-side annotations for downstream tasks can be derived through simple rule-based mappings. For example, in summarization, the summary sentence in the augmented document can be aligned to the one at the same index as in the original document. In question answering, the answer span can be identified by selecting the key phrase in the augmented document that corresponds to the same relative position as in the original input.

4 Experiment Settings

4.1 Datasets and Metrics

To evaluate the effectiveness of the proposed structure-preserving document-level data augmentation framework, we conduct experiments on two document-level NLP tasks: text summarization and question answering.

For the text summarization task, we use two datasets—the Chinese Abstractive Corpus[1] and the Newsroom Dataset (Grusky et al., 2018)—both of which provide summaries that are extracted verbatim from the original paragraphs. To improve experimental efficiency, we did not utilize the full datasets. Instead, 2,000 instances were randomly sampled from each dataset as input for generating augmented data using the method described in Section 3. An additional 2,000 instances were sampled as a control group, and 1,000 instances were reserved as the test set.

For the question answering task, we use the SQuAD1.1 (Rajpurkar et al., 2016) dataset. To improve experimental efficiency, 4,000 instances were randomly sampled as input for generating augmented data using the method described in Section 3. An additional 4,000 instances were sampled as a control group, and 1,000 instances were reserved as the test set.[2]

Different evaluation metrics are used for different datasets. For the Chinese Abstractive Corpus and the Newsroom dataset, we adopt BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) as evaluation metrics. For SQuAD 1.1, we use F1 score and Exact Match (EM) (Rajpurkar et al., 2016).

4.2 Baselines

We use a locally deployed, small-scale instruction-tuned LLM as the baseline in our experiments. In addition, we construct several training datasets to fine-tune a locally deployed base LLM and compare the performance of each fine-tuned variant against the baseline across different datasets. The training dataset configurations are defined as follows:

- **Inp**: Fine-tuning the base LLM using only the sampled input data.
- **Con**: Fine-tuning the base LLM using only the sampled control data.

[1] https://github.com/wonderfulsuccess/chinese_abstractive_corpus

[2] Possibly due to the relative simplicity of the question answering task compared to text summarization, fine-tuning the base LLM model with a small dataset leads to worse performance than directly using the instruction-following LLM model. Therefore, we sampled 4,000 instances for this task.

discourse parsing and key phrase extraction	Given a sentence pair (s_i, s_j), identify rhetorical relation into one of the predefined types in Table 1. Extract all key words or phrases from each sentence without modification. The output follows the format: $arc(s_i, s_i\_keyword, relation, s_j\_keyword, s_j)$.
key phrase replacement	Given previously replaced content and $arc(s_i, s_i\_keyword, relation, s_j\_keyword, s_j)$, strict one-to-one mapping and no completely lexical overlap between original and substituted keywords. The replaced keywords topic must be consistent with the given previously replaced content topic. The output strictly follows the format:$arc(s_i, replaced\_s_i\_keyword, relation, replaced\_s_j\_keyword, s_j)$.
paragraph generation	Given replaced $arc(s_i, replaced\_s_i\_keyword, relation, replaced\_s_j\_keyword, s_j)$, use of all provided keywords in each generated sentence, with no omission or paraphrase and only one sentence can be generated. Preservation of the specified rhetorical relation between the sentence pair (s_i, s_j). Contextual coherence with prior sentences to ensure fluent document-level discourse.

Table 2: Prompt texts used during the data augmentation process

- **Aug**: Fine-tuning the base LLM using only the augmented data.
- **Inp+Con**: Fine-tuning the base LLM using both the input and control data.
- **Inp+Aug**: Fine-tuning the base LLM using both the input and augmented data.
- **Con+Aug**: Fine-tuning the base LLM using both the control and augmented data.
- **Inp+Con+Aug**: Fine-tuning the base LLM using all three datasets: input, control, and augmented data.

4.3 Implementation Details

In the data augmentation process, we employ Qwen2.5-7B-Instruct [3], a small-scale instruction-tuned language model developed by Alibaba DAMO Academy, to perform the three tasks described in Section 3: discourse parsing, key phrase replacement, and paragraph generation. As described in Section 3, we decompose the structure-preserving data augmentation process into three separate stages, which yields significantly better performance than executing it in an end-to-end manner. As shown in Table 2, each step is executed independently using a single prompt with the LLM.

For efficiency, Qwen2.5-3B-Instruct[4] is deployed locally as the baseline model. To evaluate the effectiveness of each training dataset during fine-tuning, we use Qwen2.5-3B-Base[5] as the base LLM. Qwen2.5-3B-Base is fine-tuned using LoRA (Hu et al., 2022). The training batch size is set to 1 per device, with a gradient accumulation step of 32. A learning rate of 1×10^{-5} is used, and the model is trained for 3 epochs with a cosine learning rate scheduler and a warmup ratio of 0.1. All experiments are conducted using bfloat16 precision.

[3] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[4] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

[5] https://huggingface.co/Qwen/Qwen2.5-3B

5 Result and Analysis

5.1 Main Result

To evaluate the effectiveness of the proposed structure-preserving data augmentation framework, extensive experiments are performed on document-level text summarization and question answering tasks. Table 3 reports the BLEU and ROUGE scores for various fine-tuning configurations on both Chinese and English summarization datasets, based on human-annotated references. Table 4 presents the Exact Match (EM) and F1 scores on the SQuAD 1.1 dataset, using the same reference standard.

On the Chinese Abstractive Corpus, a clear performance gain is observed when augmented data is incorporated. The baseline Qwen2.5-3B-Instruct model without fine-tuning yields relatively low scores, indicating its limited ability to capture summarization-relevant structures. Fine-tuning with augmented data alone significantly improves performance, demonstrating the effectiveness of the proposed structure-preserving data augmentation framework for text summarization. Interestingly, combining multiple datasets generally yields the best performance. In particular, the LLM fine-tuned on both input and control data achieves the highest overall ROUGE scores. The model fine-tuned on all three datasets also performs competitively, suggesting that mixed supervision from diverse sources can enhance generalization ability.

As shown in Table 3, the performance on the Newsroom Dataset exhibits a trend similar to that

Method	Chinese Abstractive Corpus				Newsroom Datase			
	BLEU	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Baseline	0.139	0.290	0.193	0.265	0.267	0.401	0.290	0.364
Inp	0.249	0.391	0.312	0.372	0.634	0.697	0.657	0.686
Con	0.228	0.370	0.289	0.351	0.653	0.712	0.676	0.703
Aug	0.269	0.375	0.309	0.360	0.639	0.696	0.658	0.686
Inp+Con	**0.276**	**0.417**	**0.343**	**0.400**	0.686	0.741	0.709	0.733
Inp+Aug	0.247	0.385	0.304	0.366	0.775	0.796	0.765	0.787
Con+Aug	0.233	0.372	0.289	0.351	**0.787**	**0.806**	**0.778**	**0.798**
Inp+Con +Aug	0.275	0.412	0.338	0.396	0.780	0.789	0.757	0.780

Table 3: Performance of different fine-tuning datasets on the text summarization task

Method	Exact Match	F1
Baseline	0.636	**0.791**
Inp	0.708	0.780
Con	0.690	0.742
Aug	0.698	0.740
Inp+Con	0.697	0.746
Inp+Aug	0.701	0.751
Con+Aug	0.720	0.740
Inp+Con+Aug	**0.731**	0.779

Table 4: Performance of different fine-tuning datasets on the question answering task

observed on the Chinese dataset. Although the baseline starts at a relatively high level, substantial improvements are still achieved when fine-tuning the base LLM with the augmented data. The best results are achieved when the LLM is fine-tuned on a combination of control and augmented data. The second-best performance is obtained by the model fine-tuned on all three types of training data. These results suggest that the augmented data generated by the proposed method effectively improves model performance in the text summarization task.

Table 4 presents the performance of various fine-tuning configurations on the SQuAD1.1 dataset, evaluated using Exact Match (EM) and F1 scores against human-annotated gold answers. The baseline model without fine-tuning achieves an EM score of 0.636. Fine-tuning the LLM solely on augmented data increases the EM score to 0.698, demonstrating the effectiveness of the proposed data augmentation method. Fine-tuning the model with control and augmented data yields additional performance gains, while using all three datasets achieves the highest EM score of 0.731. These results suggest that the augmented data generated by the proposed method effectively improves model accuracy in the question answering task.

Interestingly, the baseline model achieves the highest F1 score (0.791), outperforming all fine-tuned models, which contrasts with the trend observed in the Exact Match (EM) metric. This discrepancy arises because fine-tuning significantly improves EM scores, reflecting enhanced precision in generating reference answers, but may simultaneously reduce the model's flexibility in handling paraphrases or semantically equivalent expressions, leading to a lower overall F1 score. As the amount of training data increases, the F1 score steadily improves. Among the fine-tuned models, the one trained on the combination of all three data types attains the highest F1 score, second only to the baseline. This finding suggests that the augmented data generated by our proposed method effectively enhances model performance.

5.2 Case Study

To illustrate the practical operation of our structure-preserving document-level data augmentation framework, Table 5 presents a step-by-step case study.

Discourse parsing and key phrase extraction: In the discourse parsing and key phrase extraction step, rhetorical relations are identified between sentence pairs within the original input document, and key phrases are extracted from each sentence. For example, sentence s_1 contains key phrases such as "Petersen Publishing Company"and "in Los Angeles,"and holds an Explanation & Clarification rhetorical relation with sentence s_2, since s_2 provides additional details about the "Circle Track" magazine published by Petersen Publishing Company.

original input document	s_1: The **Petersen Publishing Company in Los Angeles**, which already has **13 monthly magazines** catering to **leisure-time activities**, has **two more on the drawing board**. s_2: Like **six of its current titles**, **one will deal with motor vehicles - Circle Track**, whose **first issue** will have an **October** cover and a **circulation of 125,000**. s_3: **Page rates** have been set at **$1,200 for one full-page black-and-white** and **$1,920 for four-color**. s_4: **VideoWorld**, whose **first issue** will be marked **December**, will be **consumer-oriented** and **will have a circulation of 200,000**. s_5: **Page rates** will be **$2,625 and $4,200, respectively**.
discourse parsing and key phrase extraction	$A_{1,2}$: (s_1, (Petersen Publishing Company, in Los Angeles, 13 monthly magazines, $\cdots$), s_2, (six of its current titles, one will deal with motor vehicles, Circle Track, first issue, $\cdots$), **Explanation & Clarification**), $A_{2,3}$: (s_2, (six of its current titles, one will deal with motor vehicles, Circle Track, first issue, $\cdots$), s_3, (Page rates, $1,200 for one full-page black-and-white, $1,920 for four-color), **Coordination & Sequence**), $A_{1,4}$: (s_1, (Petersen Publishing Company, in Los Angeles, 13 monthly magazines, $\cdots$), s_4, (VideoWorld, first issue , December, consumer-oriented, $\cdots$), **Explanation & Clarification**), $A_{4,5}$: (s_4, (VideoWorld, first issue , December, consumer-oriented, $\cdots$), s_5, (Page rates, $2,625 and $4,200, respectively), **Coordination & Sequence**)
key phrase replacement	$A'_{1,2}$: (s'_1, (Smithsonian Institution, in Washington D.C., 13 annual reports, historical preservation, $\cdots$), s'_2, (six ongoing projects, one focusing on ancient ruins, Terra Cotta Warrior, first excavation, $\cdots$), **Explanation & Clarification**), $A'_{2,3}$: (s'_2, (six ongoing projects, one focusing on ancient ruins, Terra Cotta Warrior, first excavation, $\cdots$), s'_3, (Publication rates , $1,800 for one full-page black-and-white , $2,700 for four-color), **Coordination & Sequence**), $A'_{1,4}$: (s'_1, (Smithsonian Institution, in Washington D.C., 13 annual reports, historical preservation, $\cdots$), s'_4, (Heritage Preservation Quarterly, inaugural edition, published in January, $\cdots$), **Explanation & Clarification**), $A'_{4,5}$: (s'_4, (Heritage Preservation Quarterly, inaugural edition, $\cdots$), s'_5, (Publication rates, $2,100 and $3,150, respectively), **Coordination & Sequence**)
paragraph generation	s'_1: The **Smithsonian Institution in Washington D.C.**, has released **13 annual reports** focused on **historical preservation**, and there are **three more in development**. s'_2: Like **six ongoing projects**, **one focusing on ancient ruins**, including the **Terra Cotta Warrior**, which began its **first excavation** in **November**, **team size 50** is currently working on multiple fronts to uncover more historical secrets. s'_3: **Publication rates** have been established at **$1,800 for one full-page black-and-white** and **$2,700 for four-color**. s'_4: **Heritage Preservation Quarterly**, whose **inaugural edition** will be **published in January**, **will focus on educational initiatives** and **will reach an audience of 150,000**. s'_5: **Publication rates** will be **$2,100 and $3,150, respectively**.

Table 5: A Case study of structure-preserving document-level data augmentation

Key Phrase Replacement: In the key phrase replacement step, extracted key phrases are substituted with contextually coherent alternatives. For example, "The Petersen Publishing Company" in sentence s_1 is replaced with "Smithsonian Institution," and "Circle Track" in sentence s_2 is replaced with "Terra Cotta Warrior." Importantly, these replacements are performed in a manner that preserves the original rhetorical relations identified in the previous step.

Paragraph Generation: In the paragraph generation step, the augmented document is generated based on the previously parsed rhetorical relations and the replaced key phrases. For example, s'_1 describes "the Smithsonian Institution", which is dedicated to historical preservation, while s'_2 provides further details about the Institution's "Terra Cotta Warrior" project. Notably, the rhetorical relation between the generated sentences s'_1 and s'_2 remains consistent with that between the original sentences s_1 and s_2. The generated paragraph preserves the rhetorical structure, key phrase sequencing, and logical relations of the original document. This demonstrates that our method can produce augmented data that are both structurally faithful to the source document and lexically diverse.

6 Conclusion

In this paper, we propose a structure-preserving document-level data augmentation framework, grounded in fine-grained discourse structure parsing. Our approach identifies rhetorical relations between sentence pairs and extracts key phrases, which are then replaced with topic-unrelated content while preserving the original discourse structure. The augmented data generated by our method consistently improves performance on text summarization and question answering tasks, demonstrating the effectiveness of structure-preserving data augmentation. Moreover, our proposed method is task-agnostic and theoretically applicable to a wide range of document-level NLP tasks.

Limitations

The primary limitation of this work is that the data augmentation process is implemented through three relatively independent stages, rather than via an end-to-end machine learning approach. However, even end-to-end data augmentation methods must inherently involve the same three stages outlined in this paper: discourse parsing and key phrase extraction, key phrase replacement, and paragraph generation. Since these stages have distinct problem formulations and objectives, integrating them into a unified end-to-end framework remains a significant challenge. Addressing this challenge will be an important direction for our future work.

In addition, the core processing in our proposed method is implemented using pretrained large language models (LLMs). This design choice is motivated by the fact that our primary contribution lies in establishing a framework for analyzing rhetorical relations and sentence-level key semantics within documents, and leveraging the resulting discourse structure for data augmentation. Therefore, replacing the LLM component with more advanced discourse modeling techniques could, in principle, yield better structural representations. We leave this exploration for future work.

Although our proposed method is task-agnostic and theoretically applicable to a wide range of document-level NLP tasks, in this work we focus our empirical evaluation on text summarization and question answering. As future work, we plan to extend our evaluation to more complex document-level tasks, such as document-level machine translation and aspect-level sentiment analysis.

References

Guangsheng Bao, Zhiyang Teng, and Yue Zhang. 2023. Target-side augmentation for document-level machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10725–10742.

Kushnur Binte, Philippe Muller, and Chloé Braud. 2024. Complex question generation using discourse-based data augmentation. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*.

Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. 2021. Hiddencut: Simple data augmentation for natural language understanding with better generalization. *arXiv preprint arXiv:2106.00149*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Jiaao Chen, Yuwei Wu, and Diyi Yang. 2020. Semi-supervised models via data augmentationfor classifying interactive affective responses. *arXiv preprint arXiv:2004.10972*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Fang Zeng, Wei Liu, and 1 others. 2025. Auggpt: Leveraging chatgpt for text data augmentation. *IEEE Transactions on Big Data*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020. Dialogue discourse-aware graph model and data augmentation for meeting summarization. *arXiv preprint arXiv:2012.03502*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Xinyu Hu and Xiaojun Wan. 2023. Rst discourse parsing as text-to-text generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3278–3289.

Xiaosai Huang, Jing Li, Jia Wu, Jun Chang, and Donghua Liu. 2023. Transfer learning with document-level data augmentation for aspect-level sentiment classification. *IEEE Transactions on Big Data*, 9(6):1643–1657.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Weihao Li, Dan Jiang, Han Zhang, Kejing Xiao, and Shaozhong Cao. 2025. An adaptive fusion-based data augmentation method for abstract dialogue summarization. *PeerJ Computer Science*, 11:e2845.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou. 2019. Controlled text generation for data augmentation in intelligent artificial agents. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 90–98.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of the web conference 2020*, pages 617–628.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on rst discourse parsing? a replication study of recent results on the rst-dt. In *Conference on Empirical Methods on Natural Language Processing (EMNLP 2017)*, pages pp–1330.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 486–496.

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.

Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. 2021. Data augmentation for abstractive query-focused multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13666–13674.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, Bonnie L Webber, and 1 others. 2008. The penn discourse treebank 2.0. In *LREC*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ke Shi, Zhengyuan Liu, and Nancy F Chen. 2020. An end-to-end document-level neural discourse parser exploiting multi-granularity representations. *arXiv preprint arXiv:2012.11169*.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

David A Van Dyk and Xiao-Li Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for*

Computational Linguistics: EMNLP 2020, pages 1008–1025.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in neural information processing systems*, 36:55734–55784.