One-Dimensional Object Detection for Streaming Text Segmentation of Meeting Dialogue

Rui He [1,2], Zhongqing Wang [1], Minjie Qiang [1], Hongling Wang [1]*
Yifan Zhang [2], Hua Xu [2], Shuai Fan [2], Guodong Zhou [1]
[1] Natural Language Processing Lab, Soochow University, Suzhou, China
[2] AISpeech Ltd., Suzhou, China
rhesudanlp@stu.suda.edu.cn

Abstract

Dialogue text segmentation aims to partition dialogue content into consecutive paragraphs based on themes or logic, enhancing its comprehensibility and manageability. Current text segmentation models, when applied directly to STS (Streaming Text Segmentation), exhibit numerous limitations, such as imbalances in labels that affect the stability of model training, and discrepancies between the model's training tasks (sentence classification) and the actual text segmentation that limit the model's segmentation capabilities.

To address these challenges, we first implement STS for the first time using a sliding window-based segmentation method. Secondly, we employ two different levels of sliding window-based balanced label strategies to stabilize the training process of the streaming segmentation model and enhance training convergence speed. Finally, by adding a one-dimensional bounding-box regression task for text sequences within the window, we restructure the training approach of STS tasks, shifting from sentence classification to sequence segmentation, thereby aligning the training objectives with the task objectives, which further enhanced the model's performance. [1]Extensive experimental results demonstrate that our method is robust, controllable, and achieves state-of-the-art performance.

1 Introduction

With the advancement and widespread adoption of Automatic Speech Recognition (ASR) technology, along with the increasing societal demand for real-time recording and processing of dialogue content, Streaming Text Segmentation (STS) has become a key task in real-time dialogue preprocessing and has gained increasing attention. In scenarios such as business meetings, academic seminars, consulting services, training, and speeches, lengthy dialogue content frequently occurs, and handling this massive streaming information poses a challenge. Effectively segmenting dialogue content into coherent thematic paragraphs (Brants et al., 2002) greatly facilitates subsequent understanding and analysis, such as in dialogue summarization (Qi et al., 2021; Schneider and Turchi, 2023; Li et al., 2019), question answering (Yoon et al., 2018), and information retrieval (Zhu et al., 2019).

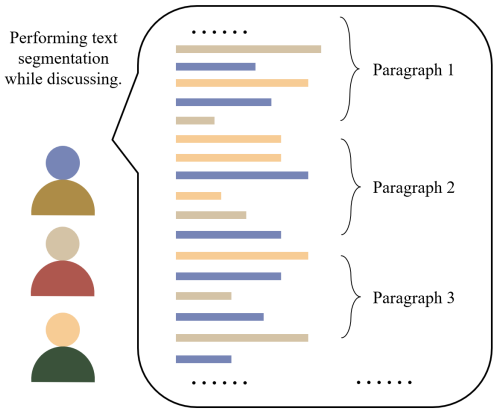

Figure 1: The dialogue content is segmented in real-time into paragraphs, with each paragraph representing a semantically complete conversation that implicitly contains a specific topic.

Although text segmentation technology has made significant progress, research on STS is still relatively scarce. While some existing segmentation techniques can be applied to STS, how to further improve the accuracy of streaming segmentation, optimize processing workflows, and reduce latency remains an urgent issue to explore.

Segmenting based on sentence granularity using text transcribed by ASR is often impractical because there may not be sufficient context to determine whether the current utterance is a segmentation boundary. To address this issue, we adopted

*Corresponding author
[1]https://github.com/DDDeeeee/1DOD

Findings of the Association for Computational Linguistics: ACL 2025, pages 4118–4130

self-adaptive sliding window (Zhang et al., 2021), performing segmentation at the window level on streaming text. Supervised text segmentation tasks are commonly interpreted as sentence classification tasks, Among these, the issue of imbalance is particularly pronounced, especially within long texts, resulting in sparse segmentation labels that impair model performance and make the training process slow and unstable. Furthermore, interpreting text segmentation tasks merely as sentence classification tasks is insufficient because this perspective overlooks paragraph content, paragraph length (Yoo and Kim, 2024), and the intrinsic structure of the text, considering only boundary locations; more importantly, there is a significant disparity between text segmentation and sentence classification tasks.

In order to solve the existing problems and improve the accuracy of STS, our research and contributions can be summarized as follows:

- To our knowledge, our method is the first to address the STS problem and resolve the issue of imbalanced training labels in text segmentation. First, we achieve STS by sampling and segmenting dialogue texts using a sliding window. Second, we introduce two levels of balanced labels to replace conventional text segmentation labels, enabling the model to converge faster and more stably during training, while effectively controlling paragraph lengths during inference.

- Inspired by object detection and semantic segmentation in images, we implemented a one-dimensional boundary-box regression task for text sequences and applied it to window-based STS, addressing the mismatch between task objectives and training objectives in text segmentation, thereby optimizing task objectives more directly. We propose the **One**-**D**imensional **O**bject **D**etection (1DOD) method, which jointly optimizes the model with a 1D boundary-box regression task and a sentence classification task, to enhance the performance of the text segmentation model.

- Our method has shown improvements over the SOTA models on the AMC dataset by 11.87 and 10.86 on the A and B respectively, and has reduced the traditional text segmentation evaluation metrics WD and P_k by 4.19 and 2.17 respectively. Furthermore, we have contributed a large-scale and extensive dataset, which has been used to validate our method. The results demonstrate that the 1DOD model possesses excellent generalization capability and universality.

2 Related Work

2.1 Text Segmentation

The popularity of deep neural networks has greatly propelled the development of the text segmentation field. Koshorek et al. (2018) and Wang et al. (2018) used Bi-LSTM to predict segment boundaries; SECTOR (Arnold et al., 2019) employs LSTM to predict sentence topics and merges them into paragraphs; Barrow et al. (2020) proposed S-LSTM for text segmentation; Xing et al. (2020) added a coherence auxiliary task to enhance the segmentation performance of the hierarchical attention BiLSTM; Yoo and Kim (2024) incorporated segment distance information in the model to control the paragraph segmentation length in novel texts; Glavas and Somasundaran (2020), Lo et al. (2021), and Lukasik et al. (2020) use hierarchical Transformer networks to extract sentence features for segmentation; Yu et al. (2023) proposed Topic-Aware Sentence Structure Prediction (TSSP) and Contrastive Semantic Similarity Learning (CSSP) to enhance topic-oriented text.

2.2 Dialogue Segmentation

Zhang and Zhou (2019) noted that speaker information helps with dialogue segmentation; Xing and Carenini (2021) and Solbiati et al. (2021) used an unsupervised Bert model to determine the thematic relevance or coherence of utterances; Zhang et al. (2021) utilized a self-adaptive sliding window for efficient text segmentation; Xu et al. (2021) proposed an unsupervised topic-aware segmentation algorithm and a Topic-Aware Dual Attention Matching (TADAM) network for dialogue segmentation; Xia et al. (2022) introduced the Neighbor Smoothing Parallel Extraction Network (PEN-NS) for dealing with segment boundary noise and ambiguity; (Gao et al., 2023) utilized neighboring utterance matching and pseudo-segmentation to enable the model to learn topic-aware utterance representations from unlabeled dialogue data.

Most existing works focus on incorporating additional topic information to aid segmentation, often neglecting the intrinsic characteristics of the text segmentation itself. In contrast, our approach does not introduce additional topic information but in-

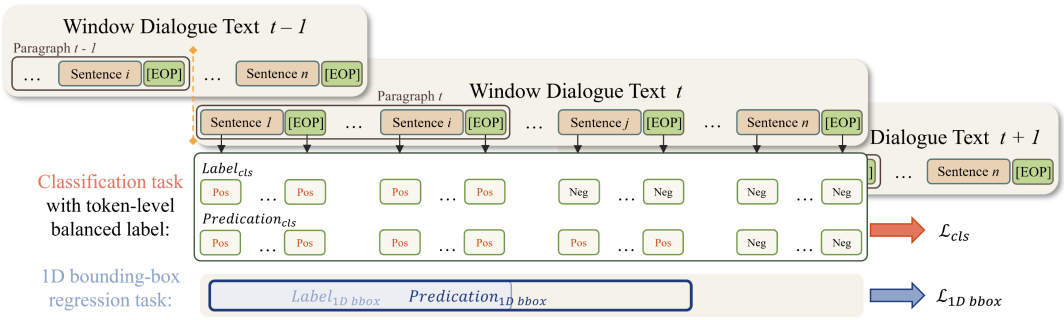

Figure 2: In each window of STS, the training objective is to minimize classification loss and 1D bounding-box regression loss. The labels used in the figure are at the token-level. It should be noted that our 1D bounding-box regression task is single-edged.

stead focuses on enhancing the model's semantic understanding for the STS task.

3 Method

3.1 Problem Formulation

In text segmentation tasks, given a dialogue text $T = [s_1, s_2, \ldots, s_n]$, we segment it into consecutive paragraphs $T = [P_1, P_2, \ldots, P_m] = [[s_1, \ldots, s_{p_1}], [s_{p_1+1}, \ldots, s_{p_2}], \ldots, [s_{p_{m-1}+1}, \ldots, s_n]]$, where $s_{p_i}, i \in [1, \ldots, m]$ is the last sentence of P_i. In the STS based on a self-adaptive sliding window, with a window size of x, for the j-th window text $T_{win_j} = [s_j, s_{j+1}, \ldots, s_{j+x-1}]$, the first paragraph of length l is segmented out as $T_{para_j} = [s_j, s_{j+1}, \ldots, s_{j+l-1}]$, with the remaining as background paragraphs, where s_j is the first sentence of the background paragraph from the previous window, and s_{j+l} is the start of the next window.

3.2 Streaming Text Segmentation with Balanced Labels

Traditional supervised text segmentation methods select the last sentence of each paragraph as a classification task, while the self-adaptive sliding window method only selects the end sentence of the first paragraph within each window range. However, these segmentation methods, when dealing with lengthy dialogue texts, result in extremely imbalanced sample labels. For example, when segmenting a conference dialogue text with a thousand utterances into ten paragraphs, the direct segmentation ratio of positive to negative samples will reach 1:99; if the window size is set to 200, the ratio in each window will reach 1:199.

Lin et al. (2017) pointed out that extreme category imbalance can lead to inefficient training and model degradation. To address this issue, under the STS architecture, we propose two levels of balanced labels to replace the original task labels, based on sentence granularity and token granularity, as shown in Figure 3.

In segmentation tasks, for utterance granularity, we add the $[EOP]$ token at the end of each sentence to indicate the end of a sentence. Typical text segmentation models classify each sentence's $[EOP]$ token to determine whether the sentence is the end of a paragraph. SeqModel (Zhang et al., 2021) uses the average pooled features of all tokens in the sentence for classification, with the same level of label sparsity as the former; in our proposed sentence-level segmentation task, within each window, the paragraph is determined by checking whether the $[EOP]$ tokens of the first l sentences belong to the paragraph; similar to the sentence-level, the token-level adjusts by judging all tokens in the first l sentences, not just the $[EOP]$ tokens.

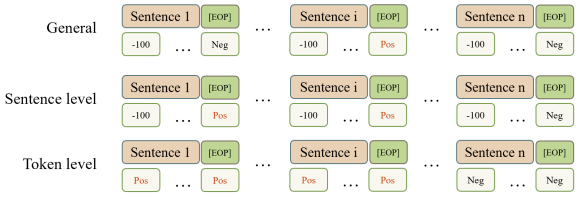

Figure 3: General text segmentation labels, sentence-level labels, and token-level labels. 'Pos' means it belongs to the target paragraph, while 'Neg' does not. '-100' refers to ignoring this label.

By pre-adjusting to obtain the appropriate window size, the overall ratio of positive to negative labels in the training data can be made as close as possible to 1:1. Training with balanced labels allows the model to fully learn from both positive and negative labels, and enables the training process to

be quick and stable. Compared to sentence-level, token-level training labels provide more and richer balanced supervision signals, which we believe can enable the model to acquire more comprehensive information.

3.3 One-Dimensional Object Detection

In every step of STS, the task objective is to segment the first paragraph of the window text. Inspired by object detection and image semantic segmentation tasks, we treat the first paragraph as a 1D target within serialized text, and adapt the object detection task to one-dimension for application in STS.

DETR (Carion et al., 2020) is the first model to apply the Transformer to object detection, using a combined loss function of L1 and GIoU (Generalized Intersection over Union) (Rezatofighi et al., 2019) for bounding-box regression training, and additionally employs Dice (Milletari et al., 2016) loss for panoptic segmentation. For STS, we retain the cross-entropy loss for fine-grained category discrimination, which is essential as it effectively guides the model to capture and learn the semantic information of the text. We replace it with $Focal$ (Lin et al., 2017) loss to better address imbalanced labels:

$$\begin{aligned}\mathcal{L}_{cls} = -\frac{1}{N}\sum_{t=1}^{N}(&\alpha(1-p_t)^{\gamma}y_t log p_t \\ &+ (1-\alpha)p_t^{\gamma}(1-y_t)log(1-p_t))\end{aligned} \quad (1)$$

Where, N is the number of samples, y_t is the true label of the t-th sample, p_t is the probability of predicting as the positive class, and α and γ are the hyperparameters of $Focal$.

We apply the smooth L1 loss function to 1D bounding-box regression, primarily modifying the 2D bounding boxes to 1D. Since the left boundary of the paragraph always coincides with the left boundary of the window, the segmentation point coordinate is equivalent to the boundary width. For the active label B of length l_t and the segmentation paragraph prediction $\widehat{B}$ of length l_p, their bounding boxes are $(\frac{1}{2}l_t, l_t)$ and $(\frac{1}{2}l_p, l_p)$ respectively, resulting in the smooth L1 loss function:

$$\mathcal{L}_{Smooth\ L1} = \frac{1}{N}\sum_{t=1}^{N}\sum_{i\in\{c,l\}} smooth_{L1}(\widehat{B}_i^t - B_i^t) \quad (2)$$

Where $\widehat{B}_i^t$ and B_i^t are the predicted and actual values of the i-th bounding box coordinate of the t-th sample respectively.

Then, we performed dimensionality reduction on the IoU (Yu et al., 2016) to one dimension and derived the loss function:

$$\begin{aligned}\mathcal{L}_{IoU} &= \frac{1}{N}\sum_{t=1}^{N}(1 - IoU_t) \\ &= \frac{1}{N}\sum_{t=1}^{N}(1 - \frac{Intersection(B^t, \widehat{B}^t)}{Union(B^t, \widehat{B}^t)})\end{aligned} \quad (3)$$

IoU has many variants, but most are not suitable for our 1D single-edge boundary regression, such as GIoU, CIoU (Zheng et al., 2020), and SIoU(Gevorgyan, 2022), because there is always an intersection when the predicted length is non-zero, and there is no aspect ratio or angle distinction. To accelerate training convergence, we have adapted the DIoU loss function to one dimension:

$$\mathcal{R}_{DIoU} = \frac{\rho(B, \widehat{B})}{Union(B, \widehat{B})} \quad (4)$$

$$\begin{aligned}\mathcal{L}_{DIoU} &= \frac{1}{N}\sum_{t=1}^{N}(1 - DIoU_t) \\ &= \frac{1}{N}\sum_{t=1}^{N}(1 - IoU_t + \mathcal{R}_{DIoU_t})\end{aligned} \quad (5)$$

Where $\rho(\cdot)$ represents the Euclidean distance, i.e., the distance between the midpoints of the two bounding boxes.

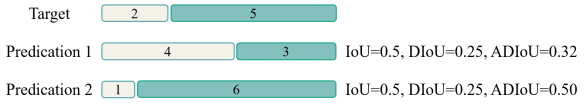

Figure 4: Comparison between the result of 1D IoU, DIoU and ADIoU.

However, both 1D IoU and DIoU do not address the issue of different scales having the same intersection over union ratio (Figure 4). Although cross-entropy loss and smooth L1 loss remain effective under such circumstances, we still explored solutions based on IoU. We simply calculated the average DIoU (ADIoU) between the segmented paragraph and the background paragraph as a metric and obtained the corresponding loss function:

$$ADIoU = \frac{1}{2}(DIoU + DIoU_{background}) \quad (6)$$

$$\mathcal{L}_{ADIoU} = \frac{1}{N}\sum_{t=1}^{N}(1 - ADIoU_t) \quad (7)$$

The 1D adaptation of the Dice loss function is similar to IoU; specifically, when using the ADIoU loss function, we use the corresponding average Dice (ADice):

$$Dice = 2 \cdot \frac{Intersection(B, \widehat{B})}{|B| + |\widehat{B}|} \quad (8)$$

$$ADice = \frac{1}{2}(Dice + Dice_{background}) \quad (9)$$

$$\mathcal{L}_{Dice} = \frac{1}{N}\sum_{t=1}^{N}(1 - Dice_t) \quad (10)$$

$$\mathcal{L}_{ADice} = \frac{1}{N}\sum_{t=1}^{N}(1 - ADice_t) \quad (11)$$

Overall, we obtain the total 1D bounding-box regression loss:

$$\mathcal{L}_{1D\ bbox} = \lambda_{SL1}\mathcal{L}_{Smooth\ L1} + \begin{cases} \lambda_{IoU}\mathcal{L}_{IoU} + \lambda_{Dice}\mathcal{L}_{Dice} \\ \lambda_{DIoU}\mathcal{L}_{DIoU} + \lambda_{Dice}\mathcal{L}_{Dice} \\ \lambda_{ADIoU}\mathcal{L}_{ADIoU} + \lambda_{ADice}\mathcal{L}_{ADice} \end{cases} \quad (12)$$

Where λ_{SL1}, λ_{IoU}, λ_{DIoU}, λ_{ADIoU}, λ_{Dice}, and λ_{ADice} are hyperparameters. Adding classification loss results in our final loss for 1DOD-based STS (Figure 2), which we apply in the training of the segmentation model:

$$\mathcal{L}_{STS} = \mathcal{L}_{cls} + \mathcal{L}_{1D\ bbox} \quad (13)$$

3.4 Tail Processing

When the window slides to the bottom, if the length of the background paragraph is less than the preset segmentation stop threshold t_{stop}, the streaming segmentation stops; otherwise, the background paragraph continues to be segmented as the next window text. But obviously, the last part of the dialogue text is not detected. This leads to a problem: the tail paragraphs are prone to being segmented too short, and setting a larger threshold cannot prevent it; a larger threshold can also lead to excessively long tail paragraphs.

In response, we propose a new retrieval-merge process as a tail handling for STS: first, set an appropriate stop threshold t_{stop} to prevent overly long tail paragraphs; secondly, set a new minimum tail length threshold t_{min}. After segmentation, if the tail paragraph is shorter than t_{min}, a forward merging algorithm is executed. In forward merging, we use a pretrained model to determine whether the consecutive segments are part of a continuous piece or two independent paragraphs; if they are continuous, they are merged.

4 Experimental Setup

Specific parameter settings can be found in the Parameter Setting.

4.1 Dataset

Due to the specificity of conferences, publicly available conference dialogue datasets are extremely scarce. The AMC corpus (AliMeeting4MUG Corpus) (Zhang et al., 2023) is currently the largest publicly available conference dialogue data set, and we used its 589 publicly segmented conference recordings for our experiments.

Furthermore, to validate the generalizability of our method, we collected and manually annotated 1553 long dialogue texts (MDT1553) covering multiple domains and topics for experimentation. In addition to conferences, MDT1553 also includes various fields such as broadcasts, live streams, speeches, and TV programs, covering multiple topics (Figure 7), complementing the AMC corpus. Statistical data for both datasets are shown in Table 5.

4.2 Evaluation Metric

Traditional text segmentation metrics P_k (Beeferman et al., 1999) and WD (Pevzner and Hearst, 2002) are sensitive to paragraph length and provide unreasonable assessments of various types of errors (Fournier and Inkpen, 2012; Fournier, 2013; Diaz and Ouyang, 2022), making them unsuitable for accurately evaluating long dialogue text segmentation tasks. B (Fournier, 2013) uses edit distance as a penalty term, improving the shortcomings of traditional metrics. A (Diaz and Ouyang, 2022) is based on paragraph matching and uses the Jaccard Index to calculate closeness, which is associated with our 1DOD task. Additionally, it is length-

independent, provides continuous penalties for segmentation boundary shifts, and applies appropriate penalties for different types of errors, which we believe better matches human evaluation methods for long dialogue texts. Therefore, we primarily rely on the A metric for evaluating our experiments, but we will also provide scores for P_k, WD, and B for reference. For P_k and WD, the sliding window size and the maximum distance for boundary consistency in B are set to half the average real paragraph length. For calculating the A index, we use the original code, while the rest use the segeval.

4.3 Baseline Models

The average length of the dialogue data exceeds 10000, which is longer than the maximum length supported by mainstream language models. To better adapt to long contexts, we experimented with two Chinese language models, Longformer (Beltagy et al., 2020) and PoNet (Tan et al., 2022), which support lengths up to 4096. Notably, the experimental results in Table 1 show that PoNet outperforms Longformer. Moreover, in the experiments, using either the Longformer or SeqModel-Longformer model failed to achieve effective segmentation, which we believe is due to the labels being too sparse for the model to be trained properly. Based on these results, we chose PoNet as the primary model for our experiments. For tail processing, we used Bert (Devlin et al., 2019) to determine whether to merge. Additionally, we selected results from other high-performance text segmentation models for comparison, where the context length for Cross-segment Bert was set at 256, and the settings for other models followed their original publications.

5 Experimental Results and Analysis

In this section, we first present the test results on AMC and the transfer test results on MDT1553, then analyze the ablation experiments, and finally demonstrate the stability of our method in various aspects.

5.1 Main Results

Tables 1 and 2 respectively show the results on the AMC test set and MDT1553. All models are supervised and trained only on the AMC training set, with the best model selected based on results from the AMC validation set. Our proposed IoU, DIoU, and ADIoU methods yield comparable results, with ADIoU performing best when using sentence-level balanced labels, and IoU performing best when using token-level balanced labels.

Models based on Bert perform poorly because they are limited by length and imbalanced segmentation labels. Furthermore, the quality of sentence embedding in hierarchical models is very important, Bert models are mainly pretrained on classification datasets, and their sentence embedding representations differ significantly from the requirements of text segmentation tasks. TSSP+CSSL results are poor, which we believe is due to the complexity of long dialogues. The inherent noise and thematic nestedness in dialogue texts make data augmentation and contrastive learning lose their original significance. SeqModel shows a significant improvement over the base model, but the imbalance in labels still restricts the performance of the model. Our 1DOD STS method based on SeqModel and significantly improves over SeqModel, with increases of 11.87 and 10.86 on the A and B respectively, thus demonstrating the effectiveness of balanced labels and 1DOD methods.

MDT1553 includes more diverse dialogue data, which can reflect the model's transferability across different dialogue topics. Without fine-tuning, we directly tested the MDT1553 dataset, and the best results of 1DOD showed significant improvements over SeqModel, with increases of 10.82 and 8.78 on the A and B respectively, still demonstrating substantial progress and exhibiting excellent generalization capabilities.

5.2 Ablation Study

Table 3 shows the results of the ablation studies. The accuracy of segmentation significantly improved just by adding balanced labels, but note that the $1 - P_k$ and $1 - WD$ slightly decreased, which we believe could be due to an increase in near-misses errors as the classification task shifted from predicting segmentation points to paragraphs; when 1DOD task was added, the model's performance improved again. Horizontally comparing, token-level balanced labels outperformed sentence-level because they introduced more supervision signals; vertically comparing, IoU and ADIoU performed better than DIoU, indicating that IoU loss was sufficient, and ADIoU performed better as it balanced between target and background paragraphs. Finally, appropriate tail processing compensated for the shortcomings of 1DOD, further improving the accuracy of STS based on the existing results.

Methods	A	B	$1-WD$	$1-P_k$
C99 (Choi, 2000)	15.07	4.85	46.94	51.80
TopicTiling (Riedl and Biemann, 2012)	19.03	9.88	45.02	51.43
$Transformer^2_{BERT}$ (Lo et al., 2021)	32.92	21.61	53.89	56.07
Cross-segment BERT (Lukasik et al., 2020)	35.20	20.76	53.65	55.46
PoNet (Tan et al., 2022)	43.12	31.17	56.99	59.98
PoNet+TSSP+CSSL (Yu et al., 2023)	42.48	30.16	56.21	58.79
SeqModel-PoNet (Zhang et al., 2021)	47.84	37.61	56.83	58.92
1DOD-Longformer-ADIoU (token-level)	58.41	43.88	58.34	58.78
1DOD-PoNet-IoU (token-level)	**59.71**	**48.47**	**61.18**	**62.15**

Table 1: Comparative experimental results based on the AMC dataset. We only listed the best results of 1DOD-Longformer and 1DOD-PoNet models, and it can be seen that our method significantly outperforms other text segmentation models. All metrics are better when higher.

Balanced Labels	Methods	A	B	$1-WD$	$1-P_k$
	C99	10.07	2.23	51.28	54.30
	TopicTiling	10.72	2.59	52.84	54.22
	$Transformer^2_{BERT}$	28.47	16.21	57.97	58.24
	Cross-segment BERT	30.51	16.16	57.22	56.69
	PoNet	36.83	27.78	**60.28**	**61.35**
	PoNet+TSSP+CSSL	36.43	27.21	59.49	60.94
	SeqModel-PoNet	39.74	29.41	58.19	59.88
Sentence-level	w/o 1DOD	47.91	35.27	52.48	54.10
	1DOD-PoNet-IoU	48.30	34.91	54.76	56.07
	1DOD-PoNet-DIoU	49.77	36.75	57.29	58.54
	1DOD-PoNet-ADIoU	49.13	35.91	57.63	58.78
Token-level	w/o 1DOD	48.28	35.42	53.38	54.94
	1DOD-PoNet-IoU	50.35	38.24	57.70	59.13
	1DOD-PoNet-DIoU	50.39	**38.62**	57.45	59.07
	1DOD-Longformer-ADIoU	46.38	31.25	52.94	53.99
	1DOD-PoNet-ADIoU	**50.56**	38.19	58.97	60.22

Table 2: Comparative experimental results for MDT1553. Models are trained only on the AMC training set to demonstrate generalization capabilities. For the 1DOD-Longformer, we only tested its best model.

Methods	Sentence-level				Token-level			
	A	B	$1-WD$	$1-P_k$	A	B	$1-WD$	$1-P_k$
$\mathcal{L}_{cls}$	55.46	45.21	55.67	56.77	55.63	45.58	55.81	56.98
$+tail\ process$	56.23	45.63	56.60	57.67	56.24	45.82	56.64	57.78
$\mathcal{L}_{cls}+\mathcal{L}_{1D\ bbox-IoU}$	58.70	46.73	59.19	59.80	59.17	48.21	60.29	61.26
$+tail\ process$	**59.69**	**47.20**	60.26	60.87	**59.71**	**48.47**	61.18	**62.15**
$\mathcal{L}_{cls}+\mathcal{L}_{1D\ bbox-DIoU}$	58.50	46.57	59.08	59.81	58.63	47.65	59.64	60.52
$+tail\ process$	59.11	46.72	59.92	60.64	59.23	47.97	60.61	61.48
$\mathcal{L}_{cls}+\mathcal{L}_{1D\ bbox-ADIoU}$	58.88	46.21	59.83	60.50	59.19	46.80	60.43	61.08
$+tail\ process$	59.29	46.22	**60.67**	**61.32**	59.62	46.83	**61.29**	61.92

Table 3: Ablation studies based on PoNet and AMC dataset, where $\mathcal{L}_{cls}$ indicates that the model was trained only on the classification task. The general method is shown in Table 1 "PoNet".

5.3 Stability Analysis

5.3.1 Stability of Training

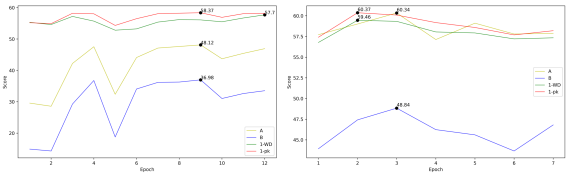

Figure 5: Evaluation results on the validation set during the training of SeqModel-PoNet (left) and 1DOD-PoNet-IoU (sentence-level, right), using the same training parameters.

In the experimental section, we selected the best model based on the evaluation results of the validation set for assessment. During training, the evaluation results of SeqModel and 1DOD for each round on the validation set are shown in Figure 5. The 1DOD model demonstrated good stability during training (with an A movement range of about 3), converging to the best level by only 2 epochs, and achieving higher evaluation scores than SeqModel, whereas SeqModel exhibited fluctuations in the early stages of training (with an A movement range of about 20). The fundamental cause of these fluctuations is the imbalance in training labels.

5.3.2 Stability of Segmentation

Tables 4 and Figure 6 display the statistical data and distribution of segment paragraph lengths for different models. The model's direct segmentation results differ most significantly from the actual labels due to the imbalance of segmentation labels and the irrelevance of paragraph length in sentence classification tasks; the self-adaptive sliding window model slightly improved this by the limitation of the window size, but it was still insufficient as the model did not learn paragraph length information. Models with added balanced labels resulted in overly stable segmentation outcomes, still differing from the actual distribution. The 1DOD method, by incorporating a 1D bounding-box regression loss, enhances the model's understanding of the "target area range", achieving an effect similar to that of Yoo and Kim (2024), who directly incorporated paragraph length information into the loss function, yet in a more intuitive manner. This makes the 1DOD model's predictions more consistent with the actual labels' distribution.

	AMC-test		MDT1553	
	Avg	Std	Avg	Std
Label	**31.25**	**19.20**	**31.10**	**29.23**
PoNet	41.78	46.34	74.67	76.89
SeqModel	35.15	35.21	45.20	44.70
SeqModel$_{TBL}$	26.34	15.40	29.36	16.57
1DOD-IoU	29.70	16.46	31.85	18.51
1DOD-DIoU	29.92	16.09	31.08	17.83
1DOD-ADIoU	32.51	19.14	34.33	23.52

Table 4: The average and standard deviation of the paragraph lengths segmented by different models. 'TBL' refers to token-level balanced label.

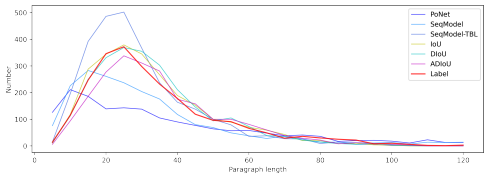

Figure 6: Comparison of segment paragraph length distributions for different models.

6 Conclusion

In text segmentation tasks, the imbalance of labels affects the stability and convergence of training, and the discrepancy between task objectives and training objectives impacts model performance. In response, we proposed balanced labels and a 1DOD task under the STS framework to address these two pain points. Experimental results show that our method far surpasses previous text segmentation models and possesses strong generalization capabilities and stability. In the future, we plan to extend this work to segmentation applications across various types of streaming media.

Limitations

During the segmentation process of streaming dialogue texts, the model may model the same text with different contexts multiple times, which reduces efficiency. Moreover, our method is suited for data where the paragraph length distribution follows a normal distribution, which is the case for most dialogue text data. For other special distributions (such as U-shaped or J-shaped) of dialogue texts, which is not the intended direction for STS as a preprocessing task for dialogue texts. For these special types of dialogue texts, our model is unable to achieve better results.

References

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Trans. Assoc. Comput. Linguistics*, 7:169–184.

Joe Barrow, Rajiv Jain, Vlad I. Morariu, Varun Manjunatha, Douglas W. Oard, and Philip Resnik. 2020. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 313–322. Association for Computational Linguistics.

Doug Beeferman, Adam L. Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210.

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

Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*, pages 211–218. ACM.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 26–33. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Gerardo Ocampo Diaz and Jessica Ouyang. 2022. An alignment-based approach to text segmentation similarity scoring. In *Proceedings of the 26th Conference on Computational Natural Language Learning, CoNLL 2022, Abu Dhabi, United Arab Emirates (Hybrid Event), December 7-8, 2022*, pages 374–383. Association for Computational Linguistics.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1702–1712. The Association for Computer Linguistics.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 152–161. The Association for Computational Linguistics.

Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Unsupervised dialogue topic segmentation with topic-aware utterance representation. *CoRR*, abs/2305.02747.

Zhora Gevorgyan. 2022. Siou loss: More powerful learning for bounding box regression. *CoRR*, abs/2205.12740.

Goran Glavas and Swapna Somasundaran. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7797–7804. AAAI Press.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 469–473. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2190–2196. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray L. Buntine. 2021. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In *Findings of the Association for Computational Linguistics: EMNLP 2021,*

Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, pages 3334–3340. Association for Computational Linguistics.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4707–4716. Association for Computational Linguistics.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 565–571. IEEE Computer Society.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguistics*, 28(1):19–36.

Mengnan Qi, Hao Liu, Yuzhuo Fu, and Ting Liu. 2021. Improving abstractive dialogue summarization with hierarchical pretraining and topic segment. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1121–1130. Association for Computational Linguistics.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 658–666. Computer Vision Foundation / IEEE.

Martin Riedl and Chris Biemann. 2012. Topictiling: a text segmentation algorithm based on lda. In *Proceedings of ACL 2012 student research workshop*, pages 37–42.

Felix Schneider and Marco Turchi. 2023. Team zoom@ automin 2023: Utilizing topic segmentation and llm data augmentation for long-form meeting summarization. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 101–107.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Calì. 2021. Unsupervised topic segmentation of meetings with BERT embeddings. *CoRR*, abs/2106.12978.

Chao-Hong Tan, Qian Chen, Wen Wang, Qinglin Zhang, Siqi Zheng, and Zhen-Hua Ling. 2022. Ponet: Pooling network for efficient token mixing in long sequences. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 962–967. Association for Computational Linguistics.

Jinxiong Xia, Cao Liu, Jiansong Chen, Yuchen Li, Fan Yang, Xunliang Cai, Guanglu Wan, and Houfeng Wang. 2022. Dialogue topic segmentation via parallel extraction network with neighbor smoothing. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2126–2131. ACM.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2021, Singapore and Online, July 29-31, 2021*, pages 167–177. Association for Computational Linguistics.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 626–636. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14176–14184. AAAI Press.

Byunghwa Yoo and Kyung-Joong Kim. 2024. Improving paragraph segmentation using BERT with additional information from probability density function modeling of segmentation distances. *Nat. Lang. Process. J.*, 6:100061.

Seunghyun Yoon, Joongbo Shin, and Kyomin Jung. 2018. Learning to rank question-answer pairs using hierarchical recurrent encoder with latent topic clustering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1575–1584. Association for Computational Linguistics.

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. 2023. Improving long document topic segmentation models with enhanced coherence modeling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 5592–5605. Association for Computational Linguistics.

Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas S. Huang. 2016. Unitbox: An advanced object detection network. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 516–520. ACM.

Leilan Zhang and Qiang Zhou. 2019. Topic segmentation for dialogue stream. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019, Lanzhou, China, November 18-21, 2019*, pages 1036–1043. IEEE.

Qinglin Zhang, Qian Chen, Yali Li, Jiaqing Liu, and Wen Wang. 2021. Sequence model with self-adaptive sliding window for efficient spoken document segmentation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 411–418. IEEE.

Qinglin Zhang, Chong Deng, Jiaqing Liu, Hai Yu, Qian Chen, Wen Wang, Zhijie Yan, Jinglin Liu, Yi Ren, and Zhou Zhao. 2023. MUG: A general meeting understanding and generation benchmark. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. 2020. Distance-iou loss: Faster and better learning for bounding box regression. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12993–13000. AAAI Press.

Lin Zhu, Xinnan Dai, Qihao Huang, Hai Xiang, and Jie Zheng. 2019. Topic judgment helps question similarity prediction in medical FAQ dialogue systems. In *2019 International Conference on Data Mining Workshops, ICDM Workshops 2019, Beijing, China, November 8-11, 2019*, pages 966–972. IEEE.

A Data Statistics and Distribution

		No.	#Utts.	#Pars.
AMC	Train	295	469.72	11.06
	Validation	65	463.79	11.32
	Test	229	326.79	10.46
MDT1553	All	1553	495.99	15.95

Table 5: Data statistics of AMC and MDT1553 datasets. # means average.

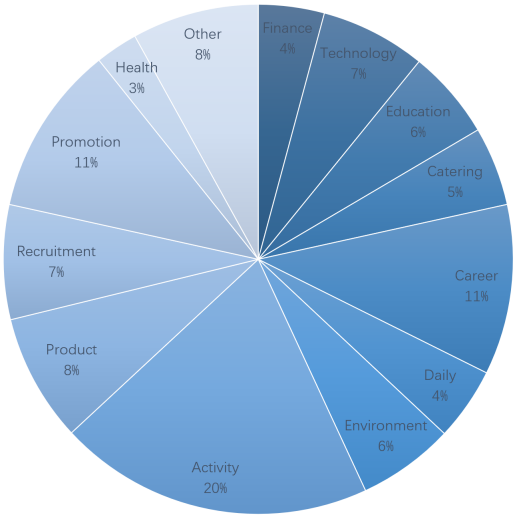

Figure 7: Topic distribution of MDT1553.

B Parameter Setting

Our experiments were conducted on the PyTorch 2.0.0. During training, we used the AdamW optimizer, with a batch size and learning rate set at 2 and 5e-5, respectively; α and γ were set at 0.75 and 2; λ_{SL1}, λ_{IoU}, and λ_{Dice} were set at 10, 0.5, and 0.5 respectively, under this setting, the balance between various losses can be maintained.

The hyperparameters for inference were obtained through experiments on the validation set, and since the AMC dataset and MDT1553 dataset are similar in length and format, the same set of parameters was used. The window size was set to 100, t_{stop} to 31, and t_{min} to 23. Our experiments were conducted on GeForce RTX-4090 GPUs and Linux OS.

C Ablation Study based on Longformer

The experimental results are shown in Table 6.

D Comparison with LLMs

We selected several mainstream online large language models for comparison (Table 7). We conducted experiments using the top 65 samples from the AMC test set, with all LLMs operating in a zero-shot setting.

E Time Consumption Comparison

We compared the baseline model and 1DOD on the AMC test set in terms of average text segmentation time under identical hardware conditions and parameter settings (batch size=1), with the results shown in Table 8.

Methods	Sentence-level				Token-level			
	A	B	$1-WD$	$1-P_k$	A	B	$1-WD$	$1-P_k$
$\mathcal{L}_{cls}$	56.29	42.33	54.86	55.43	56.60	41.25	54.94	55.40
$+tail\ process$	56.68	42.24	55.76	56.32	56.83	41.06	55.95	56.41
$\mathcal{L}_{cls} + \mathcal{L}_{1D\ bbox-IoU}$	57.21	42.91	56.74	57.29	57.03	43.68	56.30	56.91
$+tail\ process$	**57.91**	43.17	**57.83**	**58.37**	57.34	43.53	57.20	57.80
$\mathcal{L}_{cls} + \mathcal{L}_{1D\ bbox-DIoU}$	56.91	42.94	55.93	56.48	57.10	43.63	56.52	57.34
$+tail\ process$	57.58	43.15	57.02	57.56	57.65	43.70	57.54	58.37
$\mathcal{L}_{cls} + \mathcal{L}_{1D\ bbox-ADIoU}$	56.57	44.85	56.11	57.14	58.08	**43.91**	57.33	57.78
$+tail\ process$	57.25	**45.07**	57.34	58.20	**58.41**	43.88	**58.34**	**58.78**

Table 6: Ablation study based on longformer.

Methods	A	B	$1-WD$	$1-P_k$
GLM-4	24.59	20.19	18.16	47.80
Moonshot-v1-32k	32.05	26.75	28.60	48.89
GPT-4 Turbo	31.68	26.37	25.53	47.76
GPT-4o	38.21	30.41	34.71	47.81
Ours	**59.10**	**48.08**	**60.75**	**61.84**

Table 7: Comparison results with online large models.

Methods	Avg. Time (s)
PoNet	0.86
SeqModel-PoNet	1.47
1DOD-PoNet-IoU (token-level)	1.54
GPT-4o	5.84

Table 8: Experimental results of time consumption comparison.