Unsupervised Dialogue Topic Segmentation with Topic-aware Utterance Representation

Haoyu Gao[*][†]
University of Science and Technology of China
haoyugao183@gmail.com

Rui Wang[*]
Alibaba Group
wr224079@alibaba-inc.com

Ting-En Lin
Alibaba Group
ting-en.lte@alibaba-inc.com

Yuchuan Wu
Alibaba Group
shengxiu.wyc@alibaba-inc.com

Min Yang[‡]
SIAT, Chinese Academy of Sciences
min.yang@siat.ac.cn

Fei Huang, Yongbin Li[‡]
Alibaba Group
shuide.lyb@alibaba-inc.com

ABSTRACT

Dialogue Topic Segmentation (DTS) plays an essential role in a variety of dialogue modeling tasks. Previous DTS methods either focus on semantic similarity or dialogue coherence to assess topic similarity for unsupervised dialogue segmentation. However, the topic similarity cannot be fully identified via semantic similarity or dialogue coherence. In addition, the unlabeled dialogue data, which contains useful clues of utterance relationships, remains underexploited. In this paper, we propose a novel unsupervised DTS framework, which learns topic-aware utterance representations from unlabeled dialogue data through neighboring utterance matching and pseudo-segmentation. Extensive experiments on two benchmark datasets (i.e., DialSeg711 and Doc2Dial) demonstrate that our method significantly outperforms the strong baseline methods. For reproducibility, we provide our code and data at: https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/dial-start.

CCS CONCEPTS

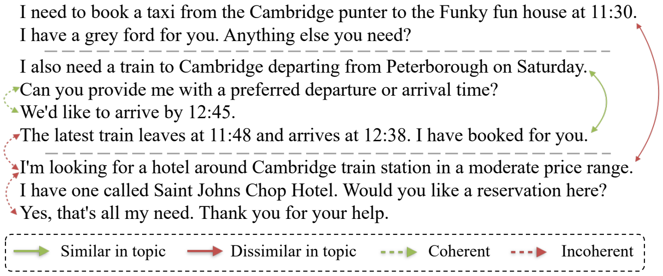
• **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

Dialogue understanding, self-supervised learning, dialogue topic segmentation, text segmentation

1 INTRODUCTION

Dialogue Topic Segmentation (DTS) aims to divide a dialogue into multiple segments, wherein the utterances within each segment are similar in topic. DTS is critical in a variety of down-steam dialogue modeling tasks, such as dialogue summarization [3, 19, 23, 27, 41], dialogue generation [18, 24, 35, 38], response prediction [16, 21, 22, 28, 36] and question answering [5, 37, 39].



I need to book a taxi from the Cambridge punter to the Funky fun house at 11:30.
I have a grey ford for you. Anything else you need?
I also need a train to Cambridge departing from Peterborough on Saturday.
Can you provide me with a preferred departure or arrival time?
We'd like to arrive by 12:45.
The latest train leaves at 11:48 and arrives at 12:38. I have booked for you.
I'm looking for a hotel around Cambridge train station in a moderate price range.
I have one called Saint Johns Chop Hotel. Would you like a reservation here?
Yes, that's all my need. Thank you for your help.

Figure 1: Comparison between topic similarity and dialogue coherence.

Most existing methods for DTS follow an unsupervised paradigm, due to the high cost of collecting accurate DTS annotations to train supervised models [13, 25, 33, 40] and the variability of annotation instructions across different domains. These methods generally involve two stages. First, various approaches are introduced to assess the topic similarity between the two sides of each potential segment boundary (i.e., an interval between two utterances). Second, a segmentation algorithm, such as *TextTiling* [17], is used to identify the segment boundaries. Previous studies usually assess topic similarity through dialogue coherence or semantic similarity computed by surface features, such as lexical overlap [4, 12, 30]. In recent years, Song et al. [32] assess semantic similarity using pre-trained word embeddings. The method is later extended to use sentence embeddings from pre-trained language models [14, 15, 31, 36], such as BERT [6] and SentenceBERT [29]. Xing et al. [34] further propose Coherence Scoring Model (CSM), which employs utterance-pair coherence to assess topic similarity.

Despite the remarkable progress of previous unsupervised DTS studies, several technical challenges related to modeling topic similarity and utilizing unlabeled dialogue data have not been fully resolved. First, prior methods typically rely on generic semantic similarity or dialogue coherence to assess topic similarity, but these measures are insufficient to capture it fully. Specifically, utterances that share the same topic may not be semantically similar, and vice versa. As illustrated in Figure 1, dialogue coherence refers to

[*] Equal Contribution.
[†] Haoyu Gao is also with SIAT, Chinese Academy of Sciences. This work was conducted when Haoyu Gao was interning at Alibaba.
[‡] Min Yang and Yongbin Li are corresponding authors.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592081

the response relation between an utterance and its preceding context [7], reflecting whether adjacent utterances are linked together. However, two non-adjacent utterances in the same topic segment may be topically similar but not coherent. Second, unlabeled dialogue data containing useful clues about utterance relationships is beneficial for unsupervised DTS. However, it has not been effectively leveraged in prior works. In the semantic similarity-based methods, word or sentence embeddings are pre-trained on generic textual corpora and supervised Natural Language Inferring (NLI) datasets [29], which are unsuitable for unlabeled dialogue data. In the coherence-based methods, CSM [34] learns dialogue coherence from the DailyDialog dataset [20] without DTS annotations. However, each of these dialogues is about one single topic, and CSM utilizes the dialogue-level topic labels to produce training samples.

To address the above issues, we propose a novel unsupervised DTS framework, called DialSTART (Unsupervised **Dial**ogue Topic **S**egmentation with **T**opic-**A**ware Utterance **R**epresen**T**ation), which learns topic-aware utterance representations from unlabeled dialogue data through neighboring utterance matching (NUM) and pseudo-segmentation. These topic-aware utterance representations are subsequently utilized in combination with the dialogue coherence to perform unsupervised segmentation. That is, neighboring utterances referring to those appearing together in one dialogue within a certain distance, are more likely to be topically similar. In unlabeled multiple-topic dialogues, such self-supervision of neighboring utterances is prone to be relatively noisy. In order to reduce the noise, we further combine the neighboring relation with pseudo-segmentation to produce refined utterance pairs that are assumed to be topically similar or dissimilar.

In practice, we first acquire topic-aware utterance representations via an utterance encoder. Second, for each utterance interval, we assess the relevance score which reflects the degree to which the two sides are within the same segment. These relevance scores are utilized by the TextTiling algorithm to perform segmentation. The segmentation results are used for inference or for generating pseudo-segmentation during training. Third, we generate topically similar and dissimilar pairs for each utterance based on its neighboring utterances and pseudo-segmentation. Finally, we fine-tune the utterance encoder to distinguish between topically similar pairs and dissimilar pairs through the marginal ranking loss. We also design a relevance modeling task to optimize the whole relevance score by distinguishing between real and synthetic fragments.

We conduct experiments on two dialogue topic segmentation datasets. The results show that our framework outperforms the state-of-the-art method by 8.03% on average in terms of Pk error. Further ablative experiments validate the effectiveness of our topic similarity modeling based on the NUM task and pseudo-segmentation for unsupervised DTS. Our contributions are three-fold:

(1) We introduce the Neighboring Utterance Matching (NUM) task to learn topic-aware utterance representations, and exploit both topic similarity and dialogue coherence to perform unsupervised dialogue topic segmentation.
(2) We propose to further reduce the self-supervision noise in the NUM task on unlabeled dialogue data by pseudo-segmentation, to obtain topically similar and dissimilar utterance pairs
(3) Experiments have demonstrated that our novel framework outperforms the state-of-the-art method significantly and the effectiveness of topic-aware utterance representation.

2 METHOD

In this section, we introduce our proposed approach. We start with the problem formulation followed by the model overview. Then, we describe topic-aware utterance representation and the training objectives.

2.1 Problem Formulation

Dialogue topic segmentation aims to identify segment boundaries in a dialogue. Formally, given a dialogue D which contains a sequence of n utterances $D = \{u_1, u_2, ..., u_n\}$, there are n - 1 intervals between adjacent utterances, denoted by $V = \{v_1, v_2, ..., v_{n-1}\}$. A segmentation algorithm predicts segment boundaries as $B = \{b_1, b_2, ..., b_k\}$, where k denotes the number of boundaries and b_i represents the dialogue is divided at b_i-th interval.

Most methods for unsupervised DTS follow a two-stage paradigm. First, for the interval v_i which locates between u_i and u_{i+1}, a relevance score r_i is computed. The higher the score, the more likely the two sides of the interval belong to the same segment. Then, given the relevance scores $R = \{r_1, r_2, ..., r_{n-1}\}$, a segmentation algorithm, such as TextTiling [17] or one of its derivatives, is utilized to determine the segmentation boundaries. Previous methods typically assess the relevance score relying on generic semantic similarity or dialogue coherence. We propose to model that relevance score exploiting both dialogue coherence and *topic similarity* derived from topic-aware utterance representations.

2.2 Model overview

As illustrated in Figure 2, our segmentation model consists of a topic encoder, a coherence encoder, and a segmentation algorithm. To get a better utterance representation initialization, we choose SimCSE[11] to initialize the topic encoder of our method. SimCSE is a simple but effective contrastive sentence embedding framework. We pass u_i into our topic encoder to obtain the topic representations of each utterance:

$$\mathrm{h_i} = \mathrm{SimCSE}(u_i), \tag{1}$$

where $\mathrm{h_i} \in \mathbb{R}^{d_h}$ denotes the pooled output of last layer of SimCSE, d_h is the dimension of hidden state. Following CSM [34], we choose the Next Sentence Prediction (NSP) BERT [6] as our coherence encoder. For each interval v_i, the coherence encoder calculates the coherence score as

$$c_i = \text{NSP-BERT}([u_{i-1}; u_i], u_{i+1}), \tag{2}$$

where u_{i+1} is the response and $[u_{i-1}; u_i]$ is the concatenated preceding context. After obtaining the topic representation $\mathrm{h_i}$ and coherence score c_i, we calculate the relevance score r_i as

$$r_i = \mathrm{sim}\left(\frac{\mathrm{h}_{i-1} + \mathrm{h}_i}{2}, \frac{\mathrm{h}_{i+1} + \mathrm{h}_{i+2}}{2}\right) + c_i \tag{3}$$

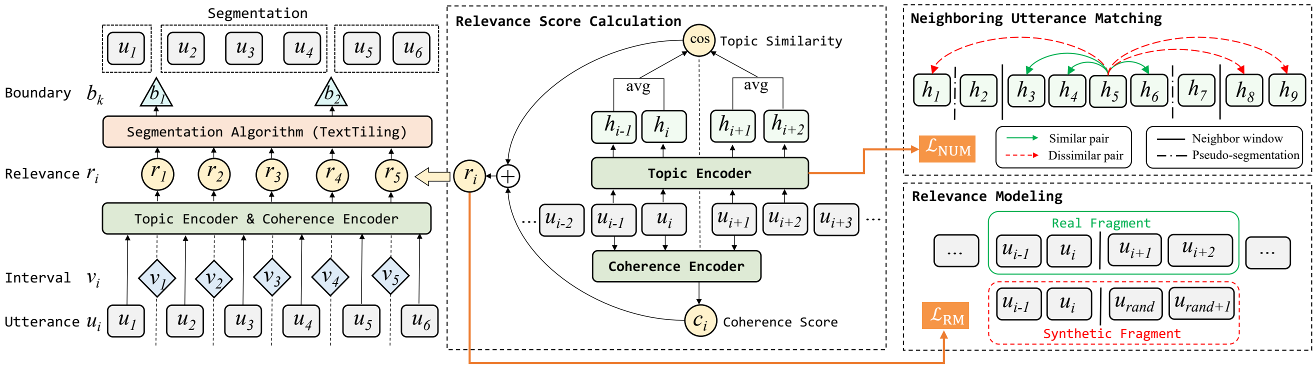

Segmentation
u_1 u_2 u_3 u_4 u_5 u_6
Boundary b_k
b_1 b_2
Segmentation Algorithm (TextTiling)
Relevance r_i
r_1 r_2 r_3 r_4 r_5
Topic Encoder & Coherence Encoder
Interval v_i
v_1 v_2 v_3 v_4 v_5
Utterance u_i
Relevance Score Calculation
cos
Topic Similarity
avg
avg
h_{i-1} h_i h_{i+1} h_{i+2}
Topic Encoder
r_i
u_{i-2} u_{i-1} u_i u_{i+1} u_{i+2} u_{i+3}
Coherence Encoder
c_i
Coherence Score
Neighboring Utterance Matching
h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9
$\mathcal{L}_{\mathrm{NUM}}$
Similar pair
Dissimilar pair
Neighbor window
Pseudo-segmentation
Relevance Modeling
Real Fragment
... u_{i-1} u_i u_{i+1} u_{i+2} ...
$\mathcal{L}_{\mathrm{RM}}$
u_{i-1} u_i u_{rand} u_{rand+1}
Synthetic Fragment

Figure 2: The overview of our model. The left part shows the illustration of segmentation. The middle part represents how to calculate the relevance score based on the output of the topic and coherence encoder. The right part shows the Neighboring Utterance Matching and Relevance Modeling tasks with positive and negative samples.

where $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity. The relevance score r_i is then used by the segmentation algorithm to perform segmentation.

Following most of previous work[32, 36], we choose TextTiling [17] as our segmentation algorithm. We apply TextTiling on $R = \{r_1, r_2, ..., r_{n-1}\}$ to obtain segment boundaries B.

$$B = \mathrm{TextTiling}(R). \tag{4}$$

To train our encoders, we present two self-supervised task, the neighboring utterance matching task for training the topic encoder and the relevance modeling task for training both encoders.

2.3 Topic-aware utterance Representation

In order to train our segmentation model with topic-aware capability, we propose a novel task named Neighboring Utterance Matching (NUM). Based on the nature of topic change, we make an assumption that utterance is more likely to be topically similar to its neighboring utterances. To further reduce the noise in unlabeled dialogues, we combine neighboring utterances and pseudo-segmentation to get refined topically similar utterance pairs and dissimilar pairs. Then we take two kinds of pairs as the positive and negative samples for the marginal ranking loss.

First, given an utterance u_i in D , we define its neighboring utterance index set U_i and non-neighboring utterance index set $\overline{U_i}$ as:

$$U_i = \{j \in [1,n] \mid w \geq |i-j| \ \wedge \ j \neq i\}, \tag{5}$$

$$\overline{U_i} = \{j \in [1,n] \mid w < |i-j|\}, \tag{6}$$

where w is the number of neighboring utterances before and after u_i, n refers to the length of the dialogue D.

In unlabeled multiple-topic dialogues, the supervision from NUM is prone to be relatively noisy. In order to reduce the noise, we further combine the neighboring relation with the pseudo-segmentation to produce refined utterance pairs that are assumed to be topically similar or dissimilar. Given u_i and its pseudo segment $segment(i)$, W_i denotes the utterances index inside $segment(i)$ and $\overline{W_i}$ denotes those outside $segment(i)$:

$$W_i = \{j \in [1,n] \mid u_j \in \mathrm{segment}(u_i) \ \wedge j \neq \ i\}, \tag{7}$$

$$\overline{W_i} = \{j \in [1,n] \mid u_j \notin \mathrm{segment}(u_i)\}. \tag{8}$$

Based on the neighboring utterances and pseudo segment of u_i, we obtain refined topically similar utterances index P_i^+ and refined topically dissimilar utterances index P_i^- for u_i as

$$P_i^+ = U_i \cap W_i, \quad P_i^- = \overline{U_i} \cap \overline{W_i}. \tag{9}$$

The topic encoder is optimized through the marginal ranking loss:

$$\mathcal{L}_{\mathrm{NUM}}(u_i) = \frac{1}{|P_i^+| \cdot |P_i^-|} \sum_{p^+ \in P_i^+} \sum_{p^- \in P_i^-} \max(0, \eta + e_{i,p}^- - e_{i,p}^+), \tag{10}$$

where η is the margin hyper-parameter, $e_{i,p}^-$ and $e_{i,p}^+$ denote for $\mathrm{sim}(h_i, h_{p^-})$ and $\mathrm{sim}(h_i, h_{p^+})$, respectively.

2.4 Training Objectives

As illustrated in Figure 2, we utilize two training objectives to enable our segmentation model to capture both dialogue coherence and topic similarity among utterances, combining them for improved segmentation performance. To model topic similarity and obtain better topic representations, we introduce the Neighboring Utterance Matching (NUM) task. Furthermore, based the utterance-pair coherence scoring used in CSM [34], we extend it to our relevance modeling (RM) task. Relevance Modeling (RM) is designed to optimize the topic and coherence encoders to produce high-quality relevance scores. To achieve this, we aim to distinguish between real and synthetic fragments based on the output of the topic and coherence encoders. Specifically, for each interval v_i, a real fragment consists of the utterances around v_i within a certain distance. In contrast, the synthetic fragment is created by randomly substituting the right-side utterances of v_i while keeping the left-side utterances fixed. To simulate various topic transitions, we employ two sampling schemes: 1) sampling utterances only within the same dialogue; 2) randomly sampling utterances from other dialogues. After obtaining the real and synthetic fragments, we model relevance by ranking real fragments higher than synthetic ones. Formally, the relevance score of real fragment r_i^+ centered at v_i is supposed to be higher than the relevance score of the synthetic fragment r_i^-. We calculate marginal ranking loss $\mathcal{L}_{\mathrm{RM}}$ as follows:

$$\mathcal{L}_{\mathrm{RM}}(v_i) = \max(0, \eta + r_i^- - r_i^+). \tag{11}$$

Table 1: Experimental results on DialSeg711 and Doc2Dial.

Method	DialSeg711		Doc2Dial	
	$P_k \downarrow$	$WD \downarrow$	$P_k \downarrow$	$WD \downarrow$
BayesSeg [8]	30.97	35.60	46.65	62.13
GraphSeg [12]	43.74	44.76	51.54	51.59
GreedySeg [36]	50.95	53.85	50.66	51.56
TextTiling (TeT) [17]	40.44	44.63	52.02	57.42
TeT + Embedding [32]	39.37	41.27	53.72	55.73
TeT + CLS [36]	40.49	43.14	54.34	57.92
TeT + NSP [34]	46.84	48.50	50.79	54.86
CSM [34]	26.80	28.24	45.23	47.32
CSM (unsup)	24.30	26.35	45.30	49.84
Ours	**17.86**	**19.80**	**38.11**	**40.72**

Overall, the final training loss is defined as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\mathrm{NUM}}(u_i) + \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{\mathrm{RM}}(v_i), \tag{12}$$

where N and M is the size of the training data of the NUM task and the RM task, respectively.

3 EXPERIMENTS

3.1 Experimental Setup

Datasets. We evaluate our method on two widely used datasets: DialSeg711 and Doc2Dial. DialSeg711 [36] is a real-world dataset including 711 English dialogues that combines dialogues from two existing task-oriented dialogue datasets, MultiWOZ [2] and KVRET [9]. This dataset contains 4.9 topic segments per dialogue and 5.6 utterances per topic segment on average. The Doc2Dial [10] dataset comprises more than 4,100 synthetic English conversations grounded in over 450 documents belonging to four domains. This dataset contains 3.7 topic segments per dialogue and 3.5 utterances per topic segment on average.

Evaluation Metrics. To ensure a fair comparison, we employ two standard metrics, namely P_k error score [1] and WinDiff (WD) [26]. Both P_k and WD are computed by measuring the overlap between the ground-truth segments and the model's predictions within a sliding window of a certain size.

Implementation details. We start from the pre-trained checkpoint of the `sup-simcse-bert-base-uncased` version of SimCSE. For both DialSeg711 and Doc2Dial, we choose the number of neighboring utterances w as 5 for performance and computational efficiency.

3.2 Main Results

We compare our proposed method with two types of unsupervised baselines: (1) those not using TextTiling including BayesSeg [8], GraphSeg [12] and GreedySeg [36], and (2) those extended from TextTiling such as TeT [17], TeT+CLS [36], TeT+Embedding [32], and Coherence Scoring Model (CSM) [34]. Note that CSM uses dialogue coherence instead of semantic similarity, which is different from other unsupervised baselines extended from TextTiling. CSM adopts DailyDialog [20] as the training data, utilizing the annotations of the dialogue-level topic and utterance-level dialogue act. We also present an improved variant of CSM denoted as CSM (unsup), which does not require topic labels and act labels. Likewise, our framework leverages the unlabeled dialogues for training, while the segmentation annotations are only used for evaluation.

Table 2: Ablation study results.

Method	DialSeg711		Doc2Dial	
	$P_k \downarrow$	$WD \downarrow$	$P_k \downarrow$	$WD \downarrow$
Ours w/o Topic encoder	27.60	29.84	41.34	44.36
Ours w/o NUM task	20.64	22.25	40.40	44.18
Ours w/o Pseudo-seg.	29.92	31.77	44.10	48.29
Ours	**17.86**	**19.80**	**38.11**	**40.72**

Table 1 presents the results of our model and baselines on two datasets. Our model achieves state-of-the-art (SOTA) performance on both evaluation datasets, with varying distances from the previous SOTA. Specifically, we were able to push the performance on DialSeg711 by another about 9% absolute both P_k error score and WD, achieving 17.86% P_k error score and 19.80 % WD. The improvements on Doc2Dial, a more complex and larger dataset, are absolute 7% on both P_k error score and WD, bringing the SOTA to 38.11% on P_k error score and 40.72% on WD. This demonstrates that our model benefits from learning topic similarity and dialogue coherence through effectively exploiting unlabeled dialogues. Furthermore, the CSM(unsup) cannot fully utilize unlabeled data for better improvement, demonstrating the challenge of exploiting unlabeled dialogue.

3.3 Ablation Study

We investigate the impact of the NUM task and pseudo-segmentation via three different settings: 1) discarding the topic encoder and only utilizing the coherence encoder; 2) keeping the topic encoder but removing the NUM task; 3) keeping the NUM task but removing the pseudo-segmentation. We present our ablation study results in Table 2. Comparing our approach with the "w/o topic encoder" setting, we observe a significant performance drop on both datasets, indicating that topic similarity is crucial for obtaining good performance, as it prevents local dialogue coherence from dominating the topic segmentation. Removing the NUM task in our method leads to a drop in performance on both datasets to varying degrees, which confirms the effectiveness of our proposed NUM task in encouraging the topic encoder to learn topic-aware utterance representations. Additionally, we found that noise in unlabeled multiple-topic dialogues could mislead the topic encoder's learning, as reflected in the decreased performance of the "w/o pseudo-segmentation" setting.

4 CONCLUSION

In this paper, we propose a novel unsupervised dialogue topic segmentation framework, which learns topic-aware utterance representations from unlabeled dialogue data through neighboring utterance matching (NUM) and pseudo-segmentation. Extensive experiments on two benchmark datasets show that our method significantly outperforms previous state-of-the-art by simultaneously utilizing dialogue coherence and topic similarity.

ACKNOWLEDGMENTS

Min Yang was partially supported by National Key Research and Development Program of China (2022YFF0902100), Shenzhen Science and Technology Innovation Program (KQTD20190929172835662), Shenzhen Basic Research Foundation (JCYJ20210324115614039 and JCYJ20200109113441941). This work was supported by Alibaba Group through Alibaba Innovative Research Program.

REFERENCES

[1] Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning* 1, 34 (1999), 177–210.

[2] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 5016–5026.

[3] Jiaao Chen and Diyi Yang. 2020. Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4106–4118.

[4] Freddy YY Choi. 2000. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

[5] Yinpei Dai, Wanwei He, Bowen Li, Yuchuan Wu, Zheng Cao, Zhongqi An, Jian Sun, and Yongbin Li. 2022. CGoDial: A Large-Scale Benchmark for Chinese Goal-oriented Dialog Evaluation. *arXiv preprint arXiv:2211.11617* (2022).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[7] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Evaluating Coherence in Dialogue Systems using Entailment. 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)*. 3806–3812.

[8] Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 334–343.

[9] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 37–49.

[10] Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8118–8128.

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics (ACL), 6894–6910.

[12] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 125–130.

[13] Alexander Gruenstein, John Niekrasz, and Matthew Purver. 2005. Meeting structure annotation: Data and tools. In *6th SIGdial Workshop on Discourse and Dialogue*.

[14] Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022. SPACE-2: Tree-Structured Semi-Supervised Contrastive Pre-training for Task-Oriented Dialog Understanding. *arXiv preprint arXiv:2209.06638* (2022).

[15] Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022. Space-3: Unified dialog model pre-training for task-oriented dialog understanding and generation. *arXiv preprint arXiv:2209.06664* (2022).

[16] Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10749–10757.

[17] Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23, 1 (1997), 33–64.

[18] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. *arXiv preprint arXiv:2211.11256* (2022).

[19] Hakan Inan, Rashi Rungta, and Yashar Mehdad. 2022. Structured Summarization: Unified Text Segmentation and Segment Labeling as a Generation Task. *arXiv preprint arXiv:2209.13759* (2022).

[20] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 986–995.

[21] Ting-En Lin, Yuchuan Wu, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Duplex Conversation: Towards Human-like Interaction in Spoken Dialogue Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3299–3308.

[22] Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8360–8367.

[23] Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1957–1965.

[24] Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li, and Fei Huang. 2022. Dial2vec: Self-Guided Contrastive Learning of Unsupervised Dialogue Embeddings. *arXiv preprint arXiv:2210.15332* (2022).

[25] Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray Buntine. 2021. Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence. *arXiv preprint arXiv:2110.07160* (2021).

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

[27] 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*. 1121–1130.

[28] Yushan Qian, Bo Wang, Ting-En Lin, Yinhe Zheng, Ying Zhu, Dongming Zhao, Yuexian Hou, Yuchuan Wu, and Yongbin Li. 2023. Empathetic Response Generation via Emotion Cause Transition Graph. *arXiv preprint arXiv:2302.11787* (2023).

[29] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. 3982–3992.

[30] Martin Riedl and Chris Biemann. 2012. TopicTiling: a text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 student research workshop*. 37–42.

[31] Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised Topic Segmentation of Meetings with BERT Embeddings. *arXiv e-prints* (2021), arXiv–2106.

[32] Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang. 2016. Dialogue Session Segmentation by Embedding-Enhanced TextTiling. *Interspeech 2016* (2016), 2706–2710.

[33] 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 *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2126–2131.

[34] 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*. 167–177.

[35] Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Discovering dialog structure graph for coherent dialog generation. 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)*. 1726–1739.

[36] Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14176–14184.

[37] 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, Volume 1 (Long Papers)*. 1575–1584.

[38] Hainan Zhang, Yanyan Lan, Liang Pang, Hongshen Chen, Zhuoye Ding, and Dawei Yin. 2021. Modeling topical relevance for multi-turn dialogue generation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 3737–3743.

[39] Sai Zhang, Yuwei Hu, Yuchuan Wu, Jiaman Wu, Yongbin Li, Jian Sun, Caixia Yuan, and Xiaojie Wang. 2022. A slot is not built in one utterance: Spoken language dialogs with sub-slots. *arXiv preprint arXiv:2203.10759* (2022).

[40] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11765–11773.

[41] Yicheng Zou, Jun Lin, Lujun Zhao, Yangyang Kang, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Unsupervised summarization for chat logs with topic-oriented ranking and context-aware auto-encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14674–14682.