Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring

Linzi Xing and Giuseppe Carenini
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada, V6T 1Z4
{lzxing, carenini}@cs.ubc.ca

Abstract

Dialogue topic segmentation is critical in several dialogue modeling problems. However, popular unsupervised approaches only exploit surface features in assessing topical coherence among utterances. In this work, we address this limitation by leveraging supervisory signals from the utterance-pair coherence scoring task. First, we present a simple yet effective strategy to generate a training corpus for utterance-pair coherence scoring. Then, we train a BERT-based neural utterance-pair coherence model with the obtained training corpus. Finally, such model is used to measure the topical relevance between utterances, acting as the basis of the segmentation inference[1]. Experiments on three public datasets in English and Chinese demonstrate that our proposal outperforms the state-of-the-art baselines.

1 Introduction

Dialogue Topic Segmentation (DTS), as a fundamental task of dialogue modeling, has received considerable attention in recent years. In essence, DTS aims to reveal the topic structure of a dialogue by segmenting the dialogue session into its topically coherent pieces. An example is given in Table 1. Topic transition happens after Turn-4 and Turn-6, where the topic is correspondingly switched from "the requirement of the insurance coverage" to "the information presented on the insurance card", and then to "the way of submitting the insurance card". Dialogue topic segmentation plays a vital role for a variety of downstream dialogue-related NLP tasks, such as dialogue generation (Li et al., 2016), summarization (Bokaei et al., 2016) and response prediction (Xu et al., 2021).

Different from the monologue topic segmentation (MTS) task (Koshorek et al., 2018; Xing et al., 2020), the shortage of labeled dialogue corpora has always been a very serious problem for DTS. Collecting annotations about topic shifting between the utterances of dialogues is highly expensive and time-consuming. Hence, most of the proposed labeled datasets for DTS are typically used for model evaluation rather than training. They are either small in size (Xu et al., 2021) or artificially generated and possibly noisy (Feng et al., 2020). Because of the lack of training data, most previously proposed methods for DTS follow the unsupervised paradigm. The common assumption behind these unsupervised methods is that the utterances associated with the same topic should be more coherent together than the utterances about different topics (Hearst, 1997; Purver et al., 2006). Hence, effectively modeling the coherence among utterances becomes the key ingredient of a successful DTS model. However, the performances of the prior unsupervised DTS models are usually limited since the coherence measurements between utterances

Turns	Dialogue Text
Turn-1: A: For how long should the liability insurance coverage remain in effect? **Turn-2**: B: As long as the registration of your vehicle remains valid. **Turn-3**: A: Does this apply for motorcycles too? **Turn-4**: B: There are some exceptions for motorcycles.	
Turn-5: A: Regarding the name on my vehicle registration application and the one on the Insurance Identification Card, do they need to be the same? **Turn-6**: B: yes, the names must match in both documents.	
Turn-7: A: Can I submit copies or faxes of my Insurance identification card to the DMV? **Turn-8**: B: yes, you can. But take into consideration that the card will be rejected if the DMV barcode reader can not scan the barcode.	

Table 1: A dialogue topic segmentation example sampled from *Doc2Dial* (Feng et al., 2020). This dialogue is segmented into three topical-coherent units (utterances in the same color are about the same topic).

[1]Our code, proposed fine-tuned models and data can be found at https://github.com/lxing532/Dialogue-Topic-Segmenter.

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 167–177
July 29–31, 2021.

are typically based on surface features (eg,. lexical overlap) (Hearst, 1997; Eisenstein and Barzilay, 2008) or word-level semantics (Song et al., 2016; Xu et al., 2021). Even though these features are easy to extract and thus make models more generally applicable, they can only reflect the coherence between utterances in a rather shallow way. More recently, there is work departing from the unsupervised setting by casting DTS as a weakly supervised learning task and utilizing a RL-based neural model as the basic framework (Takanobu et al., 2018). However, while this approach has been at least partially successful on goal-oriented dialogues when provided with predefined in-domain topics, it cannot deal effectively with more general open-domain dialogues.

To alleviate the aforementioned limitations in previous work, in this paper, we still cast DTS as an unsupervised learning task to make it applicable to dialogues from diverse domains and resources. However, instead of merely utilizing shallow features for coherence prediction, we leverage the supervised information from the text-pair coherence scoring task (i.e., measuring the coherence of adjacent textual units (Wang et al., 2017; Xu et al., 2019; Wang et al., 2020)), which can more effectively capture the deeper semantic (topical) relations between them. Due to the absence of supervision, we propose a simple yet effective strategy to generate a training corpus for the utterance-pair coherence scoring task, with the paired coherent/not-utterance pairs as datapoints. Then, after applying such strategy, we use the resulting corpus to train an utterance-pair coherence scoring model with the relative ranking objective (Li, 2011).

In practice, we create a training corpus from large conversational datasets containing real daily communications and covering various topics (proposed in Li et al. (2017) and Wang et al. (2021)). In particular, all the adjacent utterance pairs are firstly extracted to form the positive sample set. Then for each positive sample, the corresponding negative samples are generated by replacing the subsequent turn in the positive sample with (1) an non-adjacent turn randomly picked from the same dialogue, and (2) a turn randomly picked from another dialogue talking about another topic. Once the training corpus is ready, we re-purpose the *Next Sentence Prediction (NSP)* BERT model (Devlin et al., 2019) as the basic framework of our utterace-pair coherence scoring model. After fine-tuning the pretrained NSP BERT on our automatically generated training corpus with the marginal ranking loss, the resulting model can then be applied to produce the topical coherence score for all the consecutive utterance pairs in any given dialogue. Such scores can finally be used for the inference of topic segmentation for that dialogue.

We empirically test the popular *TextTiling* algorithm (Hearst, 1997) enhanced by the supervisory signal provided by our learned utterance-pair coherence scoring model on two languages (English and Chinese). The experimental results show that *TextTiling* enhanced by our proposal outperforms the state-of-the-art (SOTA) unsupervised dialogue topic segmenters by a substaintial margin on the testing sets of both languages. Finally, in a qualitative analysis, by visualizing the segment predictions of the different DTS segmenters on a sample dialogue, we show that the effectiveness of our proposal seems to come from better capturing topical relations and consideration for dialogue flows.

2 Related Work

Dialogue Topic Segmentation (DTS) Similar to the topic segmentation for monologue, dialogue topic segmentation aims to segment a dialogue session into the topical-coherent units. Therefore, a wide variety of approaches which were originally proposed for monologue topic segmentation, have also been widely applied to conversational corpora. Early approaches, due to lack of training data, are usually unsupervised and exploit the word co-occurrence statistics (Hearst, 1997; Galley et al., 2003; Eisenstein and Barzilay, 2008) or sentences' topical distribution (Riedl and Biemann, 2012; Du et al., 2013) to measure the sentence similarity between turns, so that topical or semantic changes can be detected. More recently, with the availability of large-scale corpora sampled from *Wikipedia*, by taking the section mark as the ground-truth segment boundary (Koshorek et al., 2018; Arnold et al., 2019), there has been a rapid growth in supervised approaches for monologue topic segmentation, especially neural-based approaches (Koshorek et al., 2018; Badjatiya et al., 2018; Arnold et al., 2019). These supervised solutions are favored by researchers due to their more robust performance and efficiency.

However, compared with monologue documents, dialogues are generally more fragmented and contain many more informal expressions. The dis-

course relation between utterances are also rather different from the monologue text. These distinctive features may introduce undesirable noise and cause limited performance when the supervised approaches trained on *Wikipedia* is applied. Since the lack of training data still remains a problem for DTS, unsupervised methods, especially the ones extending *TextTiling* (Hearst, 1997), are still the mainstream options. For instance, Song et al. (2016) enhanced *TextTiling* with word embeddings, which better capture the underlying semantics than bag-of-words style features. Later, Xu et al. (2021) replaced word embeddings with BERT as the utterance encoder to produce the input for *TextTiling*, because pretrained language models like BERT better capture more utterance-level dependencies. Also, to avoid a too fragmented topic segmentation, they adjusted the *TextTiling* algorithm into a greedy manner, which however requires more hyper-parameters and greatly limits the model's transferability. In contrast, here we adopt the original *TextTiling* to minimize the need of hyperparameters and use coherence signals for utterances learned from real-world dialogues to make our proposal more suitable for conversational data.

Another line of research explores casting DTS as a topic tracking problem (Khan et al., 2015; Takanobu et al., 2018), with the predefined conversation topics as part of the supervisory signals. Even though they have achieved SOTA performance on the in-distribution data, their reliability on the out-of-distribution data is rather poor. In contrast, our proposal does not require any prior knowledge (i.e., predefined topics) as input, so it is more transferable to out-of-distribution data.

Coherence Scoring Early on Barzilay and Lapata (2005, 2008) observed that particular patterns of grammatical role transition for entities can reveal the coherence of monologue documents. Hence, they proposed the entity-grid approach by using entity role transitions mined from documents as the features for document coherence scoring. Later, Cervone and Riccardi (2020) explored the potential of the entity-grid approach on conversational data and further proved that it was also suitable for dialogues. However, one key limitation of the entity-grid model is that by excessively relying on the identification of entity tokens and their corresponding roles, its performance can be reduced by errors from other NLP pre-processing tasks, like coreference resolution, which can be very noisy.

In order to resolve this limitation, researchers have explored scoring a document coherence by measuring and aggregating the coherence of its adjacent text pairs (e.g., Xu et al. (2019)), with Wang et al. (2017) being the first work demonstrating the strong relation between text-pair coherence scoring and monologue topic segmentation. In particular, they argued that a pair of texts from the same segment should be ranked more coherent than a pair of texts randomly picked from different paragraphs. With this assumption, they proposed a CNN-based model to predict text-pair semantic coherence, and further use this model to directly conduct topic segmentation. In this paper, we investigate how their proposal can be effectively extended to dialogues. Furthermore, we propose a novel method for data generation and model training, so that DTS and coherence scoring can mutually benefit each other.

3 Methodology

Following most of the previous work, we adopt *TextTiling* (Hearst, 1997) as the basic algorithm for DTS to predict segment boundaries for dialogues ((b) in Figure 1). Formally, given a dialogue d in the form of a sequence of utterances $\{u_1, u_2, ..., u_k\}$, there are $k-1$ consecutive utterance pairs. Then an utterance-pair coherence scoring model is applied to all these pairs and finally get a sequence of coherence scores $\{c_1, c_2, ..., c_{k-1}\}$, where $c_i \in [0, 1]$ indicates how topically related two utterances in the ith pair are. Instead of directly using the coherence scores to infer segment boundaries, a sequence of "depth scores" $\{dp_1, dp_2, ..., dp_{k-1}\}$ is calculated to measure how sharp a valley is by looking at the highest coherence scores $hl(i)$ and $hr(i)$ on the left and right of interval i: $dp_i = \frac{hl(i)+hr(i)-2c_i}{2}$. Higher depth score means the pair of utterances are less topically related to each other. The threshold τ to identify segment boundaries is computed from the mean μ and standard deviation σ of depth scores: $\tau = \mu - \frac{\sigma}{2}$. A pair of utterances with the depth score over τ will be select to have a segment boundary in between.

Next, we describe our novel training data generation strategy and the architecture of our new utterance-pair coherence scoring model, which are the two key contributions of this paper.

3.1 Training Data for Coherence Scoring

We follow previous work (Wang et al., 2017; Xu et al., 2019; Huang et al., 2020) to optimize the

utterance-pair coherence scoring model (described in Section 3.2) with marginal ranking loss. Formally, the coherence scoring model ***CS*** receives two utterances (u_1, u_2) as input and return the coherence score $c = \boldsymbol{CS}(u_1, u_2)$, which reflects the topical relevance of this pair of utterances. Due to the lack of corpora labeled with ground-truth coherence scores, we follow the strategy in Wang et al. (2017) to train ***CS*** based on the pairwise ranking with ordering relations of coherence between utterance pairs as supervisory signals.

In order to create the training data labeled with coherence ordering relations, we make two assumptions: (1) A pair of adjacent utterances is more likely to be more topical coherent than a pair of non-adjacent utterances but still in the same dialogue session. (2) A pair of utterances from the same dialogue is more likely to be more topical coherent than a pair of utterances sampled from different dialogues. To formalize the ordering relations, we notate a source dialogue corpus as $\mathcal{C}$ and use u_i^k to represent the ith utterance in the dialogue $d_k \in \mathcal{C}$. Then the two ordering relations based on the above assumptions can be formulated as:

$$CS(u_i^k, u_{i+1}^k) > CS(u_i^k, u_j^k), \quad j \notin \{i-1, i, i+1\} \tag{1}$$

$$CS(u_i^k, u_j^k) > CS(u_i^k, u_j^m), \quad k \neq m \tag{2}$$

Since the ranking objective is pairwise, given two utterance pairs, we deem the pair with higher/lower coherence score as the positive/negative instance. Taking eq. 1 as an example, (u_i^k, u_{i+1}^k) and (u_i^k, u_j^k) are positive and negative instance respectively.

Since the generality of the obtained coherence scoring model will significantly impact the robustness of the overall segmentation system, having a proper source dialogue corpus $\mathcal{C}$ to generate training data from is a critical step. We believe that an ideal source corpus should satisfy the following key requirements: (1) having a fairly large size; (2) covering as many topics as possible; (3) containing both formal and informal expressions. To test the strength of our proposal in a multilingual setting, we select *DailyDialog*[2] (Li et al., 2017) and *NaturalConv*[3] (Wang et al., 2021) for English and Chinese respectively. These two conversational corpora both consist of open-domain conversations about daily topics. Table 2 gives some statistics about them. Different from task-oriented dialogues, open-domain dialogues usually contain more diverse topics and expressions. From Table 2, we can see that both corpora cover multiple topics[4] and some topics like `Politics`, `Finance` and `Tech` are supposed to have more technical language, while others like `Sports`, `Entertainment` and `Ordinary Life` should include more casual expressions. Due to the lack of space, next we will only use *DailyDialog* as our running example source dialogue corpus $\mathcal{C}$ to illustrate the training data generation process for coherence scoring.

Dataset	DailyDialog	NaturalConv
Total dialogues	13,118	19,919
Language	English	Chinese
Avg. # turns per dialog	7.9	20.1
Avg. # tokens per turn	14.6	12.2
# covered topics	10	6

Table 2: Statistics of the two conversational corpora used for coherence scoring training data generation.

Given the source corpus *DailyDialog*, we first collect positive instances by extracting the adjacent utterance pairs which meet the Bi-turn Dialog Flow described in Li et al. (2017). The utterances in this corpus are labeled with the dialogue acts including {`Questions`, `Inform`, `Directives`, `Commissives`}. Among all the possible combinations, `Questions-Inform` and `Directives-Commissives` are deemed as basic dialogue act flows which happen regularly during conversations. Once positive instances $\mathcal{P} = \{(s_i, t_i^+) | i \in N\}$ have been collected, we adopt negative sampling to construct the negative instance for each positive instance by randomly picking:

— $\mathbf{t_i^-}$: an utterance not adjacent to s_i but in the same dialogue.
— $\mathbf{t_i^{'-}}$: an utterance from another dialogue different from s_i.

These utterances will replace t_i^+ in the positive instance to form two negative instances: (s_i, t_i^-) and $(s_i, t_i^{'-})$, where $\boldsymbol{CS}(s_i, t_i^+) > \boldsymbol{CS}(s_i, t_i^-) > \boldsymbol{CS}(s_i, t_i^{'-})$. In order to further enlarge the margins of coherence relations presented above, we set two constraints. Firstly, t_i^- should be labeled with

[2] yanran.li/dailydialog
[3] ai.tencent.com/ailab/nlp/dialogue/

[4] We omit topic categories of these two corpus for space, please refer original papers for more details.

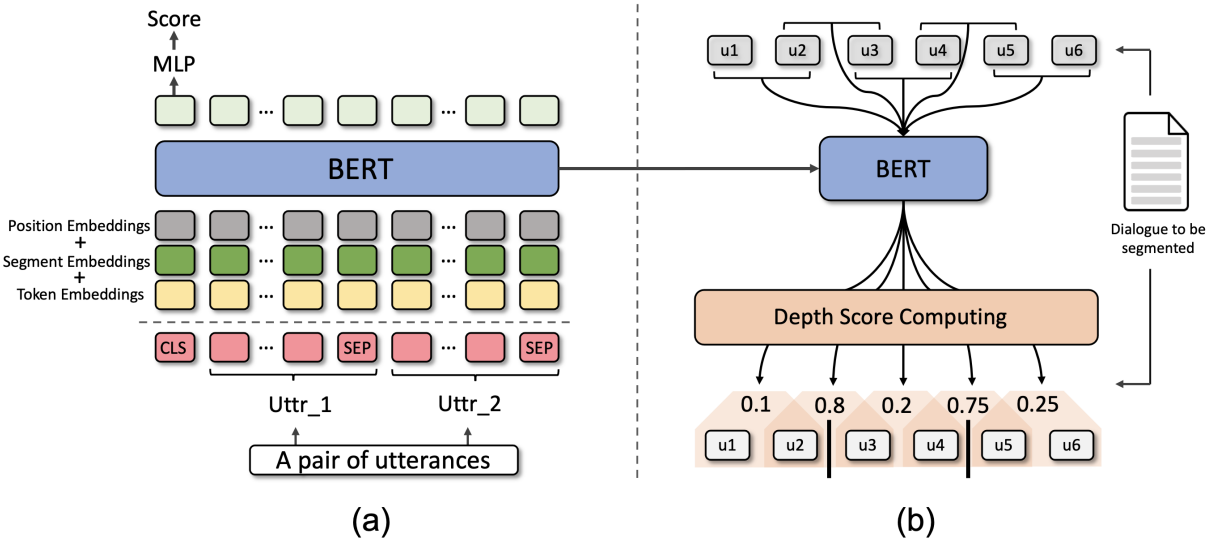

Figure 1: The overview of our proposed dialogue topic segmentation procedure. (a) Fine-tuning the NSP BERT on the training data of utterance-pair coherence scoring generated from the source dialogue corpus $\mathcal{C}$. (2) Leveraging the fine-tuned BERT as the coherence scoring model to predict coherence scores for all the consecutive utterance pairs in a testing dialogue. *TextTiling* algorithm is further utilized to infer segment boundaries.

the dialogue act different from t_i^+. Secondly, $t_i^{'-}$ should be sampled from a dialogue about a topic different from the dialogue which t_i^+ belongs to. Notice that the second corpus *NaturalConv* does not have dialogue act labels, so all the instance generation strategies with dialog acts in need are not applicable. In particular, positive instances for *NaturalConv* are simply adjacent utterances and the additional constraint for creating negative instances, in which t_i^- should be labeled with the dialogue act different from t_i^+, cannot be applied as well. By applying our novel data generation process, we obtain 91,581 and 599,148 paired pos/neg samples for *DailyDialog* and *NaturalConv* respectively. We split them into training (80%), validation (10%) and testing sets (10%) for further model training and evaluation.

3.2 Utterance-Pair Coherence Scoring Model

As illustrated in Figure 1(a), we choose the *Next Sentence Prediction (NSP)* BERT (Devlin et al., 2019) (trained for the *Next Sentence Prediction* task) as the basic framework of our utterance-pair coherence scoring model due to the similarity of these two tasks[5]. They both take a pair of sentences/utterances as input and only a topically related sentence should be predicted as the appropriate next sentence. We first initialize the model with BERT_{base}, which was pretrained on multi-billion publicly available data. At the fine-tuning stage, we expect the model to learn to discriminate the positive utterance pairs from their corresponding negative pairs. More specifically, the positive (s_i, t_i^+) and negative (s_i, t_i^-) as instances are fed into the model respectively in the form of $(\texttt{[CLS]}||s_i||\texttt{[SEP]}||t_i^{+/-}||\texttt{[SEP]})$, where $||$ denotes the concatenation operation for sequences and `[CLS]`, `[SEP]` are both special tokens in BERT. Following the original NSP BERT training procedure, we also add position embeddings, segment embeddings and token embeddings of tokens all together to get the comprehensive input for BERT. The NSP BERT is formed by a sequence of transformer encoder layers, where each layer consists of a self-attentive layer and a skip connection layer. Here we use the contextualized representation of `[CLS]` as the topic-aware embedding to predict how much the two input utterances are matched in topic. The topical coherence score will be estimated by passing `[CLS]` representation through another multilayer perceptron (MLP).

To encourage the model to learn to assign a positive instance (s_i, t_i^+) a coherence score c_i^+ higher than its paired negative instance (s_i, t_i^-) score c_i^-, we minimize the following marginal ranking loss:

[5]Instead of NSP BERT (a cross-encoder), we could have also modelled such pairwise scoring with a bi-encoder, which first encodes each utterance independently. We eventually selected the cross-encoder due to the results in Thakur et al. (2021) showing that cross-encoders usually outperform bi-encoders for pairwise sentence scoring.

$$L = \frac{1}{N}\sum_{i=1}^{N} max(0, \eta + c_i^- - c_i^+) \quad (3)$$

where N is the size of the training set, η is the margin hyper-parameter tuned at validation set.

4 Experiments

We comprehensively test our proposal by empirically comparing it with multiple baselines on three datasets in two languages.

4.1 Data for Evaluation

DialSeg_711 (Xu et al., 2021): a real-world dataset consisting of 711 English dialogues sampled from two task-oriented multi-turn dialogue corpora: MultiWOZ (Budzianowski et al., 2018) and Stanford Dialog Dataset (Eric et al., 2017). Topic segments of this dataset are from manual annotation.
Doc2Dial (Feng et al., 2020): This dataset consists of 4,130 synthetic English dialogues between a user and an assistant from the goal-oriented document-grounded dialogue corpus Doc2Dial. This dataset is generated by first constructing the dialogue flow automatically based on the content elements sampled from text sections of the grounding document. Then crowd workers create the utterance sequence based on the obtained artificial dialogue flow. Topic segments of this dataset are extracted based on text sections of the grounding document where the utterances' information comes from.
ZYS (Xu et al., 2021): is a real-world Chinese dataset consisting of 505 conversations recorded during customer service phone calls on banking consultation. Similar to DialSeg_711, gold topic segments of this dataset are manually annotated.

More details of the three datasets are in Table 3.

Dataset	DialSeg_711	Doc2Dial	ZYS
documents	711	4,130	505
language	English	English	Chinses
# sent/seg	5.6	3.5	6.4
# seg/doc	4.9	3.7	4.0
real-world	✔	✘	✔

Table 3: Statistics of the three dialogue topic segmentation testing sets for model evaluation.

4.2 Baselines

We compare our dialogue topic segmenter with following unsupervised baselines:
Random: Given a dialogue with k utterances, we first randomly sample the number of segment boundaries $b \in \{0, ..., k-1\}$ for this dialogue. Then we determine if an utterance is the end of a segment with the probability $\frac{b}{k}$.
BayesSeg (Eisenstein and Barzilay, 2008): This method models the words in each topic segment as draws from a multinomial language model associated with the segment. Maximizing the observation likelihood of the dialogue yields a lexically-cohesive segmentation.
GraphSeg (Glavaš et al., 2016): This method generates a semantic relatedness graph with utterances as nodes. Segments are then predicted by finding the maximal cliques of the graph.
GreedySeg (Xu et al., 2021): This method greedily determines segment boundaries based on the similarity of adjacent utterances computed from the output of the pretrained BERT sentence encoder.
TextTiling (TeT) (Hearst, 1997): The detailed description of this method can be found in Section 3.
TeT + Embedding (Song et al., 2016): TextTiling enhanced by GloVe word embeddings, by applying word embeddings to compute the semantic coherence for consecutive utterance pairs.
TeT + CLS (Xu et al., 2021): TextTiling enhanced by the pretrained BERT sentence encoder, by using output embeddings of BERT encoder to compute semantic similarity for consecutive utterance pairs.
TeT + NSP: TextTiling enhanced by the pretrained BERT for Next Sentence Prediction (NSP), by leveraging the output probability to represent the semantic coherence for consecutive utterance pairs.

4.3 Evaluation Metrics

We apply three standard metrics to evaluate the performances of our proposal and baselines. They are: P_k error score (Beeferman et al., 1999), *WinDiff (WD)* (Pevzner and Hearst, 2002) and F_1 score (macro). P_k and *WD* are both calculated based on the overlap between ground-truth segments and model's predictions within a certain size sliding window. Since they are both penalty metrics, lower score indicates better performance. F_1 is the standard armonic mean of precision and recall, with higher scores indicating better performance

4.4 Experimental Setup

We fine-tune the utterance-pair coherence scoring model on BERT$_{base}$ which consists of 12 layers and 12 heads in each layer. The hidden dimension of BERT$_{base}$ is 768. Training is executed with *AdamW* (Loshchilov and Hutter, 2019) as our optimizer and

Method	DialSeg_711			Doc2Dial		
	$P_k \downarrow$	$WD \downarrow$	$F_1 \uparrow$	$P_k \downarrow$	$WD \downarrow$	$F_1 \uparrow$
Random	52.92	70.04	0.410	55.60	65.29	0.420
BayesSeg (Eisenstein and Barzilay, 2008)	30.97	35.60	0.517	46.65	62.13	0.433
GraphSeg (Glavaš et al., 2016)	43.74	44.76	0.537	51.54	51.59	0.403
GreedySeg (Xu et al., 2021)	50.95	53.85	0.401	50.66	51.56	0.406
TextTiling (TeT) (Hearst, 1997)	40.44	44.63	0.608	52.02	57.42	0.539
TeT + Embedding (Song et al., 2016)	39.37	41.27	0.637	53.72	55.73	0.602
TeT + CLS (Xu et al., 2021)	40.49	43.14	0.610	54.34	57.92	0.518
TeT + NSP	46.84	48.50	0.512	50.79	54.86	0.550
Ours (w/o Dialog Flows)	32.60	37.97	0.750	48.76	50.83	0.636
Ours (w/o Dialog Topics)	26.95	28.98	0.761	46.61	48.58	0.657
Ours (full)	**26.80**	**28.24**	**0.776**	**45.23**	**47.32**	**0.660**

Table 4: The experimental results on two English testing sets: *DialSeg_711* (Xu et al., 2021) and *Doc2Dial* (Feng et al., 2020). ↑/↓ after the name of metrics indicates if the higher/lower value means better performance. The best performances among the listed methods are in **bold**.

Method	$P_k \downarrow$	$WD \downarrow$	$F_1 \uparrow$
Random	52.79	67.73	0.398
GreedySeg	44.12	48.29	0.502
TextTiling	45.86	49.31	0.485
TeT + Embedding	43.85	45.13	0.510
TeT + CLS	43.01	43.60	0.502
TeT + NSP	42.59	43.95	0.500
Ours	**40.99**	**41.32**	**0.521**

Table 5: The experimental results on the Chinese testing set proposed in Xu et al. (2021). The best performances among the listed methods are in **bold**.

the scheduled learning rate with warm-up (initial learning rate lr= 2e-5). Model training is done for 10 epochs with the batch size 16. Model's performance is monitored over the validation set and finally the margin hyper-parameter η in eq. 3 is set to 1 from the set of candidates $\{0.1, 0.5, 1, 2, 5\}$.

4.5 Results and Analysis

Table 4 compares the results of baselines and our proposal on two English dialogue topic segmentation evaluation benchmarks. The chosen baselines are clustered into the top three sub-tables in Table 4: random baseline, unsupervised baselines not extended from *TextTiling* and unsupervised baselines extended from *TextTiling*. Overall, our proposal (full) is the clear winner for both testing sets in all metrics. Another observation is that the set of segmenters *TeT + X*, which were proved to be effective for monologue topic segmentation, cannot consistently outperform the basic *TextTiling* on conversational data. The reason may be that the coherence prediction components of such approaches all rely on signals learned from monologue text (eg., GloVe and pretrained BERT). Due to the grammatical and lexical difference, signals learned from monologues tend to introduce unnecessary noise and limit the effectiveness of unsupervised topic segmemters when applied to dialogues. In contrast, our coherence scoring model trained on the dataset of coherent/non-coherent utterance pairs automatically generated from dialogues performs better than all comparisons by a substantial margin. Overall, this validates that by effectively using the topical relations of utterances in dialogue corpora, the BERT for next sentence prediction is able to produce coherence scores reflecting to what extend the two input utterances are matched in topic.

To confirm the benefit of taking dialogue flows and topics into account, we also conduct an ablation study by removing either one of these two parts from the training data generation process for coherence scoring. As reported in the bottom sub-table of Table 4, sampling positive/negative utterance pairs (t_i^+/t_i^- in Section 3.1) without using dialogue flows causes substantial performance drop on both testing sets, while sampling the other negative utterance pair ($t_i^{'-}$ in Section 3.1) without taking dialogue topics into consideration seems to have a smaller impact on the trained model's performance. This observation shows that the dialogue flow is a more effective signal than the dialogue topic. One possible explanation is that there are some basic dialogue flows that are commonly followed and gen-

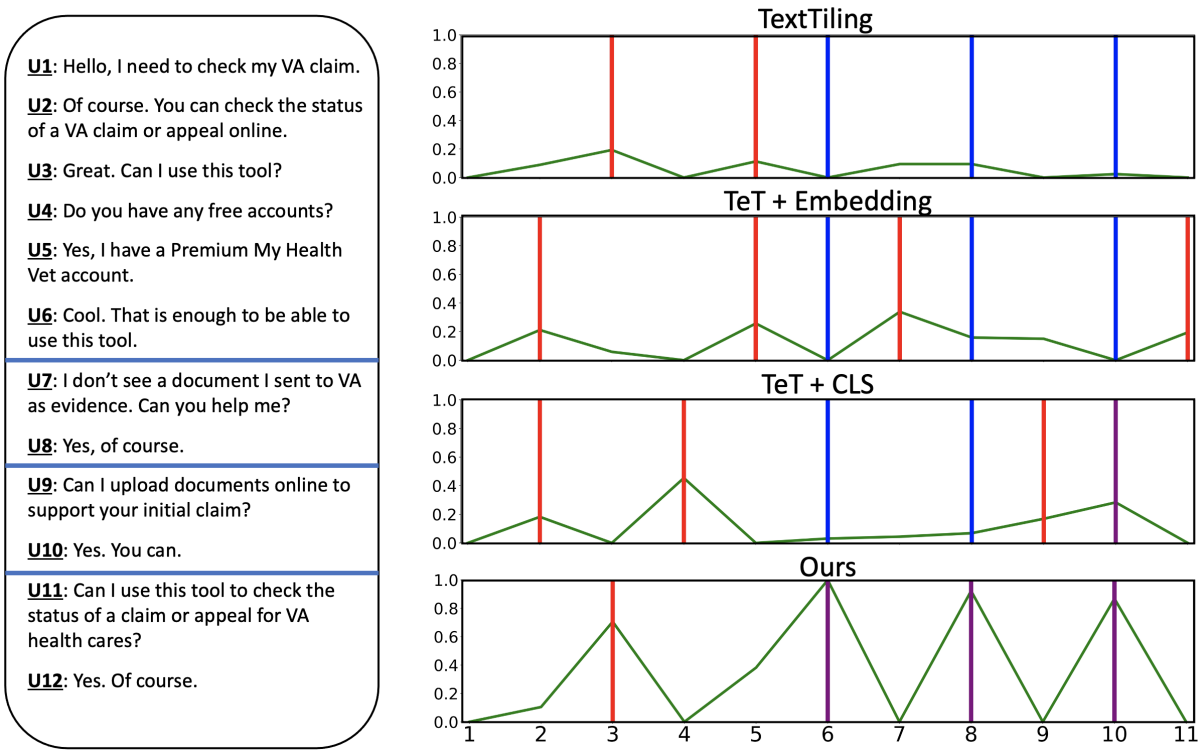

Figure 2: Behaviors of four TextTiling-based segmenters on an example dialogue selected from *Doc2Dial* (Feng et al., 2020). The horizonal axis is the index of intervals in a session, and the vertical axis is the value of depth score (higher value means more topical unrelated). The reference and prediction of topic boundaries are marked by blue and red vertical lines respectively. The overlaps of reference and prediction are marked by purple lines.

Method	DialSeg_711	Doc2Dial	ZYS
TextTiling	0.122	0.102	0.113
TeT + Embedding	0.136	0.125	0.131
TeT + CLS	0.166	0.154	0.158
Ours	**0.366**	**0.319**	**0.320**

Table 6: The average variance of depth scores on three testing sets. Highest values are in **bold**

eralize across different types of dialogues, while dialogue topics are more specific and vary much more between different dialogue corpora.

To further investigate the generality of our proposal for different languages, we train a Chinese coherence scoring model on the training data generated from *NaturalConv* (in Section 3.1) and use it together with *TextTiling* to infer segmentation for Chinese dialogues. Table 5 exhibits the performances of our method and baselines on the testing set ***ZYS***. Since the publicly available implementations for *BayesSeg* and *GraphSeg* only support English text as input, they are not included in this comparison. We note that although we observe a pattern similar to English, namely that our method surpasses all the selected baselines, gains seem to be smaller. While this still validates the reliability of our proposal for languages other than English, explaining this interlingual difference is left as future work. With a proper open-domain dialogue corpus for a particular language, *TextTiling* can be enhanced by the high-quality topical coherence signals in that language captured by our proposal.

4.6 Case Study

To more intuitively analyze the performance of our method and of the baselines, a sample dialogue is presented in Figure 2. First, notice that in models using more advanced features to compute coherence (line charts from top to bottom), the variation of depth scores (see §3) becomes more pronounced, which seem to indicate the more advanced models learn stronger signals to discriminate topically related and unrelated content. In particular, as shown again on the right-top of Figure 2, the plain *TextTiling*, which uses TF-IDF to estimate the coherence for utterance pairs, yields depth scores close to each other. With features carrying more complex semantic information, like word embeddings and BERT encoder pretrained on large-scale textual data, the difference of depth scores becomes more

obvious. Remarkably, our utterance-pair coherence scoring model optimized by marginal ranking loss further enlarges the difference. More tellingly, this trend holds in general for all three corpora as shown quantitatively in Table 6. We can observe that with more advanced features informing coherence computation, the variation of depth scores becomes more pronounced, which indicates that more advanced models can learn stronger signals to discriminate topically related and unrelated content. Remarkably, among all the presented methods, our proposal yields the largest average variance of depth scores across all three testing corpora.

A second key observation is about the benefit of our proposal taking dialogue flows into consideration in the training process. Consider (`U7`, `U8`) as an example, the first three segmenters tend to assign relatively high depth score (low coherence) to this utterance pair due to the very little content overlap between them. However, our method manages to assign this pair the minimal depth score. This is because such utterance pair is a `Questions-Inform` in the Dialog Flow, thus even if there is very limited content in common, the two utterances should still very likely belong to the same topic segment.

5 Conclusions and Future Work

This paper addresses a key limitation of unsupervised dialogue topic segmenters, namely their inability to model topical coherence among utterances in the dialogue. To this end, we leverage signals learned from a neural utterance-pair coherence scoring model based on fine-tuning NSP BERT. With no data labeled with gold coherence score, we also propose a simple yet effective way to automatically construct a training dataset from any source dialogue corpus. The experimental results on three testing sets in English and Chinese show that our proposal outperforms all the alternative unsupervised approaches.

For the future, although most recent work has built on *TextTiling*, we plan to explore if our proposal can also be integrated with other unsupervised topic segmentation methods, like *GraphSeg* and *BayesSeg*, rather than just *TextTiling*. Furthermore, we also plan to explore effective strategies to exploit external commonsense knowledge (eg., ConceptNet (Speer et al., 2017)) or user characters (Xing and Paul, 2017) in topic segmentation, since they have been shown to be beneficial in dialogue generation (Qiao et al., 2020; Ji et al., 2020b) and summarization (Ji et al., 2020a).

Acknowledgments

We thank the anonymous reviewers and the UBC-NLP group for their insightful comments and suggestions. This research was supported by the Language & Speech Innovation Lab of Cloud BU, Huawei Technologies Co., Ltd.

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. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-based neural text segmentation. In *Advances in Information Retrieval*, pages 180–193, Cham. Springer International Publishing.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210.

Mohammad Hadi Bokaei, Hossein Sameti, and Yang Liu. 2016. Extractive summarization of multi-party meetings through discourse segmentation. *Natural Language Engineering*, 22(1):41–72.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 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*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Alessandra Cervone and Giuseppe Riccardi. 2020. Is this dialogue coherent? learning from dialogue acts and entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174, 1st virtual meeting. Association for Computational Linguistics.

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

Lan Du, Wray Buntine, and Mark Johnson. 2013. Topic segmentation with a structured topic model. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Atlanta, Georgia. Association for Computational Linguistics.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii. Association for Computational Linguistics.

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*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

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)*, pages 8118–8128, Online. Association for Computational Linguistics.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.

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*, pages 125–130, Berlin, Germany. Association for Computational Linguistics.

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

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, and Minlie Huang. 2020a. Generating commonsense explanation by extracting bridge concepts from reasoning paths. 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*, pages 248–257, Suzhou, China. Association for Computational Linguistics.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020b. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736, Online. Association for Computational Linguistics.

O. Z. Khan, Jean-Philippe Robichaud, Paul A. Crook, and R. Sarikaya. 2015. Hypotheses ranking and state tracking for a multi-domain dialog system using multiple asr alternates. In *INTERSPEECH*, page 1810–1814.

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, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Hang Li. 2011. A short introduction to learning to rank. *IEICE Transactions on Information and Systems*, E94.D(10):1854–1862.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

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)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

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

Matthew Purver, Konrad P. Körding, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International*

Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 17–24, Sydney, Australia. Association for Computational Linguistics.

Lin Qiao, Jianhao Yan, Fandong Meng, Zhendong Yang, and Jie Zhou. 2020. A sentiment-controllable topic-to-essay generator with topic knowledge graph. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3336–3344, Online. Association for Computational Linguistics.

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, Jeju Island, Korea. Association for Computational Linguistics.

Yiping Song, Lili Mou, R. Yan, Li Yi, Zinan Zhu, X. Hu, and M. Zhang. 2016. Dialogue session segmentation by embedding-enhanced texttiling. In *INTERSPEECH*, page 2706–2710.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Fenglin Li, Haiqing Chen, Xiaoyan Zhu, and Liqiang Nie. 2018. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4403–4410. International Joint Conferences on Artificial Intelligence Organization.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Liang Wang, Sujian Li, Yajuan Lv, and Houfeng Wang. 2017. Learning to rank semantic coherence for topic segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1340–1344, Copenhagen, Denmark. Association for Computational Linguistics.

Weishi Wang, Steven C.H. Hoi, and Shafiq Joty. 2020. Response selection for multi-party conversations with dynamic topic tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6581–6591, Online. Association for Computational Linguistics.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14006–14014.

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*, pages 626–636, Suzhou, China. Association for Computational Linguistics.

Linzi Xing and Michael J. Paul. 2017. Incorporating metadata into content-based user embeddings. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 45–49, Copenhagen, Denmark. Association for Computational Linguistics.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Florence, Italy. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14176–14184.