Employing Coarse-grained Task to Improve Fine-grained Dialogue Topic Shift Detection

Jiangyi Lin[1], Yaxin Fan[1], Xiaomin Chu[1], Peifeng Li✉[1] and Qiaoming Zhu[1]

[1] School of Computer Science and Technology, Soochow University, Suzhou, China
{jylin, yxfansuda}@stu.suda.edu.cn
{xmchu, pfli, qmzhu}@suda.edu.cn

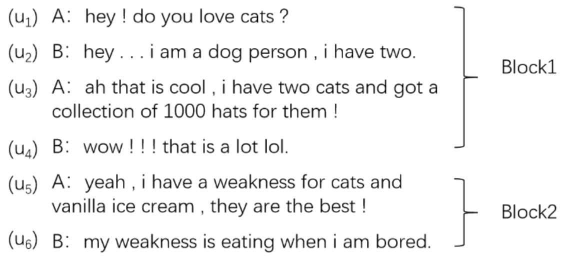
Abstract. The goal of dialogue topic shift detection is to identify whether the current topic in a dialogue has shifted or not. Previous work has focused on detecting whether a topic has shifted, without delving into the finer-grained topic situations of the dialogue. To address these issues, we further explore fine-grained topic shift detection. Based on different categories of topic semantics, a multi-task learning framework is constructed by treating the labels of both coarse and fine granularity as different tasks. The topic semantics of the two granularities reinforce each other and enhance the robustness of the model. Finally, semantic coherence learning as well as weight adaptation learning are applied to alleviate the sample imbalance problem in the dataset, so that the model can distinguish different topic shift situations more effectively. Experimental results on the Chinese dataset CNTD show that the proposed model outperforms several baseline models.

Keywords: Chinese dialogue topic, Fine granularity topic, Topic shift detection, Multi-task Learning.

1 Introduction

The topic structure explains the topical relationship between two consecutive text units (e.g., paragraphs in a discourse, turns in a dialogue). As one of the essential dialogue analysis tasks, dialogue topic shift detection refers to detecting whether a topic shift has occurred in the response of a dialogue, which can help dialogue systems to change topics and actively guide the dialogue. Based on whether the topic of the dialogue has shifted or not, some researchers further refine the topic situation to obtain fine-grained topic shift labels. The fine-grained topic shift detection task contains richer dialogue semantics while providing further assistance to the dialogue system. Since this task can help various models understand dialogue topics, it is of great benefit for many downstream tasks, such as response generation [1] and reading comprehension [2,3]. It can also assist real-time applications in generating topics that perform well in dialogue scenarios due to their response shift [4,5,6].

The goal of dialogue topic shift detection is to identify the topic of dialogue by taking into account the current response and context in real time. This task is similar to dialogue topic segmentation [7]. However, dialogue topic shift detection is a real-time task and cannot access future turns. For example, in Fig. 1, there is a part of a dialogue with two topics (e.g., favorite animal and weakness). The task of topic shift detection is to predict whether the next turn will change the topic, based on all existing turns. If we want to detect whether the topic is shifted during u_1 and u_2, we can only access two turns, i.e., u_1 and u_2.



(u_1) A: hey ! do you love cats ?

(u_2) B: hey . . . i am a dog person , i have two.

(u_3) A: ah that is cool , i have two cats and got a collection of 1000 hats for them !

(u_4) B: wow ! ! ! that is a lot lol.

(u_5) A: yeah , i have a weakness for cats and vanilla ice cream , they are the best !

(u_6) B: my weakness is eating when i am bored.

Fig. 1. An example of the topic structure in a dialogue of six turns (i.e., $u_1 - u_6$) where each block refers to a topic.

Dialogue topic shift detection is a relatively new task in the field of dialogue topics. Although those topic segmentation models can be adapted in topic shift detection, the absence of future dialogue makes it harder to distinguish the topic shift between turns. Furthermore, fine-grained dialogue topic shift detection is a more challenging task.

Only a few studies focused on dialogue topic shift detection. Xie et al. [8] defined the dialogue topic shift detection task in detail and annotated the TIAGE corpus in English. They then used T5 [9] to detect topic shifts. Xu et al. [10] built an English corpus including 711 conversations, as well as annotated a Chinese conversation topic corpus including 505 customer service call records. Lin et al. [11] annotated a Chinese Natural Dialogue Topics Corpus CNTD, which contains 1308 dialogues, totaling 26K dialogue turns. Lin et al. [11] constructed a model based on knowledge distillation to explore response-unknown dialogue topic shift detection. Lin et al.[12] proposed to facilitate response-known dialogue topic shift detection by mining dialogue information.

Multi-task learning [13] is a machine learning approach that aims to improve the performance of a model by learning multiple related tasks simultaneously. Compared to traditional single-task learning, multi-task learning allows models to learn more generalized and abstract semantic features. Based on the multi-task learning framework, we jointly train both coarse and fine granularity tasks so that the two tasks can promote each other. Relying on the guidance of the coarse-grained task and the richer semantic information in the fine-grained task, the model can better understand the topic information of the dialogue. Under this framework, the model further exploits the rich semantics of the fine-grained labels and employs semantic coherence learning to enhance the learning of the conversation. Meanwhile, the problem of imbalance of samples of different categories in the corpus is alleviated by semantic coherence learning. In addition, we use weight adaptation learning to allow the model to obtain robust results even when detecting rare topic shift situations. Due to the lack of fine-grained corpus, we only conducted experiments on the CNTD dataset. The results show that our proposed model outperforms the baseline model. Our contributions are as follows.

- We are the first to investigate the task of dialogue topic shift detection for fine-grained topic situations.
- We propose a method for topic shift detection by considering both coarse and fine granularity as different tasks allowing mutual reinforcement between topic semantics.
- We build a multi-task framework-based model to utilize different granularity topic semantics.
- Our model achieves SOTA performance on the Chinese corpus CNTD.

2 Background

We first briefly introduce the relevant dialogue topic corpus, then summarize the existing methods for dialogue topic detection, and finally introduce the related research on Mulit-tasks.

2.1 Dialogue Topic Corpora

For English, Xie et al. [8] annotated the TIAGE corpus consisting of 500 dialogues with 7861 turns based on PersonaChat [14]. Xu et al. [10] built a dataset including 711 dialogues by joining dialogues from existing multi-turn dialogue datasets: MultiWOZ Corpus [15], and Stanford dialogue Dataset [16]. Both corpora are either small or limited to a particular domain, and neither applies to the study of the natural dialogue domain.

For Chinese, Xu et al. [10] annotated a dataset including 505 phone records of customer service on banking consultation. However, this corpus is likewise restricted to a few specialized domains while natural dialogues are more complicated. Lin et al. [11] extract dialogues from the NaturalConv dataset, construct the Chinese natural conversation topic dataset CNTD, and annotate fine-grained conversation topic situations. Eventually, the annotated Chinese Natural Dialogue Topics Corpus CNTD contains 1308 conversations, totaling 26K dialogue turns.

2.2 Dialogue Topic Detection

The field of detecting topic shifts in dialogue is still in its infancy and has received limited attention thus far.

For the dialogue topic shift detection task, Xie et al. [8] are the first to define this task and predict the topic shift based on the T5 model. Additionally, Xie et al. [8] distinguish more fine-grained situations of topic shift, categorizing the topic shift as commenting on the previous context, question answering, and developing the conversation to sub-topics, and categorizing the topic not shift as introducing a relevant but different topic, and completely changing the topic, as shown in Fig. 2.

In general, the dialogue topic shift detection task is still a challenge, as it can only rely on the context information of the dialogue. Lin et al. [11] constructed a model based on knowledge distillation to explore response-unknown dialogue topic shift detection. They introduced privileged information through the teacher-student framework and constructed hierarchical contrastive learning to enhance the topic development paradigm.

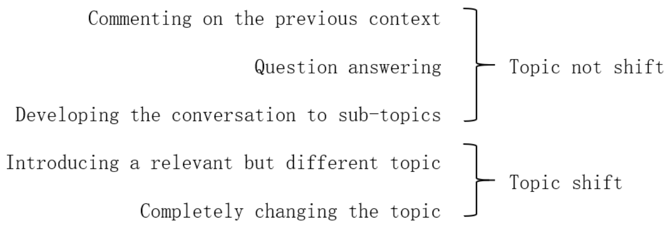

Commenting on the previous context
Question answering
Developing the conversation to sub-topics
Topic not shift
Introducing a relevant but different topic
Completely changing the topic
Topic shift

Fig. 2. The relationship between coarse and fine-grained topic situations.

Lin et al. [12] proposed to facilitate response-known dialogue topic shift detection by mining dialogue information. They utilize different dimensions of dialogue information by designing different target sentences and applying prompt learning to enable the model to learn the topic semantics.

2.3 Multi-task Learning

The multi-task learning framework [17] is a machine learning methodology designed to improve the performance of a model by simultaneously processing and learning multiple related tasks. This framework is designed so that the model can share and utilize information between tasks to achieve better results on each task. The core idea of this framework is to enable the model to learn shared features and knowledge between tasks by introducing multiple tasks in the same model. This helps to solve the problems of sample scarcity and data imbalance in single-task learning while improving the generalization ability of the model. In multi-task learning frameworks, the architecture of the model usually includes a shared layer and a task-specific layer. The shared layer is used to extract generalized features, while the task-specific layer is used to process information specific to each task.

3 Model

Our framework is shown in Fig. 3. To combine the instructive nature of coarse-grained labels with the richer semantic information in fine-grained labels, we introduce a multi-task learning framework, which consists of a shared backbone encoder as well as a multi-task classification module. The shared backbone encoder module uses different encoders to obtain dissimilar dialogue representations, while the multi-task classification module treats both coarse-grained and fine-grained dialogue topic shift detection as different tasks for joint training.

3.1 Multi-task Framework

Although fine-grained labels contain richer semantic information about topics, some topic situations also lead to detection difficulties because their semantics have similarities. To solve this problem, we construct a model based on a multi-task learning framework. The model is guided by added coarse-grained detection tasks to be able to further distinguish confusing topic situations.

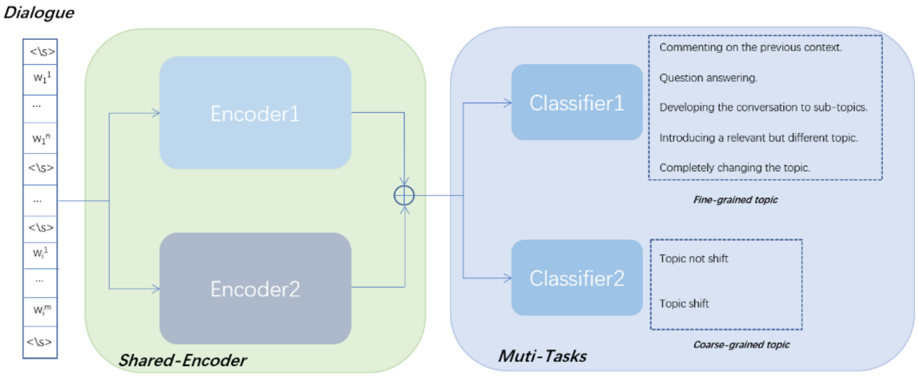


Fig. 3. Model architecture.

In this framework, multiple pre-trained models are employed as Shared-Encoder (SE) to obtain semantic representations of conversations. Specifically, let $\mathrm{C} = \{du_1, \ldots, du_i, \ldots, du_{n-1}\}$ represents a set of existing turns, where $\mathrm{n}-1$ is the number of the existing turns, and $\mathrm{d}u_i$ is the i-th turn. Let $\mathrm{R} = \{du_n\}$ represents a response turn after C. Finally, the set of all known turns includes C and R, which can be denoted as $\mathrm{DU} = \mathrm{C}, \mathrm{R}$.

We need to train a model $\mathrm{f}: \mathrm{DU} \rightarrow Y(R)$ to classify the response turn R (i.e., $\mathrm{d}u_n$) into the predefined categories $Y_c = \{0,1\}, Y_f = \{1,2,3,4,5\}$, where c stands for coarse-grained labels and f for fine-grained labels. Similar to the input of traditional classification models, we first convert DU into a string D as follows.

$$\mathrm{D} =<\backslash \mathrm{s} > \mathrm{d}u_1 <\backslash \mathrm{s} > \cdots <\backslash \mathrm{s} > \mathrm{d}u_n <\backslash \mathrm{s} > \tag{1}$$

where $\mathrm{d}u_1 = \mathrm{w}_1^1, \ldots, w_1^j$ and $\mathrm{d}u_n = \mathrm{w}_n^1, \ldots, w_n^k$ denote the sequence of tokens of $\mathrm{d}u_1$ and $\mathrm{d}u_n$, respectively. $<\backslash \mathrm{s} >$ denotes the respective special symbols of the different encoders.

Then, we feed D into the shared encoder to get the hidden state $Hidden_{1,2}$ of the particular $<\backslash \mathrm{s} >$ in the encoder, as follows.

$$Hidden_{1,2} = \mathrm{Encoder}_{1,2}(\mathrm{D}) \tag{2}$$

where $\mathrm{Hidden} \in \mathrm{R}^{\mathrm{N} \times \mathrm{d}}, \mathrm{i} \in \mathrm{I} = \{1,2, \ldots, \mathrm{N}\}$ denotes the index of the sample in the batch.

In the encoder, the model concatenates the hidden states ($Hidden_1$ and $Hidden_2$) of the $<\backslash \mathrm{s} >$ of the shared encoder to represent the corresponding topic situation representations, as shown in Equation 5.3. Then, the output V_s of the encoder is fed into the classification layer to determine the topic of the last turn $\mathrm{d}u_n$ as follows. where the classifier in multi-task learning consists of linear layers.

$$V_s = \mathrm{Con}(Hidden_1, Hidden_2) \tag{3}$$

$$\mathrm{Probabilit}y_{1,2} = \mathrm{Classifie}r_{1,2}(V_s) \tag{4}$$

3.2 Loss Function

In addition to combining the advantages of both granularities through a multi-task learning framework. We believe that the semantic information in fine-grained labels needs to be further explored. The proportion of different topic shift situations in the conversation corpus is not balanced. Therefore, we propose semantic coherence learning, as well as weight adaptation learning.

Semantic Coherence Learning We build semantic coherence learning only in the module for fine-grained tasks so that the shared backbone representation can tend to learn the semantic differences of different topic situations at fine-grained levels. As a result, semantic coherence learning is more effective in improving the final detection results. The following equation can be used to describe this loss. For a batch of N training samples, copy the last hidden state H of the conversation to get $\overline{H}$, treat it as positive, and separate its gradient. This gives $2N$ samples, and then the perceived loss of semantic coherence for all samples in a batch can be expressed as follows.

$$\mathrm{U} = \left[\mathrm{H}; \overline{\mathrm{H}}\right] \tag{5}$$

$$L^{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{exp(U_i \cdot U_p)}{\sum_{a \in A(i)} exp(U_i \cdot U_a)} \tag{6}$$

where $U \in \mathrm{R}^{2\mathrm{N} \times \mathrm{d}}$, $\mathrm{i} \in \mathrm{I} = \{1,2, \dots, 2\mathrm{N}\}$ denotes the index of the sample in the batch, and $\mathrm{P(i)} = \mathrm{I}_{\mathrm{j=i}} - \{\mathrm{i}\}$ denotes the sample that belongs to the same category as i but not to itself.

Weighted Adaptive Learning. Considering the case that the proportion difference of the number of different category samples is more disparate, the model learning process will be dominated by large-scale categories, which, because of their large number, have a large impact on model convergence and affect the classification effect of the model. Inspired by FOCAL LOSS, we add new coefficient factors to the standard cross-entropy loss to weaken the learning of large-scale category samples, strengthen the learning of rare samples, and improve the classification ability of the model. The specific calculation is shown in Equation 7, where p_t denotes the predicted probability of the true category of the sample, where γ is the hyperparameter that regulates the weight factor.

$$\mathrm{L}^{\mathrm{WAL}} = -\mathrm{y}\,(1 - \mathrm{p_t})^{\gamma} \log(\mathrm{p_t}) \tag{7}$$

As opposed to directly taking the introduced weights as fixed values, the weights are softened as a learnable weight that is calculated based on the prediction results. The introduced coefficient factor γ implements the adjustment of the degree of contribution of the sample to the loss according to the degree of difference between the predicted result and the true label, when the smaller the difference, it means the more accurate the prediction, and reduces the loss weight, and vice versa increases the loss weight. The larger the γ, the stronger the adjustment.

3.3 Loss Function

The final loss function of the model consists of three components: weight adaptation learning for fine-grained topic detection tasks (i.e., L_{Fine}^{WAL}), semantic coherence learning for fine-grained topic detection tasks (i.e., L_{Fine}^{SCL}), and weight adaptation learning for coarse-grained topic detection tasks (i.e., L_{Coarse}^{WAL}).

$$Loss = L_{Fine}^{WAL} + L_{Fine}^{SCL} + L_{Coarse}^{WAL} \tag{8}$$

4 Experimentation

In this section, we first introduce the experimental settings. Then, we report the experimental results and analysis.

4.1 Experimental Settings

Due to the limited availability of fine-grained corpus, the model is only experimented on CNTD. We followed Lin et al.'s division of the dataset CNTD and extracted (context, response) pairs from each conversation as inputs and extracted the labels of the responses as the target of the task. In the experiments, every sentence except the first sentence of the dialogue can be considered as a response. For evaluation, we report precision (P), recall (R) and Macro-F1 scores for fine-grained labels.

Our experiments use PyTorch and Huggingface [18] as deep learning frameworks. Our model uses BERT, and XLNet as an encoder and is fine-tuned during the training process. For the CNTD corpus, all pre-trained model parameters were set to default values. The models were experimented on an NVIDIA GeForce GTX 3090 with a CNTD batch size of 8 and an initial learning rate of 2e-5. The number of training epochs was set to 20 and the dropout was set to 0.5. In addition, a warm-up strategy and the AdamW optimizer were used in training and the decay factor was set to 0.01.

4.2 Experimental Results

In the dialogue topic shift detection task, the study by Xie et al. is the only one that uses the T5 model to establish a baseline on TIAGE, but it is also limited to the coarse-grained topic detection task. Therefore, we selected several pre-trained models as baseline models and also added hierarchical models. Ultimately, the following baseline was used for comparison:

- **BERT** [19], a Transformer-based bidirectional encoder for text encoding;
- **RoEBRTa** [20], an improvement on BERT;
- **XLNet** [21], an autoregressive pre-trained language model based on the Transformer architecture, which addresses limitations in autoregressive models by maximizing the joint probability of all possible permutations;
- **T5** [9], a modification of the Transformer architecture for use in various NLP tasks such as text-to-text tasks;
- **Hier-BERT** [22], a hierarchical structure based on the Transformer model;

- **BERT+BiLSTM** [22], a hierarchical structure based on the Transformer model, a combination of BERT for text encoding and Bi-directional LSTM for deep bi-directional language representation.

Table 1. Results of Baselines and Our Model on CNTD.

Model	P	R	F1
BERT	66.1	66.4	66.2
RoBERTa	62.9	63.5	63.1
XLNet	66.0	66.0	65.4
T5	65.7	65.6	65.6
BERT+BiLSTM	64.2	65.3	64.6
Hier-BERT	62.6	64.0	63.1
Ours	**67.7**	**69.7**	**68.5**

The results, as shown in Table 1, show that the pre-trained models do not perform consistently in the experiments with fine-grained topic detection results, with RoBERTa performing the worst and BERT performing the best, with an F1 score of 66.2%. However, the performance of Hier-BERT and BERT+BiLSTM with a layered structure did not improve compared to the single pre-trained model, with F1 scores However, the performance of Hier-BERT and BERT+BiLSTM with a hierarchical structure is not improved compared to the single pre-trained model, with F1 scores of 63.1% and 64.6%. We believe that in fine-grained topic detection, the use of LSTM leads to an increase in the complexity of the model, which is not worth the cost. At the same time, the internal gating mechanism of LSTM increases the complexity of the network, making the model less effective in detecting fine-grained topics.

Using a t-test with a confidence interval of 95% for significance, all improvements in Our Model relative to the baseline were significant ($p < 0.01$). In addition, our model improved its F1 score by 2.3%, which is significantly better than the optimal baseline, T5. This result validates the effectiveness of our proposed model.

5 Analysis

In this section, we first analyze our proposed model in different aspects to verify its effectiveness and then give the case study and error analysis.

5.1 Analysis of Different Categories of Samples

To analyze the optimization of the different models as well as our model, we have compiled the performance of the different models for different topic detections at a fine-grained level, as shown in Table 2, which specifically demonstrates the performance of the different models on Macro-F1.

Table 2. Model performance on different categories.

Fine-grained labels	SE(F1)	Our Model(F1)
Commenting on the previous context	87.1	88.0
Question answering	86.6	85.3
Developing the conversation to sub-topics	20.5	25.3
Introducing a relevant but different topic	50.0	52.0
Completely changing the topic	91.7	92.0

Considering the performance of the Marco-F1 values of the SE model and our model on different categories, it can be seen that on the categories of commenting on the previous context and completely changing the topic, the two models perform very similarly. Semantically, these two categories are easier to distinguish from the others. The semantics of completely changing the topic was the easiest to distinguish, with our model achieving a performance of 92.0% in this category. Although in the category of "question answering", our model was instead inferior to the SE model. However, our model's performance improves significantly in the key categories of "developing the conversation to sub-topics" and "introducing a relevant but different topic". In conclusion, our model has better performance in recognizing topic development and leading in conversations, demonstrating the model's better ability to adapt to conversations with diverse topics.

5.2 Case Study

We selected a 20-turn dialogue containing five topic situations for our case study. The detection results of the SE model and our model and the true labels are shown in Table 3. Since we do not consider the 1^{st} turn of a conversation to be a response, the result for the 1^{st} turn is defaulted to 0. The sample shown in the table has a total of four topic shifts.

As can be seen from Table 3, our model performs better than SE in detecting the occurrence of topic shift in the dialogue. our model can accurately detect topic shifts in the 3^{th}, 15^{th} and 19^{th} turn, while the SE model can only correctly detect topic shifts in the 3^{th} turn. The SE model incorrectly detects "introduced a relevant but different topic" as "question answering" and "completely changed the topic" as "introduced a relevant but different topic" in the 15^{th} and 19^{th} turn, respectively. as "introducing a relevant but different topic". Meanwhile, in the 9^{th} turn, "This affects one's metabolism, and it slows down the metabolism.", our model was able to effectively detect that the response was answering the previous question, whereas the SE model got the wrong result. The above results show that our model can effectively distinguish semantic differences between fine-grained topic situations.

5.3 Error Analysis

We focus on the turns in which the model showed an error. In the 5^{th} and 6^{th} turn, the semantic difference between the topical level key information in the text and the response is indeed very small, which is the main reason for the error. But in fact, the follow-up topic did start from "eating so little" in the \$$5^{th}$ turn. In the 7^{th} turn, the conversation topic was different from the previous turn, and the model did not effectively detect that the conversation had introduced a sub-topic of "intermittent fasting". This led to our model judgment in the category of "commenting on the previous context". However, both our model and the SE model mistakenly assume that this response did not shift the topic.

Table 3. Results of SE and our model on samples and true labels.

Turns	SE	Our Model	Label
1	0	0	0
2	1	1	1
3	5	5	5
4	2	2	2
5	1	1	4
6	4	4	2
7	4	1	3
8	1	1	1
9	3	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	4	4	1
14	4	1	1
15	1	4	4
16	1	1	1
17	1	1	1
18	1	1	1
19	4	5	5
20	1	1	1

As mentioned above, our model still has some difficulty in detecting the "introducing a relevant but different topic" topic shift in the table dialogue, which is a common problem in other dialogues. Evaluation of the model error cases showed that the discrepancies between these responses and the preceding text were often complex and ambiguous, making it difficult to obtain accurate results from the model.

6 Conclusion

In this paper, we design a multi-task learning-based framework to address the problems of complex conversation topic transfer situations, abstract semantic information, and imbalance of sample categories, as a way to achieve the goal of allowing the semantic information of fine-grained topic labels to be combined with the instructive nature of coarse-grained topic labels. On this basis, we introduce semantic coherence learning and weight adaptation learning to allow the model to further recognize semantic distinctions between different topic situations while strengthening the recognition of rare categories. The final experiments show that the final model proposed in this paper achieves the best performance on the CNTD corpus. However, there are still confusing topic situations. Meanwhile, there is still much room for improving the model's performance on categories with few samples.

Acknowledgments. The authors would like to thank the two anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62376181 and 62276177), and Key R&D Plan of Jiangsu Province (BE2021048).

References

1. Shuyang Dai, Guoyin Wang, Sunghyun Park, and Sungjin Lee. Dialogue response generation via contrastive latent representation learning. In Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pages 189–197, 2021.
2. Jiaqi Li, Ming Liu, Zihao Zheng, Heng Zhang, Bing Qin, Min-Yen Kan, and Ting Liu. Dadgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2021.
3. Yiyang Li and Hai Zhao. Self-and pseudo-self-supervised prediction of speaker and key-utterance for multi-party dialogue reading comprehension. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2053–2063, 2021.
4. Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. Approximating interactive human evaluation with self-play for open-domain dialog systems. Advances in Neural Information Processing Systems, 32, 2019.
5. Arash Einolghozati, Sonal Gupta, Mrinal Mohit, and Rushin Shah. Improving robustness of task-oriented dialog systems. arXiv preprint arXiv:1911.05153, 2019.
6. Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. Di alogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In Proceedings of NAACL-HLT, pages 2060–2069, 2018.
7. Linzi Xing and Giuseppe Carenini. Improving unsupervised dialogue topic segmen tation with utterance-pair coherence scoring. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 167–177, 2021.
8. Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. Tiage: A benchmark for topic-shift aware dialog modeling. Findings of the Asso ciation for Computational Linguistics: EMNLP 2021. 2021: 1684-1690., 2021.
9. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140):1–67, 2020.
10. Yi Xu, Hai Zhao, and Zhuosheng Zhang. Topic-aware multi-turn dialogue model ing. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14176–14184, 2021.

11. Jiangyi Lin, Yaxin Fan, Feng Jiang, Xiaomin Chu, and Peifeng Li. Topic shift de tection in chinese dialogues: Corpus and benchmark. In Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi, editors, Document Analysis and Recognition ICDAR 2023, pages 166–183, Cham, 2023. Springer Nature Switzerland.
12. Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. Multi granularity prompts for topic shift detection in dialogue. In De-Shuang Huang, Prashan Premaratne, Baohua Jin, Boyang Qu, Kang-Hyun Jo, and Abir Hussain, editors, Advanced Intelligent Computing Technology and Applications, pages 511-522, Singapore, 2023. Springer Nature Singapore.
13. Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Mod eling task relationships in multi-task learning with multi-gate mixture-of-experts. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pages 1930–1939, 2018.
14. Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2204–2213, 2018.
15. Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 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, 2018.
16. Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. Key-value retrieval networks for task-oriented dialogue. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pages 37–49, 2017.
17. Rich Caruana. Multitask learning. Machine learning, 28:41–75, 1997.
18. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement De langue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pages 38–45, 2020.
19. Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Pro ceedings of NAACL-HLT, pages 4171–4186, 2019.
20. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
21. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32, 2019.
22. Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. Text segmentation by cross segment attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4707–4716, 2020.