TIAGE: A Benchmark for Topic-Shift Aware Dialog Modeling

Huiyuan Xie\textsuperscript{1} Zhenghao Liu\textsuperscript{2} Chenyan Xiong\textsuperscript{3} Zhiyuan Liu\textsuperscript{4} Ann Copestake\textsuperscript{1}
\textsuperscript{1}Department of Computer Science and Technology, University of Cambridge, UK
\textsuperscript{2}Department of Computer Science and Technology, Northeastern University, China
\textsuperscript{3}Microsoft Research, United States
\textsuperscript{4}Department of Computer Science and Technology, Tsinghua University, China
{hx255,aac10}@cl.cam.ac.uk, liuzhenghao@cse.neu.edu.cn
chenyan.xiong@microsoft.com, liuzy@tsinghua.edu.cn

Abstract

Human conversations naturally evolve around different topics and fluently move between them. In research on dialog systems, the ability to actively and smoothly transition to new topics is often ignored. In this paper we introduce TIAGE, a new topic-shift aware dialog benchmark constructed utilizing human annotations on topic shifts. Based on TIAGE, we introduce three tasks to investigate different scenarios of topic-shift modeling in dialog settings: topic-shift detection, topic-shift triggered response generation and topic-aware dialog generation. Experiments on these tasks show that the topic-shift signals in TIAGE are useful for topic-shift response generation. On the other hand, dialog systems still struggle to decide when to change topic. This indicates further research is needed in topic-shift aware dialog modeling.\textsuperscript{1}

1 Introduction

Existing dialog models (Ghandeharioun et al., 2019; Einolghozati et al., 2019; Liu et al., 2018) have been reported to perform well in generating on-topic utterances in dialog scenarios. However, those models still struggle to proactively generate appropriate topic-shift utterances in conversations (Holtzman et al., 2020; Zhang et al., 2020a).

It is beneficial for dialog systems to be able to shift topics fluently. As shown in Figure 1, topic-shift behaviors are commonly observed in human conversations (Brown and Yule, 1983). Fluent topic shifts therefore are crucial for dialog models to be able to model or mimic human conversational patterns. Proactively using topic shifts can help chatbots guide conversations to a pre-defined target (Tang et al., 2019). Furthermore, switching topics allows chatbots to maintain engaging conversations with users. Without the ability to actively shift topics away from tired topics, chatbots risk generating dull responses or repeating themselves regarding a specific topic.

To facilitate research on topic-shift dialog modeling, we curate a Topic-shIft Aware dialoG datasEt (TIAGE) by augmenting the PersonaChat dataset (Zhang et al., 2018) with topic-shift annotations. To the best of our knowledge, TIAGE is the first dataset that focuses on topic-shift behaviors in open-domain dialog data. TIAGE contains a human annotated dataset with 7,861 gold standard topic-shift annotations, and a weak supervision dataset to adapt pretrained NLG systems to PersonaChat-style data. The inter-annotator agreement for topic-shift annotations in TIAGE is 0.479.

With TIAGE, we propose three tasks to study topic-shift behaviors: topic-shift detection, topic-shift triggered response generation and topic-aware dialog generation. The topic-shift detection task asks models to detect whether the ongoing topic has shifted or should shift. The other two tasks focus on modeling topic-shift behaviors in response generation. Specifically, the topic-shift triggered response generator receives a fixed topic-shift signal to generate topic-shift responses, whilst the topic-aware dialog generation task requires dialog systems to

\textsuperscript{1}Code and data available at: https://github.com/HuiyuanXie/tiage.
predict the topic-shift trigger by themselves.

Our experiments reveal that the topic-shift signals in TIAGE indeed improve dialog systems’ ability to generate topic-shift responses. However, it is difficult for dialog models to predict when it is appropriate to change topics. These observations highlight the need for better modeling of topic shifts in dialog generation. We hope our benchmark can motivate further research on topic-shift aware dialog modeling.

2 Related Work

Existing work in dialog systems falls into two broad categories. Task-oriented dialog systems (Budzianowski et al., 2018; Liu et al., 2018) help users complete tasks in specific domains. Open-domain dialog systems (Chen and Gao, 2017; Tang et al., 2019) allow agents to have open-ended conversations with users. Most existing dialog models (Fang et al., 2018; Zhang et al., 2020b; Ghandeharioun et al., 2019) emphasize end-to-end response generation, and do not explicitly address the topic-shift problem in dialog generation.

Early work in topic detection and segmentation (Hirschberg and Litman, 1993; Passonneau and Litman, 1997) focused on identifying cue phrases (such as on a different note) or examining lexical cohesion to segment topical chunks. Other work (Fiscus and Doddington, 2002) investigated topic detection and tracking (TDT) in a stream of broadcast news stories. More recent work (Glavas and Somasundaran, 2020) has explored utilizing neural networks to address topic segmentation. Although some of the existing work (Galley et al., 2003; Arnold et al., 2019) has investigated topic detection in dialog-style data, the generation aspect of topic-shift modeling in dialog settings is still unclear.

3 Topic-Shift Aware Dialog Dataset

In this section we introduce the rationale for our choice of data source, the human annotation process of topic-shift labelling in TIAGE and its data statistics. We also analyze the linguistic patterns of topic-shift utterances in TIAGE.

Rationale for our choice of data source. We construct TIAGE by augmenting the PersonaChat dataset (Zhang et al., 2018) with topic-shift human annotations. We view PersonaChat as a suitable dataset for topic-shift annotation for the following reasons: (1) the Personachat data was collected online in a textual form by mimicking chit-chat scenarios, where natural shifts of topics are more likely to happen; (2) dialogs in this dataset contain more than 10 dialog turns, and longer dialog contexts tend to exhibit a conversational flow with more topics; and (3) despite the fact that some participants in PersonaChat may have rushed into changing topics to quickly exchange their profile information, we observed that most of the participants still manage to change topics in a more natural and coherent way, making this dataset a favorable choice to study topic-shift behaviors.

Human annotation process. For the annotation pool, we have a total number of 25 human annotators. We randomly selected 500 dialogs from the original PersonaChat dev/test datasets, resulting in 7,861 dialog turns to label. Each dialog turn was randomly assigned to and independently labeled by 2 annotators. For each dialog turn, we asked annotators to indicate whether they think the conversational topic is changed at that turn. During the annotation process, all annotators were talked through the general aim of this annotation task and given the same annotation guidelines (see Appendix A.1 for details).

Since topic is co-constructed, it is rather limiting to analyze a turn for itself when trying to identify topic transitions. To facilitate the recognition of slowly transitioned topics, we encouraged the annotators to take into account both the previous two turns and the following two turns of the target dialog turn to make a decision. This helped decision making for cases where topics are slowly developed and transitioned.

After annotating, we obtained a dialog dataset...
with gold standard topic-shift labels for 7,861 dialog turns. The Cohen’s Kappa score for all annotations is 0.479\(^2\). Annotated examples of TIAGE dialogs are shown in Appendix A.2.

Dataset statistics. As shown in Table 1, TIAGE provides weak supervision data and human annotated data to train dialog models. Weak supervision data is selected from the original PersonaChat training set and helps adapt NLG models to PersonaChat-style data. The weak supervision data consists of 8,939 dialogs and is split into two sets: WeakSupo\(\text{train}\) and WeakSupo\(\text{dev}\). Human annotated data consists of 500 annotated dialogs with topic-shift annotations at each dialog turn. We split them into 300 ANNO\(\text{train}\), 100 ANNO\(\text{dev}\) and 100 ANNO\(\text{test}\) dialogs respectively. As each dialog has multiple dialog turns, we extract (context, response) pairs as instances for all turns in each dialog.

Analysis of topic-shift patterns. We examine a number of topic-shift utterances labeled by human annotators. We find that many of the topic-shift responses demonstrate an interesting pattern of [comment, topic shift]. More specifically, the response that changes the conversational topic is typically a brief comment on the previous dialog context, tailed by a topic-shift sentence with a different conversational focus. The comment usually corresponds to the sentiment previously expressed in the dialog.

This pattern echoes some of the findings in pragmatics research (Brown and Levinson, 1987; Goldsmith, 2007). When speakers introduce a new topic, it is a common positive politeness strategy (Leech, 2007). When speakers introduce a new topic, it is a common positive politeness strategy (Leech, 2007). When speakers introduce a new topic, it is a common positive politeness strategy (Leech, 2007). When speakers introduce a new topic, it is a common positive politeness strategy (Leech, 2007).

4 Tasks of Topic-Shift Modeling

Along with dialog utterances, TIAGE also provides gold standard topic-shift labels for dialog turns. This enables us to model topic shifts in dialog scenarios. We first introduce two tasks: topic-shift detection and topic-shift triggered response generation. They can be considered as intermediate steps of the topic-aware dialog generation task.

4.1 Preliminary of Response Generation

When considering a specific turn in a dialog, we denote the current utterance and all its previous utterances as the context \(X_T = \{x_1, ..., x_i, ..., x_N\}\) where \(x_i\) is the \(i\)-th utterance in the dialog history, and \(N\) is the context length. Then we expect the response to be generated after the current utterance \(x_N\). We denote a topic-shift response as \(s_{TS} = \{s_1, ..., s_i, ..., s_T\}\) where \(s_i\) is the \(i\)-th token in the response and \(T\) is the sentence length. Similarly, an on-topic response is denoted as \(s_{NTS} = \{\bar{s}_1, ..., \bar{s}_i, ..., \bar{s}_M\}\) where \(\bar{s}_i\) is the \(i\)-th token in the response and \(M\) is the sentence length.

4.2 Topic-Shift Detection

Topic-shift detection is a fundamental task that evaluates models’ ability to detect topic-shift occurrence at dialog turns.

Task definition. We introduce two settings for this task. In the retrospective setting, models have access to both the dialog context \(X_T\) and the corresponding response (either \(s_{TS}\) or \(s_{NTS}\)) to detect topic-shift occurrence, whilst in the predictive setting, models are asked to make topic-shift predictions based on the context \(X_T\) only.

Topic shift classifiers. We first implement three retrospective classifiers. We employ GenEnc which uses the GEN Encoder (Zhang et al., 2019) to separately encode dialog context and response into embeddings to estimate the topic-shift intents. GenEnc uses a cosine similarity threshold of 0.25 to filter out (context, response) pairs, and classify them as topic-shift occurrences. Then we implement a BERT-Wiki727k model (Devlin et al., 2019) trained on the Wiki727k dataset (Koshorek et al., 2018). We also employ a T5 model (Raffel et al., 2019) fine-tuned on the ANNO\(\text{train}\) data with topic-shift labels as our retrospective T5 topic-shift classifier (denoted as RetroTS-T5).

For the predictive setting, we implement a T5-based topic-shift manager (denoted as TSManager) and finetune it on the ANNO\(\text{train}\)
data. The major difference between RetroTS-T5 and TSManger is that RetroTS-T5 has access to both the dialog context and the response, while TSManger makes topic-shift predictions based solely on the context.

4.3 Topic-Shift Triggered Response Generation

This task examines models’ ability to generate topic-shift utterances when a need to change topics is signaled.

Task definition. Given a dialog context $X_T$, the topic-shift triggered response generation task requests models to directly generate a response $\hat{s}_{TS}$ that shifts the conversation to a different topic.

Topic-shift triggered generator. We build a T5-NLGrTS response generator using the pretrained T5 model. We first train the T5 model on the WEAKSUPOTrain data, and then further fine-tune it on the topic-shift instances (i.e., where topic shifts occur) in the ANNOTrain data.

Compared approaches. We also try a number of topic-insensitive NLG models for comparison. We train a T5-NLG model on the WEAKSUPOTrain data without any topic-shift signals. We use the DialoGPT model (Zhang et al., 2020b) finetuned on the same data as another baseline.

4.4 Topic-Aware Dialog Generation

The third task we propose targets more difficult and realistic topic-shift modeling in dialog generation.

Task definition. More formally, given a dialog context $X_T$, the goal of the topic-aware dialog generation task is to generate a topic-shift response $\hat{s}_{TS}$ if a change of topic is needed, or an on-topic response $\hat{s}_{NTS}$ if otherwise. The topic-aware dialog generation task asks models to identify the need to change topics by themselves and generate topic-shift or on-topic responses according to the prediction.

Topic-aware dialog system. Our topic-aware dialog system (TADial) is a pipeline system. We separately train two T5-based response generators: T5-NLGrTS and T5-NLGNrTS. We switch between the two response generators to produce either a topic-shift or on-topic response, guided by the topic-shift signals from TSManger. T5-NLGrTS aims to generate topic-shift responses, while T5-NLGNrTS is finetuned on non-topic-shift instances to generate on-topic responses.

Compared approaches. We use the T5-NLG and DialoGPT models finetuned on the WEAKSUPOTrain data as baselines for comparison.

5 Evaluation Results

We report here the evaluation results for baseline systems on the above three tasks.

Topic-shift detection. We test topic-shift classifiers on the annotated ANNOTest split. From Table 2 we observe that RetroTS-T5 outperforms other approaches by a large margin and is on par with human performance. This indicates that topic shifts in PersonaChat dialogs exhibit certain patterns, which can be captured from our human-labeled topic-shift annotations by our retrospective T5 classifier. We also notice that there is a clear gap in classification performance between RetroTS-T5 and TSManger. The predictive setting of TSManger is inherently harder than RetroTS-T5, as it is asked to predict topic-shift labels based solely on dialog context.

Topic-shift triggered response generation. In Table 3 we report evaluation results of our topic-shift triggered response generator (T5-NLGrTS) and two topic-insensitive models (DialoGPT and T5-NLG). Models are tested on the topic-shift instances in ANNOTest. We observe that T5-NLG yields better performance than DialoGPT. Furthermore, T5-NLGrTS achieves better evaluation results on topic-shift test instances, outperforming T5-NLG by 16.46% in BLEU-2 and 9.94% in ROUGE_L. The better performance of T5-NLGrTS

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-2</th>
<th>METEOR</th>
<th>ROUGE_L</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>DialoGPT</td>
<td>0.060</td>
<td>0.077</td>
<td>0.125</td>
<td>0.104</td>
</tr>
<tr>
<td>T5-NLG</td>
<td>0.079</td>
<td>0.086</td>
<td>0.161</td>
<td>0.170</td>
</tr>
<tr>
<td>T5-NLGrTS</td>
<td>0.092</td>
<td>0.092</td>
<td>0.177</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Table 3: Evaluation results of topic-shift triggered response generation on topic-shift instances in ANNOTest.

We use the annotations from one annotator as gold standard references, and calculate human performance on the annotations from the other annotator.

We use the nlg-eval package for automatic evaluation: https://github.com/Maluuba/nlg-eval.
validates the effectiveness of topic-shift signals in improving topic-shift response generation. It also proves that explicitly modeling topic-shift behaviors can potentially benefit dialog generation.

**Topic-aware dialog generation.** We test TADial and two topic-insensitive baselines on all instances in ANNOtest. From Table 4, we can see that TADial with a dedicated topic-shift management component does not yield better performance over the T5-NLG model which is simply trained on dialog instances with no topic-shift labels. This points out that due to the deficiency of TSManager signals, hard-wiring a topic-shift management component into the generation pipeline falls short to improve generation results. It remains a challenging task to produce well-timed and good-quality topic-shift signals based on dialog context only, which hinders overall topic-aware dialog generation.

### 6 Conclusion and Future Work

We construct the TIAGE dataset with human annotated topic-shift labels on the basis of the PersonaChat dataset. Based on TIAGE, we introduce three tasks: topic-shift detection, topic-shift triggered response generation and topic-aware dialog generation. Empirical results show that topic-shift labels in TIAGE are useful for topic-shift response generation. However, it remains a challenging task for dialog models to predict good-quality topic-shift signals based on dialog context only. Further research is needed on selecting appropriate topics to shift to among multiple references. Natural topic shifts can be both a precaution against, and a remedy to, dull and repetitive response generation in real-world dialog applications. TIAGE with its topic-shift annotations can help direct future investigation on the incorporation of topic-shift tactics in dialog models, which allows more effective control over topic-shift aware dialog generation.

### Acknowledgments

We thank the anonymous reviewers for their constructive feedback. We thank our annotators from the University of Cambridge and Tsinghua University for their help in annotating the topic-shift labels used in this paper. Huiyuan Xie is grateful for being supported by the CSC Cambridge Scholarship. Zhenghao Liu is supported by National Natural Science Foundation of China (NSFC) under grant No. 61872074 and 61772122. This work is partly supported by the National Key Research and Development Program of China (No. 2020AAA0106501).

### References


A Appendix

A.1 Human Annotation Guidelines

Here we present the annotation guidelines used for the human annotation process in this work.

Task description. Chitchat systems are expected to have the ability to proactively change conversational topics when necessary. For occasions when a chat agent runs out of things to say or the current discussion is starting to get boring, topic shifting is a common tactic to keep the conversation going on. In this work, we aim to model topic-shift phenomenon in open-domain dialog settings. To achieve this, we need to construct a new dialog dataset with topic-shift signals.

```
| Speaker1: | My dad works for the New York Times. |
| Speaker2: | Oh wow! You know, I dabble in photography; maybe you can introduce us sometime. |
| Speaker1: | Photography is the greatest art out there. \(\rightarrow\) not a topic shift |
```

(a) Commenting on the previous context.

```
| Speaker1: | Do you teach cooking? |
| Speaker2: | No, since I’m a native of Mexico, I teach Spanish. \(\rightarrow\) not a topic shift |
```

(b) Question answering.

```
| Speaker1: | Pets are cute! |
| Speaker2: | I heard that Huskies are difficult dogs to take care of. \(\rightarrow\) not a topic shift |
```

(c) Developing the conversation to sub-topics.

```
| Speaker1: | You are an artist? What kind of art, I do American Indian stuff. |
| Speaker2: | Yes, I love to draw. I love to eat too, sometimes too much. \(\rightarrow\) topic shift |
```

(d) Introducing a relevant but different topic.

```
| Speaker1: | What do you do for fun? |
| Speaker2: | I drive trucks so me and my buds go truckin in the mud. |
| Speaker1: | Must be fun! My version of that’s running around a library! |
| Speaker2: | Do you have a favourite animal? Chickens are my favourite. I love them. \(\rightarrow\) topic shift |
```

(e) Completely changing the topic.

Table 5: Different scenarios of dialog response in conversations.

Data annotation. For each utterance in a dialog, annotators are asked to decide whether the topic of the conversation changes when transiting from the current utterance to the following response. If there is a topic shift, annotators should label the response with “1”, otherwise label it with “0”.

```
<table>
<thead>
<tr>
<th>Dialog</th>
<th>TS Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Speaker1:] Hi! How are you this evening?</td>
<td>N/A</td>
</tr>
<tr>
<td>[Speaker2:] Good. I spent all afternoon walking my dogs. I’ve three Labradors.</td>
<td>0</td>
</tr>
<tr>
<td>[Speaker1:] Cool, that’s a lot of dogs. Do you like music? I love it.</td>
<td>1</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>Dialog</th>
<th>TS Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Speaker1:] I think you are great. You are my best friend.</td>
<td>N/A</td>
</tr>
<tr>
<td>[Speaker2:] My best friend is a bear, bears don’t have friends, that’s why they’re my favourite.</td>
<td>0</td>
</tr>
<tr>
<td>[Speaker1:] Webster’s dictionary defines weddings as the fusing of two metals with a hot torch.</td>
<td>1</td>
</tr>
</tbody>
</table>
```

Table 6: Annotated dialog examples in TIAGE.

In conversations, the response of a speaker to the dialog context usually falls into one of the following cases (see examples in Table 5):

- (a) Commenting on what the other participant just said (the most common scenario);
- (b) Question answering;
- (c) Developing the conversation to sub-topics;
- (d) Introducing a relevant but different topic;
- (e) Completely changing the topic.

Other tips for data labeling. A number of words and phrases are often used as indicators for topic shifts, including but not limited to: but, speaking of, talking about, anyway, by the way, that reminds me, before I forget, I want to mention, let’s talk about, we need to discuss, funny you should mention that, not to change the subject but, changing the topic slightly, totally unrelated, on a different/relevant note.

A.2 Examples of Labeled Data in TIAGE

In Table 6 we showcase examples of labeled dialogs selected from TIAGE.

A.3 Implementation Details

For the topic-shift classifiers, we use the base version of BERT and T5 models, initialized from their pretrained weights. For the dialog response generation experiments, we use the small version of DialoGPT and the base version of T5. Our implementation is based on the HuggingFace Transformers library (Wolf et al., 2020). All models are optimized using Adam with a learning rate of 5e-5 and a batch size of 64. We set the maximum input sequence length to 512. The training is carried out using 1 Nvidia RTX 8000 GPU and takes around 15 hours.