Guiding Abstractive Dialogue Summarization with Content Planning

Ye Wang1, Xiaojun Wan2,3,4∗, Zhiping Cai1∗
1 College of Computer, National University of Defense Technology
2 Wangxuan Institute of Computer Technology, Peking University
3 Center for Data Science, Peking University
4 The MOE Key Laboratory of Computational Linguistics, Peking University
{wangye19,zpcai}@nudt.edu.cn, wanxiaojun@pku.edu.cn

Abstract

Abstractive dialogue summarization has recently been receiving more attention. We propose a coarse-to-fine model for generating abstractive dialogue summaries, and introduce a fact-aware reinforcement learning (RL) objective that improves the fact consistency between the dialogue and the generated summary. Initially, the model generates the predicate-argument spans of the dialogue, and then generates the final summary through a fact-aware RL objective. Extensive experiments and analysis on two benchmark datasets demonstrate that our proposed method effectively improves the quality of the generated summary, especially in coherence and consistency.

1 Introduction

With the prevalence of dialogue texts, new challenges have arisen for abstractive dialogue summarization (Zechner, 2001). Dialogue are often informal and backchanneling, with the salient information and speaker interactions scattered across the whole chat (Chen and Yang, 2020; Liu et al., 2019). However, existing methods (Goo and Chen, 2018; Wu et al., 2021) struggle to maintain factual consistency between dialogue and summary, mainly due to the failure of capturing interactions between plot points.

As a result, we propose an abstractive dialogue summarization model that decomposes the problem into a two-step coarse-to-fine generation problem (Figure 1). We first generate a series of predicate-argument spans as content plan. We use semantic role labeling (SRL), which focuses on modeling the skeleton of a sentence, to generate predicate-argument spans. It provides a weakly supervised signal and is easier for the model to learn dependencies across events. We then feed both the dialogue and content plan to the dual-encoder model, and train it with the fact regularization objective.

We evaluate the proposed model on two benchmarks: (i) SAMSum corpus (Gliwa et al., 2019), which is a large-scale chat summarization corpus, and (ii) DialogSum corpus (Chen et al., 2021), which is a real-life scenario dialogue summarization dataset. By comparison to previous approaches, our model provides a better generation quality judgment both by humans and by automatic evaluations. Furthermore, the results show that the outputs of our model are highly consistent and coherent.

In summary, we make the following contributions in this paper: (i)We explore the helpfulness of SRL-based content plan for abstractive dialogue summarization. (ii) We propose a novel training process with fact regularization, which incorporates the information of predicate-argument span. (iii) Experimental results show that our method outperforms several strong baselines. According
to a comprehensive case study and human evaluation, our model can achieve a more coherent and consistent summary.

2 Methodology

2.1 Overview and Notations

We formalize the problem of dialogue summarization as follows. Given a dialogue $X = (x_1, x_2, \cdots, x_N)$, where $N$ is the total number of words in the dialogue. The dialogue is coupled with its corresponding summary $Y = (y_1, y_2, \cdots, y_M)$ in the length of $M$.

We implement the Transformer model (Vaswani et al., 2017) initialized with BART as our backbone architecture. As illustrated in Figure 2, our model consists of a content plan generator and a summary generator.

2.2 Content Plan Generator

Our content plan generator is based on the standard Transformer model (Vaswani et al., 2017; Wei et al., 2013), which aims to generate sequences of SRL decomposition. SRL identifies predicates and argument in sentences and the SRL decomposition retains only the core arguments in order to focus on the main semantic structure.

We obtain the gold SRL decomposition of the summary by an off-the-shelf semantic role labeler, and separate predicate-argument span with delimiter tokens. We place the predicate verb between arguments without additional signals.

Given the dialogue $X$ and the gold content plan $Z$, the learning objective of the generator is defined as

$$L_{CG} = -\log \sum_D p(Z|X)$$

where $D$ denotes the training set.

2.3 Summary Generator

Our summary generator is also built on a Transformer-based model (Dou et al., 2021) which consists of a parameter sharing dual-encoder and hierarchical attending decoder.

Given the dialogue $X$, the content plan $Z$, and the reference summary $Y$, the learning objective of the summary generator is defined as

$$L_{LM} = -\log \sum_D p(Y|X, Z)p(Z|X)$$

However, the marginalization over $p(Z|X)$ is in general intractable. Instead, following (Fan et al., 2019), we minimize a variational upper bound of the loss by constructing a deterministic posterior $q(Z|Y) = 1_{Z=Z^*}$, where $Z^*$ can be given by running an off-the-shelf semantic role labeler on summary $Y$. As a result, we optimize the following loss:

$$Z^* = \arg \max_Z p(Z|X)$$

$$L_{LM} \leq -\log p(Y|X, Z^*) - \log p(Z^*)$$

Therefore, the model can be trained separately for $p(Z^*)$ and $p(Y|X, Z^*)$.

2.4 Fact-aware Training

To encourage the model to consider the factual consistency of the sampled SRL sequences, we incorporate reinforcement learning into our training process.

Given the dialogue $X$ and the content plan $Z$, the summary generator first samples an generated summary $Y' = (y'_1, \cdots, y'_|Y'|)$ which contains $|Y'|$ words. We then update the summary generator’s parameters $\theta$ as follows:

$$L_{RL} = -\mathbb{E}_{y' \sim p(y|X,Z)} S(Y, Z)
= -S(Y, Z) \sum_{y_i \in Y'} \log p(Y|X, Z)$$

The reward function $S(Y, Z)$ measures the structure of the sampled summary $Y'$ against the reference summary $Y$, and its extracted SRL sequence $Z'$ against the input content plan $Z$. We calculate the reward function $S(Y, Z)$ as follows:

$$S(Y, Z) = R(Y, Y') + R'(Z, Z')$$

where $R(\cdot, \cdot)$ is the ROUGE score (Lin, 2004). $R'(\cdot, \cdot)$ is the improved ROUGE score of predicate-argument span, where recall is defined as how
many gold triplets are covered by the extracted fact triplets from generated summary and precision is how many extracted triplets are matched with gold fact triplets. We regard two fact triplets as matched if they contain at least two overlapping components.

For the summary generator, we first train it with $L_{LM}$, and then incorporate the fact-aware RL objective to further train the generator with $L_{LM} + L_{RL}$.

3 Experiments

3.1 Datasets

We train and evaluate our models on a conversation summarization dataset SAMSum (Gliwa et al., 2019) and a real-life scenario dialogue summarization corpus DialogSum (Chen et al., 2021). We label these datasets with an off-the-shelf semantic role labeler\(^1\), which achieves very competitive results for SRL.

3.2 Implementation Details

Our implementation is based on the Fairseq\(^2\). For content plan generator, we use the BART-large parameter to initialize. For summary generator, following (Dou et al., 2021), the top layer is initialized with pretrained parameters, but the dual-encoder are separately trained. During decoding, the first cross-attention block is randomly initialized, while the second cross-attention block is initialized with pretrained parameters.

3.3 Metrics and Baselines

We evaluate all the models with the widely used automatic metric, ROUGE F1 scores (Lin, 2004), and report ROUGE-1 (unigram), ROUGE-2 (bigram) and ROUGE-L (longest common subsequence) scores. It measures overlapping between the generated summary and the reference summary. We utilize Py-rouge\(^3\) package for evaluation.

We compare our methods with several baselines: Lead-3 is an extractive baseline that concatenates the first-3 utterances of each dialogue. PGN (See et al., 2017) is an RNN-based abstractive model with an attention mechanism that enables the system copy words from source text via pointer generator. Fast-Abs (Chen and Bansal, 2018) is an reinforcement learning method that utilizes policy gradient to connect sentence selection and summary generation. Transformer (Vaswani et al., 2017) is a randomly initialized sequence-to-sequence method based on full self-attention operations. TGDGA (Zhao et al., 2020) uses topic words and models graph structures for dialogues. MV-BART (Chen and Yang, 2020) is a BART-based method that incorporates topic and state information. CODS (Wu et al., 2021) uses pronoun categories and key phrase extracted by a constituency parser as a weakly supervised signal.

3.4 Automatic Evaluation

The results on SAMSum are shown in Table 1. It is shown that Lead-3 is less suitable for dialogue summarization. Compared to PGN, utilizing semantic structures to accommodate dialogue (TGDGA) slightly increases ROUGE scores. It indicates that adding additional information, such as semantic information and dialogue structures, can be of great help in generating summaries. Fast-Abs utilizes policy gradient, which is optimized by the token-level objective, gains noticeable improvement compared to PGN. When using a pretrained transformer-based model, all ROUGE scores improve significantly and achieve over 10 points improvement on the ROUGE-1 score, which demonstrates the superiority of pretrained methods. CODS achieves higher ROUGE score compared with other models. Our model gains an improvement of ROUGE scores compared with other methods, which verifies the effectiveness of the proposed architecture for the dialogue summarization task.

We also report the performance of our model on DialogSum dataset in Table 2. This shows that the use of semantic roles in our methods has good generalizability across different datasets.

Table 1: Test set results on the SAMSum dataset. * and † denote the results from (Feng et al., 2021) and (Wu et al., 2021) respectively. ‡ and † denote the first-ranked and second-ranked results respectively.
Table 2: Test set results on the DialogSum dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer (Vaswani et al., 2017)</td>
<td>34.78</td>
<td>8.06</td>
<td>32.37</td>
</tr>
<tr>
<td>BART (Lewis et al., 2020)</td>
<td>46.11</td>
<td>20.03</td>
<td>43.52</td>
</tr>
<tr>
<td>Ours</td>
<td>48.76</td>
<td>22.34</td>
<td>45.49</td>
</tr>
</tbody>
</table>

Table 3: Human evaluation on SAMSum. The ratings are on a Likert scale of 1(worst) to 5(best).

<table>
<thead>
<tr>
<th>Model</th>
<th>Consistency</th>
<th>Informativeness</th>
<th>Coherence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODS</td>
<td>3.70</td>
<td>3.62</td>
<td>3.78</td>
</tr>
<tr>
<td>BART</td>
<td>3.61</td>
<td>3.59</td>
<td>3.73</td>
</tr>
<tr>
<td>Ours</td>
<td>3.82</td>
<td>3.67</td>
<td>3.85</td>
</tr>
<tr>
<td>w/o CG</td>
<td>3.79</td>
<td>3.65</td>
<td>3.81</td>
</tr>
<tr>
<td>w/o RL</td>
<td>3.74</td>
<td>3.64</td>
<td>3.83</td>
</tr>
</tbody>
</table>

Table 4: Ablation Studies on the SAMSum dataset.

Riley : Chloe is on tv!!
James : on which channel?
James : never mind i’ve found it
James : what is she doing? i don’t get it
Riley : this is a programme in which women undergo a complete metamorphosis.
Riley : OMG she looks drop dead gorgeous!
BART
Chloe is on TV. James doesn’t get it.

d. Riley and James watch Chole, undergoing a metamorphosis.

(a)
Frederick : do You like your new next door neighbors ?
Frederick : they seemed really cool yesterday when we ran into them
Ricky : they’re nice people but they’re incredibly noise
Ricky : they also have parakeet that wouldn’t stop squawking all night long
hahaha
Frederick : sucks to be you
BART
Frederick and Ricky have a new next door neighbors . Ricky doesn’t like them because of their noise.
Diffs
Content Plan. Frederick and Ricky met their new neighbours, their neighbors are nice but noisy, a parakeet squawking.
Generated Summary. Frederick and Ricky don’t like their new next door neighbors because of their noise and parakeet.
Reference
Gold Content Plan. Ricky’s new neighbour are nice but loud, a parakeet makes a lot of noise throughout the night.
Gold Summary. Ricky’s new neighbours are nice but loud . They own a parakeet that makes a lot of noise throughout the night.

(b)

Table 5: Sample summaries for SAMSum.

and gains higher factual consistency. For the example (a), we can see that the generated summary shows the correct sentiment and content, although ROUGE scores may not be high. For the example (b), it is shown that our model grasp important information - "parakeet".

4 Conclusion and Future Work

In this paper, we explore the helpfulness of SRL-based content plan composed of predicate-argument spans and propose a fact-aware RL training process for the dialogue summarization task. We observe that the use of semantic roles can improve the performance of the BART architecture. In the future, we plan to directly integrate semantic role information into other pretrained large generative models like GPT-3 and T5 to further improve the performance.
Limitations

There are several limitations of our proposed method:

- The method is in pipeline way that requires to generate content plan first. Compared with end-to-end method, it requires more tedious steps. It is worthwhile to explore more appropriate end-to-end methods for abstractive dialogue summarization.
- The method depends on the effect of the semantic role labeler. Using methods that do not rely on a solid labeler is also a direction worth exploring.

References


Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In Proceedings of the 28th