Analyzing and Evaluating Faithfulness in Dialogue Summarization

Bin Wang†, Chen Zhang†, Yan Zhang†, Yiming Chen†, Haizhou Li‖, †, §
†National University of Singapore, Singapore
‡The Chinese University of Hong Kong, Shenzhen, China
§Kriston AI, China
bwang28c@gmail.com

Abstract

Dialogue summarization is abstractive in nature, making it suffer from factual errors. The factual correctness of summaries has the highest priority before practical applications. Many efforts have been made to improve faithfulness in text summarization. However, there is a lack of systematic study on dialogue summarization systems. In this work, we first perform the fine-grained human analysis on the faithfulness of dialogue summaries and observe that over 35% of generated summaries are faithfully inconsistent respective the source dialogues. Furthermore, we present a new model-level faithfulness evaluation method. It examines generation models with multi-choice questions created by rule-based transformations. Experimental results show that our evaluation schema is a strong proxy for the factual correctness of summarization models. The human-annotated faithfulness samples and the evaluation toolkit are released to facilitate future research toward faithful dialogue summarization. Code available: https://github.com/BinWang28/FacEval.

1 Introduction

Text summarization aims to condense a document into a short paragraph or a single sentence while conveying the core information (El-Kassas et al., 2021). It can be either extractive or abstractive. Extractive summarization methods identify salient sequence spans from the source document and pasting them together (Dorr et al., 2003; Kobayashi et al., 2015; Zhong et al., 2020). Abstractive summarization methods generate completely new summary in a coherent manner (Paulus et al., 2018; Lewis et al., 2020; Zhang et al., 2020; Liu et al., 2021a; Wang et al., 2022). Abstractive summarization methods generate completely new summary in a coherent manner (Paulus et al., 2018; Lewis et al., 2020; Zhang et al., 2020; Liu et al., 2021a; Wang et al., 2022). Previous work discovered that abstractive summarization suffers from unfaithful outputs, limiting its applicability in real-world scenarios (Kryscinski et al., 2020; Falke et al., 2019; Zhu et al., 2021a; Ladhak et al., 2022).

An essential way of exchanging information, conversations usually involve multiple participants, informal language usage, repetition, and negations (Sacks et al., 1978; Chen and Yang, 2020). Therefore, dialogue summarization is vulnerable to factual issues due to its abstractive nature. Table 1 gives an example of factually incorrect dialogue summaries. The problem of factual correctness is broadly studied for text summarization in news and article domains (Nallapati et al., 2016; Narayan et al., 2018). The progress is primarily because of the availability of factually annotated data at both summary and token levels (Kryscinski et al., 2020; Wang et al., 2020; Pagnoni et al., 2021; Cao et al., 2022). Many studies are proposed to evaluate and reduce factual errors in the generated summaries. However, due to the interactive nature of dialogues, we cannot simply transfer these methods to dialogue summarization.

In this work, we first categorize the most frequently occurred factual errors for dialogue sum-

Table 1: A real example from SAMSum dataset. Span of factual errors are marked with underline.
marization into 6 types. Then, we collect fine-grained factual annotations for human reference and the output of 4 recent dialogue summarization systems (§3). At least two annotators are involved, and a verification process is incorporated to ensure the annotation quality. As a result, our study on human-annotated data suggests that over 35% of the generated dialogue summaries contain at least one factual error. Similar observations have been made in the news summarization domain where 30%-80% of generated text are factually inconsistent (Cao et al., 2018; Pagnoni et al., 2021). More research attention should be made toward faithful dialogue summarization.

The unavailability of faithful evaluation methods hinders the development of effective dialogue summarization models. In this work, we present a model-level evaluation schema, FacEval, targeting dialogue summarisation models’ faithfulness (§4). First, we synthesize a set of positive and negative summaries for each dialogue with back-translation or rule-based transformations. Then, a summarization model is asked to distinguish positive and negative summaries based on conditional generation probabilities. More correct judgements indicate the model is more factually competent.

To compare the model-level performance of evaluation methods, we leverage two ad-hoc training schema to synthesize a series of models with different capability ranks. Then, the evaluation methods are used to predict the ranking of trained models. Seven non-factual and factual evaluation methods have been examined, followed by a detailed discussion of their properties. The effectiveness of FacEval is also proven by showing a strong correlation with the factual correctness of summarization models.

2 Related Work

2.1 Summarization Methods

Text summarization is one of the most important tasks in natural language generation (NLG). With the development of pre-trained language models, a lot progress has been made to abstractive text summarization (See et al., 2017; Zhang et al., 2020; Liu et al., 2022), especially in news domain (Hermann et al., 2015; Narayan et al., 2018). With the availability of datasets (Carletta et al., 2005; Gliwa et al., 2019; Zhu et al., 2021), dialogue summarization research has attracted a lot of attention. For dialogue summarization, fine-tuning pre-trained generation models including T5 (Raffel et al., 2020), PEGASUS (Zhang et al., 2020) and BART (Lewis et al., 2020) are served as a strong baseline, where BART achieves the SOTA performance on ROUGE scores. Some recent works consider the dialogue properties for more advanced summarization models. Chen and Yang (2020) and Liu et al. (2021a) incorporate the conversational structures into the semantic encoding process of dialogue. Conversations involve lots of co-references. Therefore, Liu et al. (2021b) proposes injecting co-reference information into the transformer layers by adapting attention maps or through graph convolutional networks (GCN). We include the outputs of recent dialogue summarization models in our analysis.

2.2 Faithfulness Analysis

Previous works spot that the factual consistency problem is one key aspect of improving text summarization (Kryscinski et al., 2020; Cao et al., 2020). The analysis of factual errors in summaries is mainly performed in the news domain. Kryscinski et al. (2019) and Falke et al. (2019) conducted the initial crowdsourcing of binary factual annotations and found that nearly 30% of the generated summaries are factually inconsistent. Recent extensions focus on more fine-grained analysis (Cao and Wang, 2021; Pagnoni et al., 2021) and also discovering factual evidences at entity level (Cao et al., 2022) or span level (Huang et al., 2020; Maynez et al., 2020; Goyal and Durrett, 2021).

Recently, CONFIT presented the first study on the faithfulness of dialogue summaries (Tang et al., 2022b). Similar to our work, they also define a taxonomy of factual errors and conduct fine-grained annotations. However, they focus on comparing reference summaries and generated summaries without referring to the whole dialogue. It is sub-optimal because the reference summary cannot fully represent the entire dialogue and also can be incorrect according to our analysis in Section 3. Besides, the missing and redundant information is categorized as factual errors, which we consider less proper. More recent advanced dialogue summarization models are also not included in their analysis.

2.3 Faithfulness Evaluation

The default evaluation metric for summarization, ROUGE, is based on n-gram overlaps between a generated summary and the corresponding references, rendering it less sensitive for capturing fac-
Speaker 1: Fiona

What should I prepare for my dad’s birthday? 
Turning 50.

I know, but I don’t have any idea. 
He watches a lot of military movies. 
No, not even close.

I don’t know how my mum will react but I like it :D

Speaker 2: Jonathan

How old is he? 
Wow, a round birthday, it must be something big.

What does he like? 
U said he likes military... maybe paintball?

Jonathan suggests a paintball match.

Table 2: An illustration of the taxonomy on factual error types.

<table>
<thead>
<tr>
<th>Ref. Summary</th>
<th>Subject Object Error (SubObjE)</th>
<th>Pronoun Error (ProE)</th>
<th>Negation Error (NegE)</th>
<th>Particulars Error (ParE)</th>
<th>Hallucination Error (HalE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiona doesn’t know what she should give to her dad as a birthday gift. He likes military. Jonathan suggests a paintball match.</td>
<td>Jonathan doesn’t know what she should give to her dad as a birthday gift. He likes military. Jonathan suggests a paintball match.</td>
<td>Fiona doesn’t know what he should give to her dad as a birthday gift. He likes military. Jonathan suggests a paintball match.</td>
<td>Fiona doesn’t know what she should give to her dad as a birthday gift. He likes military. Jonathan suggests a paintball match.</td>
<td>Fiona doesn’t know what she should give to her dad as a Christmas gift. He likes military. Jonathan suggests a paintball match.</td>
<td>Fiona doesn’t know what she should give to her dad as a birthday gift. He likes military. Jonathan invites Fiona to watch a military movie.</td>
</tr>
</tbody>
</table>

3 Fine-grained Faithfulness Analysis

Previous studies of factuality analysis in summarization mainly focus on the news domain. The typology of factual errors for dialogues can be very different. Therefore, we first define a taxonomy of frequently occurred factual errors for dialogue summaries. A fine-grained analysis is then performed by measuring the factual consistency within dialogue summary pairs.

3.1 Taxonomy of Factual Errors

We collect the generated summaries using four SOTA dialogue summarization models on the popular dialogue summarization dataset, SAMSum (Gliwa et al., 2019). The selected models are BART (Lewis et al., 2020), MV-BART (Chen and Yang, 2020), Coref-BART (Liu et al., 2021b) and CondigSum-BART (Liu et al., 2021a). We define five most frequently occurred error types in dialogue summaries as below. An example for each error type is shown in Table 2.

Subject Object Error (SubObjE): The subject(s) or object(s) involved for an event is (partially) wrong. It includes substitution, addition and deletion of any related subject(s) or object(s).

Pronoun Error (ProE): Pronoun references are frequently occurred in dialogue summarization. This error includes wrong references and ambiguous ones that cannot be fully understandable relying on the summary.

Negation Error (NegE): Dialogues can contain confirmation utterances. This error means that the generated summary makes wrong conclusions when contradictory or unconfirmed events are presented in the dialogue.

Particulars Error (ParE): The summary presents related events, but some details are inaccurate or faulty. It can include incorrect information like date, time and location.

Hallucination Error (HalE): Generation models have the imaginary ability and can be triggered by certain prompt words in the dialogue. The hallucination error refers to the cases where the summary contains events not presented in the dialogue.

Other Error (OthE): It is used to classify factual errors that do not belong to any of the above types. Note that the above-mentioned error types are not exclusive to each other. That is, one summary may contain multiple error types.

3.2 Annotation Process

We random sample 150 dialogues from the test set of SAMSum. Five summaries are listed for each dialogue, including the human-written one and four model-generated summaries.

Falke et al. (2019) founds that it needs at least 12 annotations to reach an inter-annotator agreement of coefficient $k = 0.75$, which can lead to high annotation costs and unreliable results with fewer annotators (Kryscinski et al., 2020). There-
Figure 1: The proportion of summaries with different types of factual errors. Note that one summary can contain multiple error types.

Figure 2: The proportion of summaries with at least one factual error.

Therefore, we perform a two-step verification process to ensure the annotation quality. First, each sample is annotated by two distinct annotators. If there is a disagreement about whether a summary contains factual errors, a third annotator is involved in making the final decision while considering inputs from the previous two annotators. As a result, we have collected 750 fine-grained faithfulness annotations from 30 participants.

3.3 Results and Analysis

The detailed annotation results are shown in Figure 1. There are several exciting findings: 1) the human annotations contain non-negligible factual errors at around 17%; 2) 36% to 50% of generated summaries from dialogue summarization models contain at least one factual error; 3) three advanced dialogue summarization models perform worse than their baseline on factual consistency.

First, the popular SAMSum dataset (Gliwa et al., 2019) associates each dialogue with one human-written reference summary. However, we found that 17% of reference summaries have factual errors. Therefore, we encourage people to be aware of the issue, especially for evaluation. It is because the dialogue annotation process for SAMSum only involved one annotator per sample, and no further verification process was executed. We notice that the source of factual errors for human summaries is also different from machine-generated ones. Some factual errors in human-written summaries are caused by typos, which rarely occur in machine-generated summaries.

For dialogue summarization models, we found that 35%-50% of generated summaries contain factual errors. The most frequent error types are SubObjE and ParE. Because dialogue often involves scattered information exchange with multiple speakers in multiple turns, it is very challenging to accurately locate who and whom in who-did-what-to-whom. That is the leading cause of SubObjE. ParE is the second most frequent error type, indicating that the generated summaries express the same topic but do not accurately capture the details. OthE occurs less frequently. It shows that our taxonomy of factual errors can cover the most frequent error types for dialogue summarization.

Surprisingly, we found that MV-BART, Coref-BART and CondigSum-BART perform even worse than the baseline model, with an increase of around 10% overall factual error rate. They are accepted as more advanced summarization models and perform better on ROUGE scores. It indicates that enhancing topical information is not necessarily contributing much to factuality (Chen and Yang, 2020; Liu et al., 2021a). Coref-BART aims to improve BART with co-reference information (Liu et al., 2021b). However, our result shows it does not bring obvious benefits. In conclusion, we encourage the future development of summarization models to pay more attention to the factuality perspective, and a more diverse evaluation schema.
4 Model-level Faithfulness Evaluation

Some efforts have been made toward sample-level factual error evaluation. An example is shown in Figure 3. The sample-level evaluation methods are model-agnostic and examine a model solely based on its output sequences. Most existing evaluation methods, including ROUGE score, human evaluation and recent factual evaluation methods, belong to this type. One ultimate goal for factuality evaluation is to discriminate better summarization models. We propose directly probing models’ generation probability with a constrained search space. First, FacEval generates a set of positive and negative samples with variant factual errors by rule-based transformations. Then, the generation probabilities of positive and negative summaries are compared for each dialogue. A better summarization model should be more likely to generate positive summaries than negative ones.

4.1 Dialogue-summary Pair Generation

We design transformations to synthesize negative samples with factual errors. Given the source text and target text, one or more modifications are performed to the target text while referring to the information of the source text. It is because the frequently occurred errors are conceptual confusions from the source. Our designed transformations are listed as follows:

- **Speaker Swap (SS):** We first spot the name of speakers from the source text by colon symbol and then swap the names at the target text.

- **Entity / Pronoun / Date / Number Swap (ES / PS / DS / NS):** An NER system is first applied to both source and target text. The entities from the target text are randomly swapped with entities from the source text if they share the same entity type.

- **Negation (NG):** Negation is performed using a set of hand-crafted rules. Auxiliary verbs are first scanned. Then, positive verbs are negated by adding not or n’t. Similarly, negative sentences are inverted by negation removal.

First, we paraphrase the summary to create more positive samples through back-translation (BT). The Google Cloud API is leveraged for this task\(^1\). Then, we generate new summaries with factual errors by corrupting positive summaries, which means the summaries are treated as the target text, and the dialogue is the source text.

Ideally, the negative summaries should be prone to errors generated in real-world scenarios. Therefore, our designed transformations try to mimic that. In the context of the analysis presented in Section 3, we have the following list of correspondences: 1) SS-SubObjE; 2) PS-ProE; 3) NG-NegE; 4) ES/DS/NS-ParE.

4.2 Comparison of Generation Probabilities

An illustration of probability comparison is shown in Figure 4. Given a dialogue \(D\), a summary \(S\) = \([y_1, \ldots, y_L]\) and a summarization model \(f_s(\cdot)\), we can compute a generation score (GS) for \(D\)-\(S\) pair.

\(^1\)https://cloud.google.com/translate/
from the generation probability:

\[ GS(S|D) = \frac{1}{L^\alpha} \log P(y_1, \ldots, y_L|D) \]

\[ = \frac{1}{L^\alpha} \sum_{t=1}^{L} \log P(y_t|y_1, \ldots, y_{t-1}, D) \]

where the generation probability for each token is as follows:

\[ P(y_t|y_1, \ldots, y_{t-1}, D) = P(f_s(y_1, \ldots, y_{t-1}, D) = y_t) \]

We leverage the above generation score from decision process of beam search algorithm (Graves, 2012), where the sequence length is taken into consideration. In default, we set the length penalty parameter \( \alpha \) as 1.0.

For dialogue \( D_i \), there is positive summary set \( S = [S_1, \ldots, S_M] \) and negative summary set \( \hat{S} = [\hat{S}_1, \ldots, \hat{S}_N] \). We evaluate the number of times the positive samples have higher scores than the negative samples concerning the same dialogue. The factuality score (FS) of model \( f_s(\cdot) \) is then computed as follows:

\[ FS(f_s) = \frac{1}{|D|} \sum_{i=1}^{N} \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} I[GS(S_m|D_i) > GS(\hat{S}_n|D_i)] \]

where \( |D| \) is the number of dialogues.

4.3 Evaluation Preparation

A series of models need to be prepared with different faithfulness capabilities to evaluate the effectiveness of model-level evaluation methods. One option is to collect as many well-trained models as possible and refer to human annotations to rank models based on factuality. However, it is hard to reach a high agreement and may not be trustworthy with limited annotators, as indicated by Falke et al. (2019) and Kryscinski et al. (2020). Therefore, instead, we construct a series of models using the following two ad-hoc methods:

**Limited data training (LDT).** One joint agreement is that more training data lead to better model performance. Therefore, we train 20 models using different proportions of the training data from 5% to 100%.

<table>
<thead>
<tr>
<th># Diag</th>
<th># Spk</th>
<th># Turn</th>
<th>Sum. Len.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>14,732</td>
<td>2.40</td>
<td>11.17</td>
</tr>
<tr>
<td>Val</td>
<td>818</td>
<td>2.39</td>
<td>10.83</td>
</tr>
<tr>
<td>Test</td>
<td>819</td>
<td>2.36</td>
<td>11.25</td>
</tr>
</tbody>
</table>

Table 3: The detailed statistics of the SAMSum dataset. The header refers to the number of dialogues, the average number of speakers, the average number of dialogue turns, and the average summary lengths.

**Mixed data training (MDT).** In this setting, we randomly replace the human-labelled training samples with noisy ones. The noisy samples are created by corrupting only the dialogue using transformations introduced in Section 4.1. Here, the source and target are both the dialogue. The trained model is more likely to be confused and generate more factuality errors with noisy data. Here, we obtain 21 models with different replacement ratios from 100% to 0%.

LDT will cause a model to be less competent for generation in all aspects. In comparison, MDT will lead the model to generate summaries with more factuality errors while less affecting other properties like fluency. Therefore, we expect a better factuality evaluator to correlate more with MDT models. All correlations are computed on model-level instead of sample-level judgements.

5 Experiments

5.1 Experimental Settings

SAMSum dataset (Gliwa et al., 2019) is used for all experiments. It consists of 16,369 dialogue-summary pairs written by expert linguistics. One human-written reference summary is provided for each dialogue. The detailed dataset statistics are listed in Table 3. The samples from the test set are used for all evaluation methods.

For backbone models, we exam with BART_large, BART_base (Lewis et al., 2020), T5_base and T5_small (Raffel et al., 2020), which are SOTA summarization models. Each model is trained with both LDT and MDT methods. As a result, we obtained 164 trained models, divided into eight groups. The models in each group are associated with increasing levels of capabilities. The Spearman’s rank correlation coefficient (\( \rho \)) between these models and evaluation scores is reported. For sample-based evaluation methods, the scores on all test set samples are averaged as the model-level performance. We ensure
all models are appropriately trained and avoid training collapses by examining their ROUGE scores. The best hyper-parameters are used and kept the same for models from the same group.

5.2 Results and Analysis

Table 4 shows the fine-grained results of FacEval. First, we found that FacEval has a higher correlation with MDT models than LDT models. The LDT models are less competent in all aspects as fewer data are involved with training. The generated summaries are weaker in multiple elements, including factuality, fluency, coherence, and granularity. In contrast, the MDT models mainly deteriorate in factuality with factually corrupted training data. Therefore, it is desired that FacEval shows a higher correlation to MDT models.

Second, when considering each negative sample type, a relatively higher correlation is shown with negation (NG), pronoun swap (PS) and speaker swap (SS). It is because more comparison pairs are created with these methods. Also, for chit-chat dialogues, almost all summaries contain reasoning concerning speakers and personnel in the dialogue. And the confirmation of action is happening in multiple utterances. As a result, these several error types are more commonly witnessed in dialogue summarization, as illustrated in Figure 1. In contrast, the negative pairs generated by entity swap (ES), date swap (DS) and number swap (NS) show a lower correlation. It is because these samples are more related to particular errors which appear in various formats and are more challenging to simulate. Even though solely considering these samples shows a lower correlation, we still include them in the overall comparison process to have a more comprehensive evaluation.

5.3 Comparison with Other Metrics

We include a list of popular evaluation methods for summarization to compare our evaluation schema with existing ones. It contains three generic evaluation methods and four dedicated faithfulness evaluation methods.

5.3.1 Baseline Metrics

Three generic evaluation methods are as follows: ROUGE (Lin, 2004) score is the default evaluation metric for summarization. We experiment with the F-measure of ROUGE-1, ROUGE-2 and ROUGE-L, which are derived from the uni-gram overlap, the bi-grams overlap and the longest common subsequence (LCS) between generated and reference summaries, respectively.

BLEU (Papineni et al., 2002) score is the primary evaluation metric for machine translation. It is mainly designed for corpus-level similarity computation derived from n-gram overlaps. In the following experiments, we report the most commonly used BLEU-4 score.

BERTScore (Zhang* et al., 2020) leverages the pre-trained contextual embeddings from BERT and computes the similarity between text sequences by matching words in candidate and reference by cosine similarity.

Four faithful evaluation methods are as follows: FactCC v1 (Kryscinski et al., 2020) first augment summaries by applying rule-based transformations

### Table 4: Detailed correlation analysis between model series and negative sample types. For each column, one negative type is involved. ‘all’ indicates the usage of all negative types.

<table>
<thead>
<tr>
<th>Train. Strategy</th>
<th>Model</th>
<th>NG</th>
<th>PS</th>
<th>SS</th>
<th>ES</th>
<th>DS</th>
<th>NS</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDT</td>
<td>BART_Large</td>
<td>79.79</td>
<td>31.28</td>
<td>87.82</td>
<td>-18.65</td>
<td>77.89</td>
<td>-2.41</td>
<td>84.70</td>
</tr>
<tr>
<td></td>
<td>BART_Base</td>
<td>89.62</td>
<td>75.64</td>
<td>98.65</td>
<td>73.83</td>
<td>40.84</td>
<td>-2.26</td>
<td>99.40</td>
</tr>
<tr>
<td></td>
<td>T5_Base</td>
<td>63.16</td>
<td>46.62</td>
<td>93.98</td>
<td>50.68</td>
<td>73.38</td>
<td>7.98</td>
<td>89.62</td>
</tr>
<tr>
<td></td>
<td>T5_Small</td>
<td>96.54</td>
<td>90.53</td>
<td>97.29</td>
<td>91.13</td>
<td>87.52</td>
<td>43.98</td>
<td>98.05</td>
</tr>
<tr>
<td>MDT</td>
<td>BART_Large</td>
<td>64.03</td>
<td>53.12</td>
<td>99.87</td>
<td>-21.30</td>
<td>35.84</td>
<td>-99.74</td>
<td>99.74</td>
</tr>
<tr>
<td></td>
<td>BART_Base</td>
<td>40.91</td>
<td>95.97</td>
<td>100.0</td>
<td>77.53</td>
<td>0.78</td>
<td>22.18</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td>T5_Base</td>
<td>79.22</td>
<td>90.91</td>
<td>99.87</td>
<td>92.73</td>
<td>-4.29</td>
<td>43.34</td>
<td>99.87</td>
</tr>
<tr>
<td></td>
<td>T5_Small</td>
<td>89.74</td>
<td>88.31</td>
<td>100.0</td>
<td>90.78</td>
<td>-12.27</td>
<td>-38.26</td>
<td>99.74</td>
</tr>
<tr>
<td># Neg Samples</td>
<td></td>
<td>3,094</td>
<td>2,990</td>
<td>2,100</td>
<td>643</td>
<td>547</td>
<td>98</td>
<td>9,472</td>
</tr>
</tbody>
</table>
Table 5: Comparison of a series of automatic evaluation metrics. The result shown is Spearman’s rank correlation between model ranks and predicted scores.

<table>
<thead>
<tr>
<th>Non-Factual Evaluation Schema</th>
<th>BART Large</th>
<th>BART Base</th>
<th>T5 Base</th>
<th>T5 Small</th>
<th>BART Large</th>
<th>BART Base</th>
<th>T5 Base</th>
<th>T5 Small</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1</td>
<td>81.35</td>
<td>95.79</td>
<td>94.44</td>
<td>95.94</td>
<td>84.16</td>
<td>91.04</td>
<td>95.58</td>
<td>85.84</td>
<td>90.52</td>
</tr>
<tr>
<td>ROUGE-2</td>
<td>86.77</td>
<td>96.84</td>
<td>96.39</td>
<td>96.09</td>
<td>90.13</td>
<td>97.01</td>
<td>95.45</td>
<td>93.90</td>
<td>94.07</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>75.64</td>
<td>96.24</td>
<td>96.39</td>
<td>92.63</td>
<td>86.23</td>
<td>98.31</td>
<td>97.14</td>
<td>94.16</td>
<td>92.09</td>
</tr>
<tr>
<td>BLEU</td>
<td><strong>91.88</strong></td>
<td>90.08</td>
<td>92.33</td>
<td>86.02</td>
<td>89.87</td>
<td>94.16</td>
<td>93.38</td>
<td>84.03</td>
<td>90.22</td>
</tr>
<tr>
<td>BERTScore</td>
<td>88.87</td>
<td>97.14</td>
<td><strong>97.29</strong></td>
<td>95.79</td>
<td>91.69</td>
<td>94.42</td>
<td>95.97</td>
<td>92.47</td>
<td>94.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factual Evaluation Schema</th>
<th>Factual Evaluation Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>FactCC_v1</td>
<td>--</td>
</tr>
<tr>
<td>FactCC_v2</td>
<td>6.02</td>
</tr>
<tr>
<td>FEQA</td>
<td>39.40</td>
</tr>
<tr>
<td>NLI</td>
<td>83.70</td>
</tr>
</tbody>
</table>

**FacEval** (ours) from the document sentences and fine-tune a pre-trained language model BERT to classify whether the summary is consistent or inconsistent with the documents. It is initially trained in the news summarization domain.

**FactCC_v2** is an adapted FactCC_v1 to the dialogue domain by us. The negative summaries are generated using our transformations discussed in Section 4.1. We train a T5_small model as the classifier and take dialogue and summary as input to predict their consistency.

**FEQA** (Durmus et al., 2020) is a question generation and answering method for faithfulness evaluation. It first extracts question-answer pairs from summaries with pre-trained models. Then, a QA model pulls answers from the document with the same questions. A matching method is deployed to measure the similarity between both answer responses from the summary and document as the factuality score. Note that the model is designed for documents.

**NLI** (Falke et al., 2019) is an entailment-based method which takes the maximal entailment probability between summary and document sentence as the factual consistency score. As no dialogue-based entailment model is available, we compute the entailment probability between reference and generated summaries with a BERT-based entailment model trained on SNLI and MultiNLI datasets.

5.3.2 Results and Analysis
The experimental results are shown in Table 5.

**Non-factual Evaluator:** The non-factual evaluation methods measure the similarity between reference and generated summaries. ROUGE and BLEU are derived from n-gram overlaps, which indicate the overall generation quality. It is expected that evaluators have a reasonable correlation with LDT models as training with fewer data will result in quality degradation of the summary in all aspects. For MDT models, they also show a good correlation. We observe that R-2 and R-L are better indicators than R-1 for factuality evaluation. It is because simply replacing isolated tokens can easily change the factual correctness of a summary without much influence on the R-1 score.

**Factual Evaluator:** As FactCC_v1 is trained for the news summarization, we found that the released model is incapable of making predictions for dialogues. Similarly, FEQA is not a good indicator of model performance because the question and answer generation models are optimized for documents, which limits its transferability to the dialogue domain. In comparison, FactCC_v2 and NLI are better evaluation methods for factuality and can make good predictions on MDT models.

**FacEval Properties:** FacEval is the only model-level evaluation schema. The examined model requires reasonable predictions on single sentences and differentiation between positive and negative pairs. Therefore, FacEval shows a strong correlation with LDT and MDT models. The exceptional performance on MDT models indicates that FacEval can effectively reflect model’s capability on factuality.
<table>
<thead>
<tr>
<th>Model</th>
<th># Params</th>
<th>All</th>
<th>NG</th>
<th>PS</th>
<th>SS</th>
<th>ES</th>
<th>DS</th>
<th>NS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART&lt;sub&gt;Large&lt;/sub&gt;</td>
<td>400M</td>
<td>86.87</td>
<td>89.03</td>
<td>89.08</td>
<td>84.19</td>
<td>89.46</td>
<td>84.22</td>
<td>68.37</td>
</tr>
<tr>
<td>BART&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>140M</td>
<td>85.73</td>
<td>87.73</td>
<td>89.43</td>
<td>79.46</td>
<td>89.96</td>
<td>82.50</td>
<td>68.63</td>
</tr>
<tr>
<td>T5&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>220M</td>
<td>85.32</td>
<td>86.81</td>
<td>88.27</td>
<td>79.92</td>
<td>90.71</td>
<td>84.90</td>
<td>67.75</td>
</tr>
<tr>
<td>T5&lt;sub&gt;Small&lt;/sub&gt;</td>
<td>60M</td>
<td>79.87</td>
<td>82.09</td>
<td>85.94</td>
<td>69.68</td>
<td>79.88</td>
<td>79.17</td>
<td>62.96</td>
</tr>
</tbody>
</table>

Table 6: Benchmarking results on 4 different models

### 5.3.3 Benchmarking and Future Directions

It is beneficial to provide benchmarking performance and analysis on popular dialogue summarization models. As discussed in Sec. 3, dedicated dialogue summarization models do not outperform their baseline models in terms of faithfulness. Therefore, we evaluate on T5 and BART models instead.

The benchmarking results are shown in Tab 6. There are several interesting findings. First, BART<sub>Large</sub> has the largest model size as well as the overall best performance. We can also conclude that larger pre-trained models are more faithful based on our evaluation. Second, BART model is generally better than T5 in factuality with model size taken into consideration. This may be because that BART is designed for the generation with various denoising objectives, while T5 is a sequence-to-sequence model for different tasks including but not limited to generation. Third, from fine-grained analysis, we can see that speaker information (from SS) is a major challenge for dialogue summarization. This is because dialogue involves multiple speakers and their roles are tightly involved in the ideal summarization. Therefore, how to improve the model’s understanding capability on speaker roles is an interesting direction to explore (Liu et al., 2021b). Meantime, because some faithful errors are coming from lack of commonsense for existing models (Wang et al., 2021). How to effectively combine hidden semantics (Wang and Kuo, 2020) and well-structured knowledge (Ge et al., 2022) are also worth exploration.

### 6 Conclusion

We believe our faithfulness analysis and evaluation method can facilitate the development of dialogue summarization systems. Instead of measuring faithfulness on generated summaries, we directly assess the model’s capability by multi-choice questions. We expect FacEval to be effectively extended to other generation scenarios.

### 7 Limitations

The testing samples used in our method are obtained by rule-based transformations of the reference and back-translated summaries. It is still limited to the types of transformations designed. More transformation methods need to be proposed to have a comprehensive evaluation. To obtain more natural summaries, we can gather generated summaries and perform annotation by humans. The model can be evaluated in more aspects and closer to real-world scenarios with more available samples.

Verifying the effectiveness of the model-level evaluation schema requires various models and their corresponding rankings. However, such model rankings are currently unavailable because 1) there are not enough varieties of dialogue summarization models as it is still a developing field; 2) the annotations on the faithfulness of dialogue summaries are not adequate. Therefore, in this work, we refer to heuristic methods to manually create a series of models with desired capability levels. When new evaluators are proposed, the best practice is to leverage model-level human rankings for performance benchmarking.

### Acknowledgement

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A2b0046) and Science and Engineering Research Council, Agency of Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 (Grant No. 192 25 00054). This work is also supported by the Internal Project Fund from Shenzhen Research Institute of Big Data under Grant T00120220002.
References


ZiQiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In thirty-second AAAI conference on artificial intelligence.


