# DialSummEval: Revisiting Summarization Evaluation for Dialogues

Anonymous ACL submission

## Abstract

Dialogue summarization is receiving increasing attention from researchers due to its extraordinary difficulty and unique application value. We observe that current dialogue summarization models have flaws that may not be well exposed by frequently used metrics such as ROUGE. In our paper, we re-evaluate 18 categories of metrics in terms of four dimensions: coherence, consistency, fluency and relevance, as well as a unified human evaluation of various models for the first time. Some noteworthy trends which are different from the conventional summarization tasks are identified. We will release DialSummEval, a multi-faceted dataset of human judgments containing the outputs of 14 models on SAMSum.

## 1 Introduction

Neural network based approaches and sizable datasets have led to significant progress in researches towards conventional summarization tasks such as news and scientific papers (Lin and Ng, 2019). Compared with conventional summarization tasks, dialogue summarization has received increasing attention from researchers due to its great difficulty and unique application value (Feng et al., 2021a). With the proposal of dialogue summary datasets such as SAMSum (Gliwa et al., 2019), DialogSum (Chen et al., 2021) and MediaSum (Zhu et al., 2021), a number of models for automatic generation of dialogue summaries have emerged (Feng et al., 2021b; Liu and Chen, 2021; Zou et al., 2021; Qi et al., 2021; Chen and Yang, 2020; Chen and Yang, 2021; Zhao et al., 2020; Liu et al., 2021).

There is no denying that these studies have made promising progress, but it remains a challenge to evaluate these advances comprehensively. Current studies generally use the SAMSum dataset and adopt ROUGE (Lin, 2004), an n-gram-based automatic evaluation metric using reference summaries, as the overall evaluation criterion for summary quality, complemented by manual evaluation. Schluter (2017) and Graham (2015) illustrate the limitations of ROUGE in evaluating summarization tasks. Also the manual evaluation protocols vary from one research to another based on our observations.

We argue that the inadequate evaluation mechanism may have become a major obstacle to the progress of dialogue summarization researches. Many studies, such as Chen and Yang (2020) and Tang et al. (2021), have pointed out that the current dialogue summarization models still have many shortcomings, such as wrong references, incorrect reasoning and improper gender pronouns, and ROUGE may not reflect these problems effectively. For example, Gabriel et al. (2021) note that ROUGE-1 and ROUGE-L fail to accurately measure factual inconsistency across domains. Our case study in Table 1 also illustrates this point. However, it is impractical to perform frequent time-consuming and costly manual evaluation. The alternative is to introduce or propose more reliable automatic evaluation metrics to evaluate the models in a more comprehensive and fine-grained manner.

Although there are automatic evaluation metrics for measuring the quality of all aspects of summaries on conventional summarization tasks, especially for factual consistency (Huang et al., 2021), it is difficult to guarantee that they will still perform well on dialogue summarizarion. Recently proposed automatic metrics for evaluating generic natural language generation tasks such as BERTScore (Zhang* et al., 2020), BARTScore (Yuan et al., 2021) have also not been experimented on dialogue summarization. The high abstraction level, low extraction rate, and the requirement for complex reasoning power of the dialogue summarization task present new challenges to automatic evaluation metrics. There

| Dialogue | Reference Summary | Generated Summary | R-1 | R-2 | R-L |
|---|---|---|---|---|---|
| **Kirsten**: Youth group this Friday, don't be late. **Alex**: What time? **Kirsten**: 7 pm. We're going bowling, so we'll meet up and then all go together. **Alex**: Cool. See you. **Kirsten**: Bye | Kirsten reminds Alex that the youth group meets this friday at 7 pm and go bowling. | Kirsten is going bowling with her youth group this Friday at 7 pm. | 0.69 | 0.44 | 0.61 |
| **Ola**: Hey running late **Ola**: I should be free by 8 **Kurt**: Sure no prob, call me | Ola should be free by 8. Kurt wants her to call him. | Ola will be late. She should be free by 8. Kurt will call her. | 0.69 | 0.42 | 0.67 |

Table 1: Case study of some outputs of BART on SAMSum. The ROUGE values of these outputs have substantially exceeded the state of the art on SAMSum. The summary in the first row fails in relevance, and the second has a factual error.

have been a number of manual evaluation datasets and analytical studies for conventional summarization tasks ((Dang and Owczarzak, 2008); Fabbri et al., 2021b; Bhandari et al., 2020), but very little work has been done on systematic analysis of dialogue summarization models and evaluation metrics. Our work will fill the gap in this area and includes the following contributions: 1) We identify evaluation problems in the field of dialogue summarization and point out the urgent need of automatic evaluation metrics that better adapt to dialogue summarization. 2) We collect and provide a sizable, multi-faceted dataset of manual evaluations for dialogue summarization, which contains the output of 14 models, and the dataset will be released. 3) We re-evaluate the performance of 18 types of automatic evaluation metrics on dialogue summarization. 4) We evaluate a variety of dialogue summarization models (extractive, abstractive, and recently based on pre-trained language models) in a unified manner.

## 2 Related Work

**Meta-Evaluation with Human Judgments** Automatic evaluation Metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020)) were compared with other metrics when proposed. However, they are basically not using the dialogue summarization dataset as an experimental corpus, and rarely provide new human judgments data. Bhandari et al. (2020) used *pyramid* (Nenkova and Passonneau, 2004), a widely used human evaluation method on several conventional summarization datasets to obtain relevance scores for some of the system outputs and re-evaluated the metrics in 6 categories. Similarly, Fabbri et al. (2021b) used CNN/DailyMail dataset (Hermann et al., 2015) and the output of some models for human evaluation covering four facets of relevance, consistency, fluency, and coherence, and then re-evaluated the metrics in 14 categories. None of these involved dialogue summarization datasets. Gabriel et al. (2021) is one of the few current studies using the dialogue summarization dataset SAMSum (Gliwa et al., 2019) for meta-evaluation, but it focuses on factual consistency and selects a small number of metrics.

**Evaluation for Dialogue Summarization Models** Tang et al. (2021) and Chen and Yang (2020) sample the output of models on SAMSum and analyzes the error types when proposing a new model. Due to the different manual evaluation protocols and the small number of models included, it is difficult to comprehensively compare the strengths and weaknesses of different models.

## 3 Preliminaries

In this section, we introduce the involved dataset, metrics and models.

### 3.1 Dataset

**SAMSum** (Gliwa et al., 2019) is the first manually annotated, high-quality chat summarization dataset, containing over 16k dialogues. We use it in this study as it is most widely used and has

greatly promoted the research in the field of dialogue summarization, and we are able to collect the outputs of various models on this dataset.

### 3.2 Evaluation Metrics

We selected a number of evaluation metrics that are frequently used on summarization or other natural language generation tasks. Some are for overall quality; others are specific to a particular aspect. Some require reference summaries or source documents; some only need the summary itself. Here is a brief categorization and description.

Metrics based on n-gram overlap include:

**ROUGE** (Lin, 2004) is the most widely used automatic evaluation metric in summarization. Researchers mainly adopt ROUGE-1, ROUGE-2 and ROUGE-L, which measure the unigram-overlap, bigram-overlap and longest common sequence between two texts respectively. [1]

**BLEU** (Papineni et al., 2002) is the primary evaluation metric for machine translation. It calculates n-gram overlap between texts using precision scores and includes a brevity penalty. [2]

**METEOR** (Banerjee and Lavie, 2005) computes an alignment by mapping unigrams in two texts, based on surface forms, stemmed forms, and meanings.

**CHRF** (Popović, 2015) computes character based n-gram overlap between two texts. [3]

Metrics based on pre-trained language models include:

**BERTScore** (Zhang* et al., 2020) measures the soft-overlap between two texts at token level using contextual embeddings from BERT. [4]

**MoverScore** (Zhao et al., 2019) applies the semantic distance between two texts at n-gram level using n-gram embeddings pooled from BERT. [5]

**BARTScore** (Yuan et al., 2021) assumes that models trained to convert the generated text to/from a reference output or the source text will achieve higher scores when the generated text is better. It can be flexibly applied to evaluation of text from different perspectives using BART. [6]

**BLANC** (Vasilyev et al., 2020) is a reference-less metric. It hypothesizes that a good summary is beneficial for a pre-trained language model to conduct language understanding tasks on the source document. Specifically, it measures the performance boost of BERT by utilizing the summary in two different ways. [7]

**PPL**, namely perplexity, is often used to evaluate the quality of a language model or the fluency of an utterance. We adopt GPT-2 (Radford et al., 2019) as the language model for computing the perplexity for the whole summary. [8]

Metrics based on word embeddings include:

**SMS** (Clark et al., 2019), namely Sentence Mover Similarity, extends Word Movers Distance (Kusner et al., 2015) to measure the distance between two texts which are represented as a bag of sentence embeddings. [9]

**Embedding average** (Sharma et al., 2017) is an embedding based metric computing the cosine similarity between the embeddings of two texts. A sentence-level embedding is represented by averaging the embeddings of the words composing the sentence.

**Vector extrema** (Forgues et al., 2014) is also an embedding based metric similar to Embedding average. The metric computes a sentence-level embedding by taking the most extreme value of the embeddings of the words composing the sentence for each dimension of the embedding.

**Greedy matching** (Rus and Lintean, 2012) is another embedding based metric. The metric does not compute a sentence-level embedding. It directly compares the embeddings of words in the two sentences using a greedy matching algorithm to calculate similarity.

Metrics based on question-answering include:

**FEQA** (Durmus et al., 2020) employs a BERT-based question-answering model to answer questions using source document. Questions are generated by a fine-tuned BART model using generated summaries with masked named entities as inputs. The metric reports F1 scores against the gold answer, which are often regarded as a measure of factual consistency. [10]

[1] https://github.com/Diego999/py-rouge

[2] Used code at https://github.com/Maluuba/nlg-eval, the same for Embedding average, Vector extrema, Greedy matching and METEOR

[3] https://github.com/m-popovic/chrF

[4] https://github.com/Tiiiger/bert_score

[5] https://github.com/AIPHES/emnlp19-moverscore

[6] https://github.com/neulab/BARTScore

[7] https://github.com/PrimerAI/blanc

[8] https://huggingface.co/docs/transformers/perplexity

[9] https://github.com/eaclark07/sms

[10] https://github.com/esdurmus/feqa

**SummaQA** (Scialom et al., 2019) is also a QA-based metric. Unlike FEQA, it generates questions from source documents instead of summaries to be evaluated and then uses summaries to answer them. The F1 overlap score and QA-model confidence are reported. [11]

**QuestEval** (Scialom et al., 2021) is another a QA-based metric. This metric can be considered as a combination of FEQA and SummaQA. It takes into account the scores obtained from both styles. For comparison purposes, We use the reference-less mode.[12]

Metrics based on entailment classification include:

**FactCC** (Kryscinski et al., 2020) is an entailment classification metric. We follow the way Pagnoni et al. (2021) used it. Each sentence of the summary is fed into the classifier together with the document to determine whether the facts are consistent, and the proportion of consistent sentences is used to indicate how consistent the summary is. [13]

**DAE** (Goyal and Durrett, 2020; Goyal and Durrett, 2021) is an entailment classification metric based on dependencies. We use it in a similar way to FactCC. When a sentence cannot be parsed by the metric, we default it factually inconsistent. [14]

### 3.3 Summarization Models

We select some representative models and get the outputs of them on the test set of SAMSum. We choose LEAD-3 and LONGEST-3 as representatives of the simple extractive approaches. PGN (See et al., 2017) and Transformer (Vaswani et al., 2017) are selected as representatives of the earlier neural summarization models. For generic pre-trained generative models, we use BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020) and UniLM (Dong et al., 2019). We retrain these models above to obtain the outputs and the automatic evaluation results are close to Gliwa et al. (2019) and Wu et al. (2021) in default settings. For models specifically designed for dialogue summarization, we choose CODS (Wu et al., 2021), ConvoSumm (Fabbri et al., 2021a), MV-BART (Chen and Yang, 2020), PLM-BART (Feng et al., 2021c), Ctrl-DiaSumm (Liu and Chen, 2021), S-BART (Chen and Yang, 2021) and the outputs are all provided by their authors. We also regard the reference summary as a kind of model output.

## 4 Data Annotation

### 4.1 Annotation Setup

Since human evaluation is expensive and time-consuming, we decide to randomly sample 100 dialogues from the test set of SAMSum and evaluate the summaries generated by all models on these dialogues. To comprehensively evaluate each metric and model, we perform human evaluation in four aspects, as in Kryscinski et al. (2019):

**Coherence** measures the quality of all sentences in the summary as a whole. It focuses on whether the summary is coherent and natural.

**Consistency** measures how well the summary aligns with the dialogue in facts. It focuses on whether the summary contains factual errors.

**Fluency** measures the quality of individual sentences in the summary compared to Coherence. It focuses on whether the sentences are well-written and grammatically correct.

**Relevance** measures how well the summary captures the key points of the dialogue. It focuses on whether all and only the important aspects are contained in the summary.

To ensure the quality of the annotation, we tried to annotate some of the data ourselves at the beginning to judge the difficulty of the task and the approximate time spent.

### 4.2 Annotation Process

We initially tried to annotate the data using crowdsourcing platforms. We published the annotation task on Amazon Mechanical Turk [15]. The interface contained instructions and definitions of the four aspects. A dialogue and a corresponding summary were included in the interface, and the summaries of different models on the same dialogue were presented to the annotators in a sequence to facilitate comparison. For each dimension/aspect, annotators were asked to rate the summary on a Likert scale from 1 to 5. Each summary was evaluated by 5 different annotators, and For each dimension we would receive a total of $100 \times 14 \times 5 = 7000$ human annotations. The annotation was done quickly in one day, but the qual-

[11] https://github.com/ThomasScialom/summa-qa

[12] https://github.com/ThomasScialom/QuestEval

[13] https://github.com/salesforce/factCC

[14] https://github.com/tagoyal/factuality-datasets

[15] https://www.mturk.com

ity was not satisfactory. We calculated the average score of each model in each aspect based on these annotation data and found that the scores of the models are close in each dimension, which is not in the accordance with the reality. For example, in terms of consistency, the reference summary and the extractive approaches should have had a definite advantage, but this failed to be reflected from the data. The result is shown in Table 5. For reliability reasons, we do not use these annotations for our analysis.

Then, we decided to recruit annotators from the school forum who are required to be capable of reading daily conversations and articles in English fluently. We recruited three annotators, using a similar annotation interface and approach as in the crowd-sourcing platforms. These annotators were college students and they are fluent in English. The differences with the crowd-sourcing platform annotation are as follows: 1) For a student who wanted to participate in the annotation, we would ask him to annotate all models on the first 10 conversations ($10 \times 14 = 140$ annotations), and let her/him continue the annotation only when these annotation results were checked by us to confirm that the annotator had understood the task correctly and could finish the annotation responsibly. Otherwise, we paid the annotator directly for this part and terminated his annotation task. 2) We required each annotator to annotate all data ($100 \times 14 = 1400$ annotations) to ensure the consistency within the annotator. 3) During the annotation process, we kept in touch with the annotators via email or instant messaging app to answer their questions at any time.

It took around 10 days to finish the annotation. We received $100 \times 14 \times 3 = 4200$ annotations for each perspective. For each aspect of each summary, if two scores were the same and the other was different from them, we considered the different one as noise. For each dimension, we removed the noise separately and calculated the the Krippendorffs alpha coefficient (Krippendorff, 2011). We found the inter-annotator interval kappa to be within an acceptable range - from 0.5621 to 0.7564, as detailed in Table 2. The raw annotated data will be released and we use the cleaned data for analysis. At last, we use the average of the remaining data to represent the human evaluation score of an summary on a dimension.

# 5 Metric Evaluation

In this section, we will introduce several definitions in meta-evaluation and re-evaluate the metrics mentioned in Section 3.2.

## 5.1 Task Formulation

As mentioned by Bhandari et al. (2020), there are two common ways to measure the correlation of automatic evaluation metrics to manual evaluation: system-level and summary-level.

Assuming there are $N$ dialogues, the $i$-th dialogue is represented as $d_i$. For a dialogue $d_i$, there are $J$ summaries generated by $J$ models, and we denote each of them as $s_{ij}, j = 1 \cdots J$. There are $K$ evaluation metrics in total, and $m_k$ refers to an automatic evaluation metric or human evaluation of a certain dimension. $m_k(s_{ij})$ means the score of $k$-th metric towards a pair of dialogue and summary $(d_i, s_{ji})$. We use $R(m_i, m_j)$ to denote the correlation coefficient between two metrics $m_i$ and $m_j$.

System-level correlation is defined as follows. The corresponding p-value which indicates statistical significance can be obtained:

$$R_{sys}(m_p, m_q) = R( \quad (401)$$

$$[\frac{1}{N}\sum_{i=1}^{N} m_p(s_{i1}), \cdots, \frac{1}{N}\sum_{i=1}^{N} m_p(s_{iJ})], \quad (402)$$

$$[\frac{1}{N}\sum_{i=1}^{N} m_q(s_{i1}), \cdots, \frac{1}{N}\sum_{i=1}^{N} m_q(s_{iJ})]) \quad (403)$$

(404)

Summary-level correlation is defined as follows, and the p-value cannot be derived here because the Summary-level correlation is an average value:

$$R_{sum}(m_p, m_q) = \frac{1}{N}\sum_{i=1}^{N} R([m_p(s_{i1}), \cdots, m_p(s_{iJ})], \quad (408)$$

$$[m_q(s_{i1}), \cdots, m_q(s_{iJ})]) \quad (409)$$

(410)

## 5.2 Discussion

Comparing the performance of various metrics reveals some trends in Table 3. In each dimension, metrics which are strongly correlated with human judgments exist, but few metrics show significant

| | Coherence | Consistency | Fluency | Relevance |
|---|---|---|---|---|
| **remaining** | 3161 | 3360 | 3050 | 3439 |
| **total** | 4200 | 4200 | 4200 | 4200 |
| **kappa** | 0.7564 | 0.6709 | 0.6782 | 0.5621 |

Table 2: The inter-annotator agreement for each dimension.

strengths in all four dimensions. Of all the metrics, QuestEval has the most comprehensive capabilities at the system level. Generally metrics that perform better on coherence and fluency perform worse on consistency and relevance, and vice versa. This can be attributed to the definition of the dimensions, i.e. there is some correlation between the four dimensions themselves. In all dimensions, automatic evaluation metrics based on pre-trained language models generally outperform metrics based on n-gram overlap and context-independent word embedding. Among them, the recently proposed BARTScore and the increasingly popular QA-based metrics perform the best. This suggests that both directions have the potential to be explored in terms of evaluation for dialogue summarization. Across dimensions, almost all metrics correlate better with human judgments at the system level than at the summary level, and both showed good agreement with each other. This indicates that the summary-level correlations are also worth referring to when enough data are not available for system-level analysis. In addition, metrics such as BLEU and CHRF, which are frequently used in other natural language generation tasks (e.g., machine translation, dialogue, etc.), do not show advantages on dialogue summarization.

The characteristics presented by the automatic evaluation metrics on the dialogue summarization differ from those of the conventional summarization tasks. For ROUGE, we find that increasing the size of n in ROUGE-n is not better in almost all dimensions, which is different from the findings of Rankel et al. (2013) and Fabbri et al. (2021b). The ability of ROUGE to reflect content selection, i.e., relevance, as we usually believe, is also questionable. Compared to the results of Fabbri et al. (2021b), metrics based on n-gram overlap such as ROUGE and CHRF perform worse on dialogue summarization, while some metrics that use source documents such as BLANC perform better. We need to focus on the limitations of ROUGE and the role of the source dialogues in evaluating dialogue summaries.

We have also observed some interesting phenomena. Entailment classification metrics such as FactCC and DAE outperform many metrics in terms of consistency, but not as well as BARTScore and QA-based metrics. This may be due to the large gap between the corpus used in training and dialogues, and the need to slice the summaries by sentence when using them. FEQA, which is designed for factual consistency, however, performs best in coherence and fluency, and rather poorly in consistency and relevance. Comparing its performance with QuestEval and SummaQA, generating questions from the original dialogue may be more reliable in measuring consistency, which corroborates with the points of Gabriel et al. (2021). It is surprising that metrics based on the language model such as PPL, BARTScore-h performs poorly in measuring both coherence and fluency. The exact reasons for this need further investigation.

## 6 Model Evaluation

In each dimension, we evaluate each model mentioned in Section 3.3 using the average of the human evaluation scores of all summaries. Analyzing Table 4, we conclude the following.

The reference summaries in SAMSum are not perfect, and the annotators felt that they also contained some factual inconsistencies compared to the source dialogues, as well as important elements of the dialogues that were not all captured by them. However, comparing the human evaluation scores of the reference summaries in CNNDM (Fabbri et al., 2021b), the quality is already superior.

Extractive models produce summaries that differ in style from abstractive models, and many conversations contain ungrammatical utterances, which can affect the reading experience and impair their fluency and coherence. In particular, LONGEST-3, which extracts some potentially discontinuous sentences from dialogues, has low coherence. However, since they do not modify the content, they still perform well in terms of con-

| | Coherence | | Consistency | | Fluency | | Relevance | |
|---|---|---|---|---|---|---|---|---|
| **Metrics** | sys | sum | sys | sum | sys | sum | sys | sum |
| **ROUGE-1** | **0.59**$^*$ | **0.30** | 0.42 | 0.33 | **0.58**$^*$ | **0.27** | 0.40 | 0.30 |
| **ROUGE-2** | 0.47 | 0.26 | 0.41 | 0.32 | 0.43 | 0.22 | 0.41 | 0.30 |
| **ROUGE-3** | 0.39 | 0.22 | 0.39 | 0.30 | 0.33 | 0.17 | 0.40 | 0.30 |
| **ROUGE-4** | 0.33 | 0.20 | 0.37 | 0.27 | 0.27 | 0.14 | 0.38 | 0.28 |
| **ROUGE-L** | **0.57**$^*$ | **0.32** | 0.39 | 0.30 | **0.54**$^*$ | 0.27 | 0.37 | 0.27 |
| **BERTScore-p** | **0.57**$^*$ | **0.37** | 0.11 | 0.10 | 0.50 | **0.31** | 0.08 | 0.06 |
| **BERTScore-r** | 0.43 | 0.21 | 0.45 | 0.38 | 0.42 | 0.20 | 0.46 | 0.39 |
| **BERTScore-f1** | 0.53 | **0.31** | 0.28 | 0.24 | 0.48 | **0.27** | 0.27 | 0.22 |
| **MoverScore** | 0.50 | 0.28 | 0.39 | 0.32 | 0.46 | 0.25 | 0.38 | 0.31 |
| **SMS** | 0.33 | 0.18 | 0.38 | 0.28 | 0.27 | 0.14 | 0.40 | 0.29 |
| **BARTScore-s-h +** | 0.09 | 0.08 | **0.62**$^*$ | **0.44** | 0.24 | 0.15 | **0.60**$^*$ | **0.42** |
| **BARTScore-h -** | 0.08 | 0.05 | -0.09 | -0.09 | 0.16 | 0.13 | -0.18 | -0.12 |
| **BARTScore-h-r** | 0.50 | 0.21 | **0.55**$^*$ | **0.46** | 0.51 | 0.21 | 0.56$^*$ | **0.46** |
| **BARTScore-r-h** | **0.67**$^{**}$ | **0.42** | 0.31 | 0.23 | **0.67**$^{**}$ | **0.40** | 0.26 | 0.17 |
| **BLANC-help +** | -0.32 | -0.21 | 0.54 | **0.45** | -0.13 | -0.08 | **0.60**$^*$ | **0.50** |
| **BLANC-tune +** | -0.37 | -0.23 | 0.50 | 0.38 | -0.18 | -0.10 | 0.56$^*$ | **0.43** |
| **FEQA +** | **0.82**$^{**}$ | 0.27 | 0.32 | 0.16 | **0.84**$^{**}$ | **0.26** | 0.25 | 0.10 |
| **QuestEval +** | 0.50 | 0.15 | **0.85**$^{**}$ | **0.39** | **0.75**$^{**}$ | 0.20 | **0.83**$^{**}$ | 0.37 |
| **SummaQA-conf +** | -0.08 | -0.03 | **0.64**$^*$ | **0.39** | 0.03 | -0.01 | **0.67**$^{**}$ | **0.39** |
| **SummaQA-fscore +** | -0.26 | -0.11 | **0.58**$^*$ | 0.26 | -0.06 | -0.06 | **0.62**$^*$ | 0.29 |
| **PPL -** | -0.13 | -0.01 | -0.49 | -0.30 | -0.34 | -0.15 | -0.43 | -0.30 |
| **CHRF** | 0.42 | 0.20 | 0.46 | 0.38 | 0.41 | 0.20 | 0.47 | 0.39 |
| **BLEU-1** | 0.35 | 0.15 | 0.34 | 0.29 | 0.30 | 0.13 | 0.36 | 0.30 |
| **BLEU-2** | 0.31 | 0.16 | 0.35 | 0.29 | 0.25 | 0.12 | 0.37 | 0.30 |
| **BLEU-3** | 0.28 | 0.15 | 0.33 | 0.27 | 0.21 | 0.11 | 0.36 | 0.28 |
| **BLEU-4** | 0.25 | 0.14 | 0.33 | 0.25 | 0.17 | 0.09 | 0.36 | 0.28 |
| **METEOR** | 0.37 | 0.19 | 0.42 | 0.35 | 0.33 | 0.17 | 0.43 | 0.35 |
| **Embedding average** | 0.43 | 0.17 | 0.17 | 0.20 | 0.52 | 0.22 | 0.15 | 0.19 |
| **Vector extrema** | 0.47 | 0.22 | 0.35 | 0.28 | 0.43 | 0.21 | 0.35 | 0.26 |
| **Greedy matching** | 0.43 | 0.21 | 0.35 | 0.31 | 0.43 | 0.21 | 0.36 | 0.30 |
| **FactCC +** | -0.29 | -0.09 | 0.46 | 0.19 | -0.23 | -0.09 | 0.49 | 0.19 |
| **DAE +** | -0.24 | -0.07 | 0.50 | 0.29 | -0.15 | -0.02 | 0.54$^*$ | 0.29 |

Table 3: The correlation (Pearson's r) of annotations computed on system level and summary level along four quality dimensions between automatic metrics and human judgments. For evaluation, all metrics require at least the summaries to be evaluated as input. Metrics with + indicate that the source dialogues are used, metrics with - means no other input are required, others need to use the reference summaries. The five most-correlated metrics in each column are bolded (For system level, **=significant for $p \leq 0.01$, *=significant for $p \leq 0.05$). We add suffixes to distinguish the different variants of metrics. For BARTScore, h, r and s are abbreviations of hypothesises, references and source dialogues respectively. BARTScore-s-h measure the probability to generate hypothesises using source dialogues as inputs, while BARTScore-h measures the probability to generate hypothesises without other inputs, and so on. For BLANC, BLANC-tune refers to the way of fine-tuning on a generated summary and then conducting nature language understanding tasks on source dialogues, while BLANC-help refers to the way of inferring with a generated summary concatenated together. For SummaQA, SummaQA-fscore measures the average overlap between predictions and ground truth answers, and SummaQA-conf corresponds to the probability of the true answer.

| Models | Coherence | Consistency | Fluency | Relevance |
|---|---|---|---|---|
| reference summary | **4.500** | **4.370** | 4.560 | **4.210** |
| **LONGEST-3** | 3.230 | **4.393** | 4.100 | **4.363** |
| **LEAD-3** | 4.370 | 4.093 | 4.200 | 3.843 |
| **PGN** | 3.568 | 2.103 | 3.657 | 2.293 |
| **Tranformer** | 3.403 | 1.573 | 3.673 | 1.650 |
| **BART** | 4.480 | 3.667 | **4.667** | 3.500 |
| **PEGASUS** | **4.590** | 3.730 | 4.640 | 3.417 |
| **UniLM** | 4.303 | 3.320 | 4.523 | 3.290 |
| **CODS** | 4.268 | 3.637 | 4.567 | 3.397 |
| **ConvoSumm** | 4.507 | 3.743 | 4.643 | 3.437 |
| **MV-BART** | 4.320 | 3.937 | 4.660 | 3.747 |
| **PLM-BART** | 4.360 | 3.717 | **4.680** | 3.500 |
| **Ctrl-DiaSumm** | 4.320 | 3.893 | 4.650 | 3.670 |
| **S-BART** | 4.227 | 3.307 | 4.520 | 3.337 |

Table 4: Human ratings of summaries along four evaluation dimensions using cleaned annotations from campus recruitment. scores are averaged over three annotators, broken down by the approximate classification in Section 3.3. The two highest-rated models in each column are in bold.

sistency. Since the average length of dialogues in SAMSum is small, extracting a few sentences from it can generally include important contents, so the relevance is also high.

The early neural summrization models represented by PGN and Transformer perform relatively poorly in all dimensions compared to the reference summaries, especially consistency and relevance. This is to be expected because of the high difficulty of dialogue summarization and the small size of SAMSum dataset.

An important finding is that the generic pre-trained language models represented by BART, PEGASUS and UniLM, and various recently proposed models specifically designed on the dialogue summarization task do not have significant differences in each dimension. They are already comparable, and in some cases better, in terms of coherence and fluency compared to the reference summaries. They have improved dramatically compared to earlier neural summarization models with respect to consistency and relevance, but there is still some room for enhancement. On the one hand, this finding affirms the capability of these models; On the other hand, it urges us to reflect on how much these recently proposed complex models or fancy techniques are an improvement over the generic pre-trained language models.

## 7 Conclusion

We point out the problems with the evaluation in the dialogue summarization and introduce DialSummEval, a multi-faceted dataset containing the output of various models and the corresponding human judgments. Based on this dataset, we provide a comprehensive re-evaluation and analysis of the performance of widely used automatic evaluation metrics and each model. There are three important findings: 1) Few metrics are excellent in all dimensions, and the recently proposed BARTScore and QA-based metrics are comparatively outstanding and worth exploring. 2) The automatic evaluation metrics and their variants present some trends that differ from conventional summarization. 3) A variety of models specifically designed for dialogue summarization perform comparably to reference summaries in terms of coherence and fluency, but still have shortcomings in consistency and relevance. We hope that researchers in the field recognize the importance of evaluation in current research, choose some metrics other than ROUGE when evaluating models, propose automatic evaluation metrics that can be better adapted to the field of dialogue summarization based on our work.

## 8 Ethical Considerations

Whether recruiting annotators through Amazon Mechanical Turk or campus, we paid them 15 dollars per hour, more than the local average mini-

mum wage. We removed all content in the dataset that might contain personal information about the annotators.

## References


Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *TAC*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021a. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021a. A survey on dialogue summarization: Recent advances and new frontiers. *Computing Research Repository*, arXiv:2107.03175. Version 1.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021b. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3808–3814. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021c. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491, Online. Association for Computational Linguistics.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive

summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *Computing Research Repository*, arXiv:2104.14839. Version 2.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9815–9822.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*,

pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

MengNan Qi, Hao Liu, YuZhuo Fu, and Ting Liu. 2021. Improving abstractive dialogue summarization with hierarchical pretraining and topic segment. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1121–1130, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

Vasile Rus and Mihai Lintean. 2012. An optimal assessment of natural language student input using word-to-word similarity metrics. In *International Conference on Intelligent Tutoring Systems*, pages 675–676. Springer.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *Computing Research Repository*, arXiv:2112.08713. Version 1.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Computing Research Repository*, arXiv:2106.11520. Version 2.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong

Kong, China. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

# A Appendix

## Instructions

In this task you will evaluate the quality of summaries written for a dialogue from daily life.

To correctly solve this task, follow these steps:

1. Carefully read this dialogue, be aware of the information it contains.
2. Read the proposed summaries A-N (14 in total).
3. Rate each summary on a scale from **1** (worst) to **5** (best) by its *relevance, consistency, fluency, coherence*.

Figure 1: Instruction for annotators in data collection interface.

## Definitions

**Relevance**

The rating measures how well the summary captures the key points of the dialogue.

Consider whether all and only the important aspects are contained in the summary.

**Consistency**

The rating measures the whether the facts in the summary are consistent with the facts in the dialogue.

Consider whether the summary does reproduce all facts accurately and does not make up untrue information.

**Fluency**

The rating measures the quality of individual sentences, are they well-written and grammatically correct.

Consider the quality of individual sentences.

**Coherance**

The rating measures the quality of all sentences collectively, to the fit together and sound naturally.

Consider the quality of the summary as a whole.

Figure 2: Definition for annotators in data collection interface.

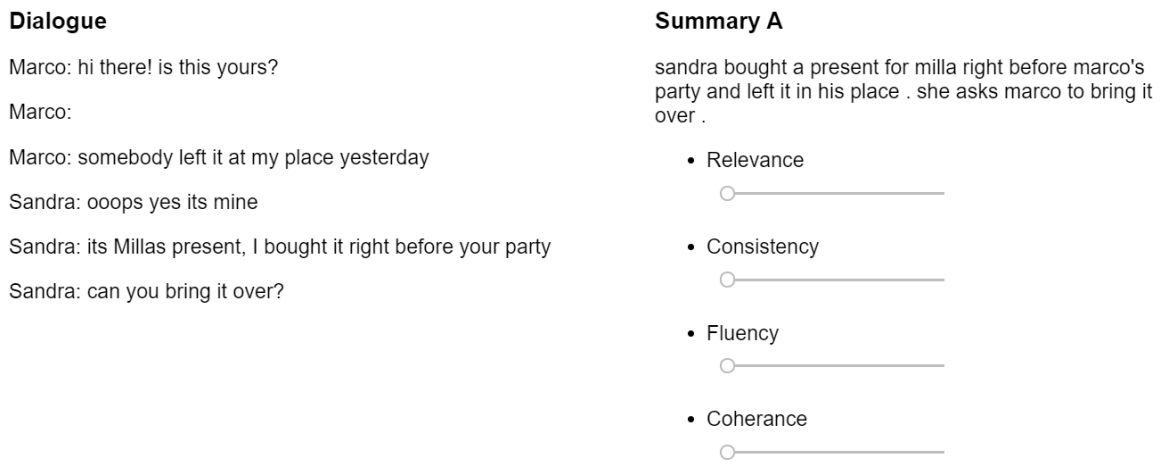


Figure 3: Annotation example in data collection interface.

| Models | Coherence | Consistency | Fluency | Relevance |
|---|---|---|---|---|
| reference summary | 3.308 | 3.300 | 3.396 | 3.380 |
| **LONGEST-3** | 3.220 | 3.230 | 3.286 | 3.306 |
| **LEAD-3** | 3.256 | 3.228 | 3.312 | 3.334 |
| **PGN** | 3.260 | 3.206 | 3.336 | 3.280 |
| **Tranformer** | 3.240 | 3.248 | 3.294 | 3.320 |
| **BART** | 3.286 | 3.298 | 3.410 | 3.358 |
| **PEGASUS** | 3.354 | 3.360 | 3.356 | 3.302 |
| **UniLM** | 3.288 | 3.342 | 3.390 | 3.364 |
| **CODS** | 3.346 | 3.328 | 3.384 | 3.396 |
| **ConvoSumm** | 3.368 | 3.334 | 3.420 | 3.426 |
| **MV-BART** | 3.232 | 3.260 | 3.366 | 3.344 |
| **PLM-BART** | 3.302 | 3.284 | 3.360 | 3.432 |
| **Ctrl-DiaSumm** | 3.232 | 3.300 | 3.360 | 3.348 |
| **S-BART** | 3.358 | 3.400 | 3.354 | 3.380 |

Table 5: Human ratings of summaries along four evaluation dimensions using data from Amazon Mechanical Turk. scores are averaged over five annotators, broken down by the approximate classification in Section 3.3