Reference Matters: Benchmarking Factual Error Correction for Dialogue Summarization with Fine-grained Evaluation Framework

Mingqi Gao[1,2,3]**, Xiaojun Wan**[1,2,3]**, Jia Su**[4]**, Zhefeng Wang**[4]**, Baoxing Huai**[4]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Center for Data Science, Peking University
[3]The MOE Key Laboratory of Computational Linguistics, Peking University
[4]Huawei Cloud
{gaomingqi,wanxiaojun}@pku.edu.cn
{sujia3,wangzhefeng,huaibaoxing}@huawei.com

Abstract

Factuality is important to dialogue summarization. Factual error correction (FEC) of model-generated summaries is one way to improve factuality. Current FEC evaluation that relies on factuality metrics is not reliable and detailed enough. To address this problem, we are the first to manually annotate a FEC dataset for dialogue summarization containing 4000 items and propose FERRANTI, a fine-grained evaluation framework based on reference correction that automatically evaluates the performance of FEC models on different error categories. Using this evaluation framework, we conduct sufficient experiments with FEC approaches under a variety of settings and find the best training modes and significant differences in the performance of the existing approaches on different factual error categories. [1]

1 Introduction

Factuality (also known as factual consistency, faithfulness) is a crucial dimension in evaluating summary quality. The summaries generated by current summarization models, and even some reference summaries, still have much room for improvement in factuality (Maynez et al., 2020; Fabbri et al., 2021b; Pagnoni et al., 2021). Dialogue summarization, a recently popular subfield of text summarization, has more challenging factual issues involved (Wang et al., 2022; Gao and Wan, 2022).

The prior approaches to enhance the factuality of summaries can be broadly classified into two categories: one is to introduce factuality-related objectives in training or inference process to make the summarization models more faithful, which is a direct generation of factually better summaries (Falke et al., 2019; Liu and Chen, 2021; Wan and Bansal, 2022; Tang et al., 2022; Liu et al., 2021); the other is to design a factual error correction (FEC) model independent of the summarization models, which takes the source document and the summary to be corrected as input and outputs a corrected summary (Cao et al., 2020; Dong et al., 2020; Zhu et al., 2021; Chen et al., 2021a; Fabbri et al., 2022b; Balachandran et al., 2022). There are a number of studies on news summarization that can fall into both categories. To the best of our knowledge, there has been no work on factual error correction for dialogue summarization. Considering the importance of factual issues in dialogue summarization, we would like to try to correct factual errors in dialogue summaries.

However, after carefully examining and considering the motivations and practices of previous FEC studies, we argue that there are flaws in the way FEC models are evaluated, which may have diverted the FEC for summarization from its original purpose. Previous studies evaluate the effectiveness of FEC models mainly by judging whether the scores of factuality metrics (e.g. FactCC (Kryscinski et al., 2020)) of the corrected summaries increase compared to the original summaries. First, this evaluation mechanism is so vague that it is difficult to evaluate the effectiveness of factual error correction accurately: we neither know which parts of the original summary have factual errors nor whether the corrected summary addresses them as expected. Second, this evaluation mechanism also blurs the line between FEC for summarization and the direct generation of factually better summaries: the factual error correction model can ignore the content of the original summary and directly generate a different but more factually correct summary.

We argue that it is necessary to introduce manually annotated reference correction to address the above issues. Factual error correction for summarization has its basic requirement: to correct factual errors in the original summary by as few substitution, insertion, and deletion operations as possible to obtain a fluent and non-redundant summary.

[1] Code and data will be available at https://github.com/kite99520/DialSummFactCorr

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13932–13959
July 9-14, 2023 ©2023 Association for Computational Linguistics

This can be reflected in the manual annotation. The introduction of reference correction, on the one hand, provides more valuable data for the training of FEC models compared to pseudo data; on the other hand, and more importantly, it creates the condition for a more comprehensive and accurate evaluation of the performance of FEC models. We construct an evaluation framework that can assess the performance of FEC models on different factual error categories based on manually annotated references. Using this framework, we are able to comprehensively evaluate and analyze the performance of various FEC methods on dialogue summarization. Our work has the following three main contributions:

1) We collect the outputs of four common models on two dialogue summarization datasets and are the first to correct the factual errors in them manually. The dataset containing 4000 data items will be released to facilitate further research.

2) We propose FERRANTI, a fine-grained evaluation framework based on reference correction that provides a comprehensive assessment of the performance of FEC models on different categories of factual errors.

3) Based on the above dataset and evaluation framework, we conduct a comprehensive evaluation and analysis of the performance of multiple FEC methods for dialogue summarization under different settings to illustrate the role of manually annotated data and the weaknesses of current models.

2 Related Work

2.1 Dialogue Summarization Models

As datasets such as SAMSum (Gliwa et al., 2019) were proposed, many models designed for dialogue summarization sprang up. Many of them build on generic pre-trained generative models such as BART (Lewis et al., 2020), incorporating dialogue structure information such as multiple views (Chen and Yang, 2020), summary sketch (Wu et al., 2021), argument mining (Fabbri et al., 2021a), personal named entity (Liu and Chen, 2021), and discourse relations (Chen and Yang, 2021). The summaries generated by these systems contain factual errors. They are what the FEC model needs to correct.

2.2 FEC for Summarization

Cao et al. (2020) and Dong et al. (2020) can be considered as the first work on FEC for text summarization. Cao et al. (2020) apply data augmentation methods to transform the reference summary, obtain pseudo data to fine-tune the pre-trained model, and generate the corrected summary directly. In contrast, Dong et al. (2020) use a more conservative strategy: masking the entities in summary and training a QA model to select span as the answer from the source document. Balachandran et al. (2022) follow the idea of Cao et al. (2020) and generate harder pseudo data through infilling language models. A similar approach based on data augmentation is Zhu et al. (2021), which makes use of the knowledge graph extracted from the source document. Chen et al. (2021a) replace named entities and numbers in the summary to generate candidates, from which the best one is selected as the corrected summary. In addition, Fabbri et al. (2022b) train the model using sentence-compressed data and remove hallucinated entities from the summary. We will test some of these methods on real annotated data of dialogue summarization.

2.3 Factuality Evaluation for Summarization

There are two main types of metrics widely used to evaluate the factuality of summaries. A class of metrics based on natural language inference, which formulate factuality as the result or confidence of binary classification, such as FactCC (Kryscinski et al., 2020), DAE (Goyal and Durrett, 2020; Goyal and Durrett, 2021), and SUMMAC (Laban et al., 2022). The other class is QA-based metrics, which usually contain a module for question generation and a module for question answering, with different implementation details, such as FEQA (Durmus et al., 2020), SummaQA (Scialom et al., 2019), QuestEval (Scialom et al., 2021), and QAFactEval (Fabbri et al., 2022a). Besides, BARTScore (Yuan et al., 2021) is also used to assess factuality. Many of them are used to evaluate the effectiveness of FEC models for summarization.

2.4 Evaluation for Post-editing and Correction

Evidence-based factual error correction is to correct the factual errors in a claim with evidence texts from trustworthy knowledge bases (Thorne and Vlachos, 2021; Shah et al., 2020; Chen et al., 2022). Reference-based evaluation metrics SARI

(Xu et al., 2016) and ROUGE correlate highly with human judgments on evidence-based FEC (Thorne and Vlachos, 2021). Automatic post-editing (APE) of machine translation and grammar error correction (GEC) also mainly use reference-based metrics (Chollampatt et al., 2020). For APE, they are BLEU, TER (Snover et al., 2006), and CHRF (Popović, 2015). For GEC, they are M^2 (Dahlmeier and Ng, 2012) and ERRANT (Bryant et al., 2017). From the above, it is clear that these post-editing or correction tasks use reference-based evaluation metrics if manual annotation data are available.

3 Data Annotation

3.1 Source Data Selection

We select SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021b), the two most widely used datasets in the field of short dialogue summarization, and collect summaries generated by four systems, BART (Lewis et al., 2020), UniLM (Dong et al., 2019), MV-BART (Chen and Yang, 2020) and CODS (Wu et al., 2021), on their test sets. The outputs of each system on the SAMSum test set are obtained from DialSummEval (Gao and Wan, 2022). For DialogSum, the outputs of BART and UniLM are provided by the authors of the dataset, and we retrain MV-BART and CODS on DialogSum with default settings to obtain their outputs.

We randomly sample 500 dialogues from each of the test sets of SAMSum and DialogSum, and the corresponding summaries of the above four systems, for a total of $2 \times 500 \times 4 = 4000$ dialogue-summary pairs, as the raw data to be annotated.

3.2 Annotation Process

We recruited college students as annotators. Annotators are required to be able to read and understand English daily conversations and articles fluently and have good English writing skills.

We designed the annotation interface by tagtog [2] to allow annotators to easily annotate multiple types of data. One dialogue and four system summaries are shown to the annotator at the same time. For each summary, the annotators will determine whether it is factually correct first. If there are factual errors in the summary, they will drag the mouse to mark the words and phrases which are factually inconsistent with the dialogue and then assign an error category by clicking the word and phrases they select. A summary may contain more than one error. Finally, if the summary contains any factual errors, they will write a corrected summary. Otherwise, the corrected summary will be the same as the original.

A detailed annotation guide was given to annotators to help them be familiar with the annotation interface and the definition of the task. Here we follow the taxonomy of factual errors proposed by Pagnoni et al. (2021). There are eight kinds of factual errors: (1) Entity Error (**EntE**); (2) Predicate Error (**PredE**); (3) Circumstance Error (**CircE**); (4) Coreference Error (**CorefE**); (5) Discourse Link Error (**LinkE**); (6) Out of Article Error (**OutE**); (7) Grammatical Error (**GramE**); (8) Others (**OthE**). Please see examples in Appendix A.

When correcting factual errors, the annotators needed to follow the three principles: (1) Correct factual errors with as few modifications as possible. (2) Making substitutions for words and phrases is preferred. When substitution is difficult, deletion can be performed. (3) The corrected summary should be grammatically correct, coherent, and non-redundant as possible.

We divided the original data into 10 batches, each containing 100 dialogues ($100 \times 4 = 400$ items). In order to ensure annotation quality, those who wished to participate in the annotation were required to complete the annotation of all the summaries corresponding to the 10 dialogues ($10 \times 4 = 40$ items) first. After completing this small part, we evaluated the annotation results, pointed out any inappropriate annotations, and told them our suggestions. After confirming that the annotation task was correctly understood, the participants were allowed to continue annotation. In subsequent annotation, we sampled the results to check. Throughout the process, we kept in touch with the annotators via email and instant messaging software.

3.3 Data Analysis

It is necessary to illustrate the difference between the manually annotated corrected summaries and the reference summaries in the dialogue summarization dataset. We focus on their relationship to the summaries to be corrected. Since the summaries that do not contain factual errors do not need to be corrected, i.e., the corrected summaries are the same as the original summaries, we only count data for samples where the original summaries contain factual errors. For these samples, it can be seen from Figure 1 that the corrected summaries

[2] https://www.tagtog.com/

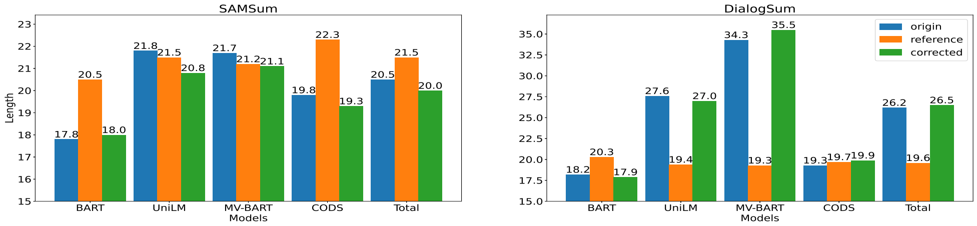


Figure 1: The average length of original summaries, reference summaries, and corrected summaries. Only items with factual errors in the original summary are counted.

	BART	UniLM	MV-BART	CODS	Total
SAMSum	0.43 / 0.85	0.39 / 0.82	0.45 / 0.85	0.46 / 0.84	0.43 / 0.84
DialogSum	0.62 / 0.73	0.54 / 0.68	0.56 / 0.76	0.61 / 0.72	0.58 / 0.72

Table 1: BLEU score comparison (origin vs. reference / origin vs. corrected). Only items with factual errors in the original summary are counted.

	BART	UniLM	MV-BART	CODS	Total
SAMSum	26.00	51.20	37.00	44.40	39.65
DialogSum	31.20	44.80	58.00	40.60	43.65

Table 2: Percentage of summaries with factual errors.

are closer in length to the original summaries compared to the reference summaries. This is more obvious on DialogSum. As shown in Table 1, the corrected summaries are closer to the original summaries in terms of n-gram overlap compared to the reference summaries. This result is in line with our annotation principles.

For the percentage of factual inconsistencies and error categories, as shown in Table 2, around 40% of generated summaries contain factual errors. This ratio is similar to the annotation results of Wang et al. (2022). Figure 2 shows that, **EntE** and **PredE** are the two most dominant types of errors. It is important to note that the percentage of **GramE** (difficult to understand due to grammatical errors) is less. This is in line with the findings of Gao and Wan (2022): the current dialogue summarization systems based on pre-trained models generate summaries that are already good in terms of fluency.

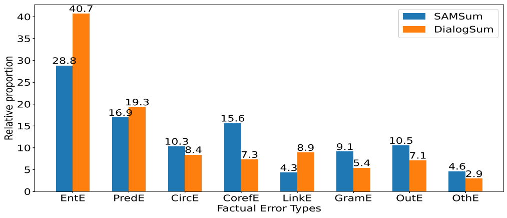


Figure 2: The relative percentage of factual error types. A factually incorrect span is counted as one error.

4 Test for Factuality Metrics

We perform a simple test of the reliability of factuality metrics using the above dataset. In general, the factuality metric F takes the source document S and the summary H as inputs and outputs a score $F(S, H)$. A reliable factual indicator needs to satisfy the condition that for summaries with factual errors, the factual score of the corrected summary C is greater than that of the original summary O, i.e., $F(S, C) > F(S, O)$.

We select four commonly used factuality metrics: FactCC (Kryscinski et al., 2020), DAE (Goyal and Durrett, 2020; Goyal and Durrett, 2021), QuestEval (Scialom et al., 2021), and BARTScore (Yuan et al., 2021). Table 3 illustrates that it is unreliable to evaluate the factuality of the original and corrected summaries using these metrics. The factuality scores of the corrected summaries are not significantly better than those of the original summaries under these metrics, either in mean or pairwise comparisons.

5 Reference-based Evaluation Framework

We find that it is difficult for manual annotation to determine the boundaries of erroneous spans accurately sometimes, which hinders the fine-grained evaluation of FEC models by error categories. Considering these error categories have clear linguistic characteristics, it is more feasible to use a rule-based approach to automatically align and determine error categories when reference correction is already available.

We propose FERRANTI, a **F**actual **ERR**or **AN**notation **T**oolk**I**t designed for FEC. Noting the great practice of ERRANT (Bryant et al., 2017), our implementation builds on it. As shown in Figure 3, it mainly consists of three steps: alignment, classification, and comparison.

SAMSum (N=793)						DialogSum (N=873)					
	origin	**correct**	<	=	>		**origin**	**correct**	<	=	>
FactCC	0.136	0.139	0.04	0.93	0.03	**FactCC**	0.286	0.276	0.04	0.91	0.05
DAE	0.076	0.077	0.02	0.97	0.02	**DAE**	0.199	0.207	0.04	0.93	0.03
QuestEval	0.392	0.380	0.30	0.23	0.47	**QuestEval**	0.486	0.486	0.34	0.31	0.34
BARTScore	-3.084	-3.123	0.25	0.53	0.22	**BARTScore**	-2.826	-2.810	0.21	0.57	0.22

Table 3: Test results of factuality metrics. The higher the output scores of the metrics, the better the factuality. Only summaries with factual errors are counted. **origin** and **correct** refer to the average of the output scores of the original summaries and the corrected summaries. **<**, **=**, and **>** refer to the proportion of scores of the original summary that are less than, equal to, and greater than that of the corrected summary, when compared as a pair.

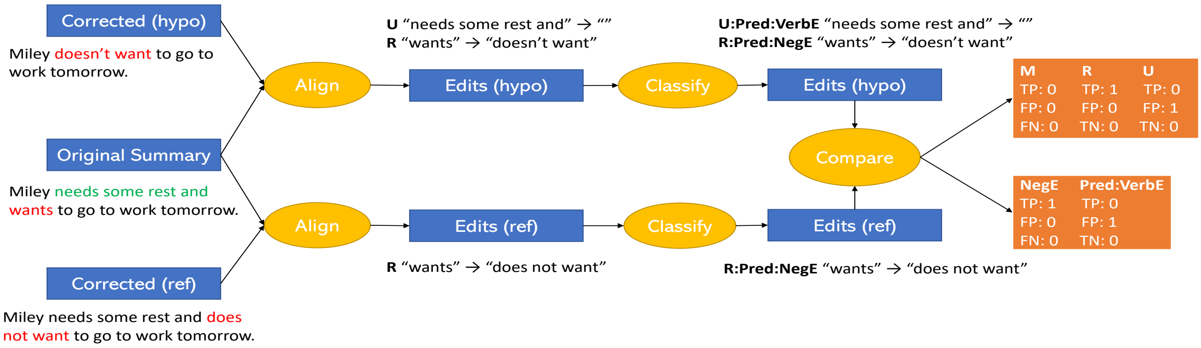

Corrected (hypo)
Miley doesn't want to go to work tomorrow.
Original Summary
Miley needs some rest and wants to go to work tomorrow.
Corrected (ref)
Miley needs some rest and does not want to go to work tomorrow.
Align
Align
U "needs some rest and" → ""
R "wants" → "doesn't want"
Edits (hypo)
Edits (ref)
R "wants" → "does not want"
Classify
Classify
U:Pred:VerbE "needs some rest and" → ""
R:Pred:NegE "wants" → "doesn't want"
Edits (hypo)
Compare
Edits (ref)
R:Pred:NegE "wants" → "does not want"
M TP: 0 FP: 0 FN: 0
R TP: 1 FP: 0 TN: 0
U TP: 0 FP: 1 TN: 0
NegE TP: 1 FP: 0 FN: 0
Pred:VerbE TP: 0 FP: 1 TN: 0

Figure 3: Diagram of our evaluation framework, FERRANTI. Red parts indicate replacement (R). Green parts indicate deletion (U). Addition edits do not appear in this example (M).

5.1 Taxonomy of Factual Errors

To automatically classify factual errors for FEC, we propose a new taxonomy of factual errors. Compared to existing classifications of factual errors, such as Pagnoni et al. (2021), Tang et al. (2022) and Wang et al. (2022), our taxonomy differs in three main ways: (1) we point out that there are two classifications of factual errors of different perspectives, content-based and form-based; (2) we hierarchize the content-based classification of factual errors; (3) our error classification is implemented by explicit linguistic rules rather than manual annotation.

The content-based categories are shown in Table 5. In this classification, the category to which an edit belongs needs to draw on the POS of the words in the sentence as well as on the dependencies. Compared to the classification we used in the annotation, we subdivide **EntE** and **PredE**, add **NumE**, and do not use **OutE** and **GramE** that have unclear POS and dependency features. By this, we cover special categories such as negation errors (**NegE**) that received attention in summarization factuality without losing generality.

The form-based categories are shown in Table 4. They are called form-based because, in factual error correction, it is basically only necessary to align the original summary and the corrected summary by whether the words are the same to determine whether an edit is an addition, deletion, or modification. Devaraj et al. (2022) adopt a similar way when analyzing the factuality of text simplification.

It is necessary to point out that the form-based and content-based classifications are not mutually exclusive. They can be combined, such as **R:Pred:Neg** in Figure 3.

5.2 Alignment

In this step, the corrected summaries are aligned with the original ones and the edits are extracted automatically. We follow ERRANT by using an alignment algorithm that considers linguistic features as a cost function (Felice et al., 2016). However, unlike ERRANT, we merge all adjacent edits considering that a small number of factually corrected edits are longer. Before alignment, the summary is pre-processed with Spacy[3] for tokenization, POS tagging, etc. Form-based error categories are automatically assigned to each edit after alignment.

5.3 Classification

After edits are extracted, they are assigned content-based categories based on the linguistic features of the original span and the corrected span (mainly

[3] version 2.3.0, https://spacy.io/

Code	Meaning	Description	Examples
M	Missing	Missing information that needs to be added.	*with Ms.* → *with Ms. Blair*
R	Replacement	Wrong information that needs to be modified.	*reminds* → *teaches*
U	Unnecessary	Redundant information that needs to be deleted.	*Derek and Phil* → *Derek*

Table 4: Form-based categories of factual errors.

Code	Description	Example
Ent:ObjE	Object errors in entity errors, mainly nouns.	*Laura* → *Paul*
Ent:AttrE	Attribute errors in entity errors, mainly adjectives.	*proud* → *happy*
Pred:ModE	Modality errors in predicate errors, mainly modal verbs that express possibilities.	*is* → *may be*
Pred:TensE	Tense errors in predicate errors.	*is* → *was*
Pred:NegE	Negation errors in predicate errors.	*will* → *won't*
Pred:VerbE	General predicate errors that do not fall into the above categories.	*lent* → *gave*
CircE	Circumstance errors, mainly adverbs, prepositional phrases, etc.	*after* → *during*
CorefE	Coreference errors, mainly pronouns.	*her* → *Ann*
LinkE	Link errors, conjunctions	*but* → *because*
NumE	Errors in numbers	*15* → *30*
OthE	Other errors that are not all of the above types of errors.	*, so she* → *. She*

Table 5: Content-based categories of factual errors. The examples in the table are all replacements, but deletions and additions are also possible.

POS and lemma). The detailed rules are not listed here.

5.4 Comparison

In this step, hypothesis edits and reference edits are compared and scores are computed in different categories for form-based and content-based categories. Edits that appear in both hypothesis and reference are true positive (**TP**). For TP, we use the category of edits in reference as the final category. Edits that appear only in the hypothesis or reference are false positive (**FP**) or false negative (**FN**). Further, we can obtain precision, recall, and F-values. We report $F_{0.5}$ out of a penalty for over-correction.

6 Experiments

6.1 FEC approaches

We select a few representative FEC approaches. Among them, we are most interested in such methods: generating corrected summaries directly based on data augmentation because of their flexibility.

Rule-based transformation Cao et al. (2020) use a set of rules that swap the entities, numbers, dates, and pronouns of the reference summaries to construct the summaries to be corrected for training. We call this approach **rule**.

Infilling-based transformation Balachandran et al. (2022) mask and predict the subjects, relations, and objects of sentences in the source documents to train an infilling model. The reference summaries are then masked in the same way, and the trained infilling model is used to fill the masked reference summaries to construct the summaries to be corrected. For the infilling model, we experiment with two different setups: (1) using the trained infilling model from the original study, denoted as **infill**. (2) retraining the infilling model , denoted as **infill-r**. Please see Appendix C for the details of retraining.

In addition to the method of generating a corrected summary directly, we also select other approaches, which aim at correcting extrinsic hallucinations:

CCGS Chen et al. (2021a) replace named entities and numbers in reference summary with the compatible semantic type of content from the source document to generate candidates to train a factual classifier based on BART. At the time of inference, the candidates for the summary to be corrected are generated in a similar way, the trained classifier is used to re-rank the candidates, and the best one is selected as the corrected summary.

FactPegasus Wan and Bansal (2022) propose a component for correcting factual errors without training data: based on manually written rules and

SAMSum						DialogSum					
BART as the pre-trained model						BART as the pre-trained model					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**	**Type**	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
M	0.00	0.00	0.00	**2.08**	2.02	**M**	**5.49**	0.00	0.00	0.75	2.32
R	4.26	7.58	**15.00**	2.34	1.44	**R**	**3.48**	1.72	1.74	1.58	1.34
U	7.04	6.07	**13.66**	3.89	4.66	**U**	**12.05**	4.32	4.57	3.02	2.43
Total	4.15	5.63	**13.01**	2.54	2.33	**Total**	**4.24**	2.43	2.31	1.76	1.66
PEGASUS as pre-trained models						PEGASUS as pre-trained models					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**	**Type**	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
M	0.00	0.00	0.00	1.41	**1.59**	**M**	**14.93**	0.00	0.00	0.00	0.78
R	12.15	1.58	**13.72**	2.58	4.68	**R**	**9.32**	5.75	4.44	2.10	2.19
U	**7.46**	4.05	7.04	1.17	4.25	**U**	**13.33**	3.70	0.00	2.84	1.41
Total	9.48	2.15	**10.82**	1.99	4.18	**Total**	**10.25**	4.58	3.50	1.98	1.87
T5 as the pre-trained model						T5 as the pre-trained model					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**	**Type**	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
M	0.00	0.00	0.00	0.00	**0.88**	**M**	**13.33**	0.00	0.00	1.45	3.26
R	10.54	0.00	**16.18**	3.52	4.66	**R**	7.33	1.35	**8.29**	2.46	4.26
U	7.94	18.99	**24.10**	6.26	7.99	**U**	7.46	16.95	**18.18**	3.36	4.18
Total	8.89	4.72	**15.69**	3.74	4.57	**Total**	7.89	2.98	**8.33**	2.50	4.12

Table 6: Performance (FERRANTI: form-based categories defined in Table 4) of different training modes on SAMSum and DialogSum. The values are all $F_{0.5}$ scores. The best results under the same pre-trained model are bolded. The data augmentation approach is set to **rule**. The pseudo data for the two datasets are constructed separately. Please see Table 18, Table 19, Table 20, and Table 21 in Appendix E for precision, recall, or TP, etc.

the Spacy library, and it removes or replaces entities and related content in the summary that do not appear in the source document.

6.2 Training Modes

For different data augmentation approaches (**rule**, **infill**, and **infill-r**), we conduct experiments with different training modes to explore some factors of interest. To compare the role played by pseudo data (generated by data augmentation) and real data (manually annotated) in training, we designed the following training modes: (1) Training with pseudo data only (**Pseudo**). (2) Training with real data only (**Real**). (3) Training with pseudo data first, then with real data (**Pseudo + Real**). In order to compare the difference between the reference correction and the reference summary of the summarization dataset, we also design the following training modes: (4) Replace the reference correction in the real data with the reference summary for training (**RefS**). (5) Training with pseudo data first, then proceed to (4) (**Pseduo + RefS**).

6.3 Datasets and Settings

We split our annotated dataset (which we call the real data) into a training set, a validation set, and a test set. Specifically, for the 500 dialogues of SAMSum, we split them according as 300/100/100. Each dialogue has the corresponding four model-generated original summaries and corrected summaries. The total size is 1200/400/400. For the 500 dialogue of DialogSum, the split is the same as SAMSum. We train and test models separately on the two parts (datasets). Please see Appendix D for model settings and training details.

6.4 Evaluation

We use the evaluation framework presented in Section 5, FERRANTI to automatically evaluate FEC approaches on the test set. For comparison, we also adopt factuality metrics mentioned in Section 4.

7 Results and Analysis

7.1 Performance across Training Modes

Here we show the results on the category of form-based errors. Content-based results are shown in Table 22 and Table 23 in Appendix E.

Reference summary vs. Reference correction Table 6 illustrates that in most cases where FERRANTI is used as the evaluation framework, training FEC models using the reference summary as the final correction target (**RefS**, **Pseudo+RefS**) does not yield good results. Tables 19 and 21 in Appendix E illustrate that both modes present many FPs on various error types, i.e., false edits. This is to be expected since we have shown in Section 3 that there is a large difference between the reference correction and the reference summary. Interestingly, if evaluated using factuality metrics, we find that training with the reference summary gives the best results in most cases (the results are shown in Table 17 in Appendix E). This suggests

	BART as the pre-trained model						PEGASUS as the pre-trained model						T5 as the pre-trained model					
	Pseudo			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**		
	rule	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**
Ent:ObjE	2.02	**11.36**	4.72	10.59	9.09	9.62	14.71	5.21	8.93	**16.83**	11.90	9.62	14.34	4.03	3.47	**16.20**	3.47	7.46
Ent:AttrE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:ModE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:TensE	-	-	0.00	0.00	-	-	-	0.00	0.00	-	0.00	-	-	-	-	-	-	-
Pred:NegE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:VerbE	4.76	2.65	2.70	**12.12**	7.30	11.30	0.00	3.76	3.18	0.00	**9.52**	4.59	0.00	0.00	0.00	13.76	12.00	**14.29**
CircE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CorefE	7.04	0.00	0.00	**25.32**	10.99	7.04	0.00	0.00	0.00	5.75	6.67	**17.24**	7.04	0.00	0.00	**24.27**	6.02	7.46
LinkE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NumE	31.25	0.00	0.00	**41.67**	0.00	0.00	41.67	0.00	0.00	**50.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OthE	-	0.00	-	-	-	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	4.15	5.23	2.92	**13.01**	8.10	8.96	9.48	3.33	4.44	**10.82**	8.54	8.62	8.89	1.57	1.27	**15.69**	6.28	6.98

Table 7: Performance (FERRANTI: content-based categories defined in Table 5 of different data augmentation approaches on SAMSum. The results on DialogSum are shown in Table 28 in Appendix E. The values are all $F_{0.5}$ scores. The best results under the same pre-trained model are bolded. The pseudo-data corpus is SAMSum. Please see Table 26 and Table 27 in Appendix E for precision, recall, or TP, etc.

SAMSum				**DialogSum**			
Pre-trained Models	BART	BERT	RoBERTa	Pre-trained Models	BART	BERT	RoBERTa
M	0.00	0.00	0.00	**M**	0.00	0.00	0.00
R	1.98	0.00	1.52	**R**	1.06	2.22	2.42
U	0.00	0.00	0.00	**U**	0.00	0.00	0.00
Total	1.41	0.00	1.14	**Total**	0.87	1.80	1.91

Table 8: Performance (FERRANTI: form-based categories) of CCGS on SAMSum and DialogSum.

SAMSum				**DialogSum**			
Spacy Models	sm	md	lg	Spacy Models	sm	md	lg
M	0.00	0.00	0.00	**M**	0.00	0.00	0.00
R	0.00	0.00	0.00	**R**	0.00	0.00	0.00
U	1.80	1.62	3.99	**U**	0.58	1.32	0.36
Total	1.42	1.21	2.98	**Total**	0.37	0.91	0.26

Table 9: Performance (FERRANTI: form-based categories) of FactPegasus on SAMSum and DialogSum.

that it is essential to introduce reference correction in FEC evaluation. Otherwise, FEC for summarization may lose its meaning, since the seemingly best results can be obtained by using reference summaries unrelated to the original summaries as training targets.

Real data vs. Pseudo data Table 6 shows that training with pseudo data first and then with real data (**Pseudo+Real**) or training with only pseudo data (**Pseduo**) are the two best training modes. The former is better on SAMSum and the latter is better on DialogSum. Here we cannot say that real data is less effective because there is a huge difference in the size between real and pseudo data: real training data is only 1200 items on each dataset; while the generated pseudo data are 40451 and 35174 items on SAMSum and DialogSum, respectively. This on the one hand corroborates the effectiveness of the FEC approach based on data augmentation in the past, and on the other hand, implies that the combination of real and pseudo data is promising.

Regarding the performance on the form-based error categories: On both datasets, most of the edits are in the **Replacement** category (see Table 19 and Table 21 in Appendix E). Table 6 illustrates that using the reference correction as the final training goal (**Real**, **Pseudo+Real**) performs poorly on the **Missing** category. This indicates that it is difficult for models to learn addition operations in manual correction.

In addition, we also try to mix SAMSum and DialogSum as a corpus for constructing pseudo data. Table 36 in Appendix E illustrates that in some cases, the mixed construction has better results than the separate construction. For comparison, we still construct the pseudo data separately in the subsequent experiments.

7.2 Performance across FEC Approaches

Here we mainly show the results of data augmentation approaches on the category of content-based errors. Form-based results are shown in Table 31 in Appendix E. Training modes are set to **Pseudo** and **Pseudo+Real**.

Ent:ObjE and **Pred:VerbE** are the two main error types (see Tables 27 and 30 in Appendix E), which coincide with our annotation results in Section 3. An important finding is that Tables 7 (and Table 28 in Appendix E) show that these methods based on data augmentation for generating corrected summaries directly show error-correcting power only for a few categories: **Ent:ObjE**, **Pred:VerbE**, **CorefE**, **NumE**, and **OthE**. We argue that this cannot be attributed only to the chance brought by the small percentage of some error categories. The strategy of data augmentation is an important factor. Because we notice the fact that the rule-based data augmentation approach performs swapping on numbers, and it has a relatively great performance on **NumE** on SAMSum, even though the percentage of **NumE** is small.

The infilling-based data augmentation method is generally inferior to the rule-based data augmentation method. Its performance also changes insignificantly after retraining. The particular structural information in the conversation summaries has to be further exploited. The infilling-based method sometimes performs better on **Pred:VerbE**. This may be due to the fact that it masks and predicts the relations in the reference summary when constructing pseudo data, with verb phrases in the relations.

In addition, both CCGS and Factpegasus perform poorly. Table 8 illustrates that CCGS can only correct errors in the form of substitution. Table 9 illustrates that Factpegasus can only correct errors by deletion. This is consistent with their algorithms. Table 32 and Table 33 in Appendix E illustrate that they can almost correct only one type of errors, **Ent:ObjE**.

However, the above findings would not have been available if we had used only factuality metrics (see Table 24, Table 25, Table 34 and Table 35 in Appendix E). This illustrates the superiority of FERRANTI.

8 Conclusion

Our work establishes a new benchmark for model-agnostic factual error correction for dialogue summarization. Unlike previous studies, we manually correct factual errors in summaries. We point out the shortcomings of factuality metrics in FEC evaluation: They are not reliable enough and cannot provide more detailed information. For better evaluation, we propose FERRANTI, a reference-based evaluation framework and conduct thorough experiments on the performance of multiple FEC approaches under various settings. We have the following important findings:

1) Training FEC models with reference summaries from dialogue summarization datasets yields the best results of unreliable factuality metrics. There is an urgent need to change the evaluation methods for FEC models.

2) Introducing human-corrected summaries during the training of FEC models for dialogue summarization can improve their performance. Combining human-annotated data with synthetic data is a promising direction.

3) Current FEC models struggle to correct factual errors by addition and cannot address attribute errors, modality errors, link errors, etc.

For future work, it is feasible to apply FERRANTI to FEC for other summarization tasks.

Limitations

Due to limited resources, the size of our annotated dataset is not large, with only 4000 items. In addition, we use an annotation paradigm where direct writing is the main focus with error labeling as a supplement. This is good for the coherence of the corrected summary and gives larger freedom to the annotator. In this case, it may be better to increase the number of reference corrections per sample. The datasets we select, SAMSum and DialogSum, are both short daily chat summarization datasets. For other domains or long dialogue summarization, our conclusion may not apply.

About FERRANTI, it can be continuously improved since we automatically classify and label factual errors for the first time. It also relies on the lexical and syntactic nature of English.

Ethics Statement

We recruit annotators through the campus BBS. They are completely free to decide whether to participate and can quit in the middle. They are paid $15 per hour, more than the local minimum wage. No participants' personal information or payment information will be released. Some of the information is temporarily stored on the server and will be deleted at the end of the study.

The application of datasets, models, and tools in our study is consistent with their intended use and license. We hope the artifacts we release are to be used for academic research (non-commercial licence: CC BY-NC 4.0).

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFF0901502), National Science Foundation of China (No. 62161160339), State Key Laboratory of Media Convergence Production Technology and Systems and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. *Computing Research Repository*, arXiv:2210.12378.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, 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.

Jiangjie Chen, Rui Xu, Wenyuan Zeng, Changzhi Sun, Lei Li, and Yanghua Xiao. 2022. Converge to the truth: Factual error correction via iterative constrained editing. *Computing Research Repository*, arXiv:2211.12130.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021a. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. 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.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020. Can automatic post-editing improve NMT? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2736–2746, Online. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

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 Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022a. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R. Fabbri, Prafulla Kumar Choubey, Jesse Vig, Chien-Sheng Wu, and Caiming Xiong. 2022b. Improving factual consistency in summarization with compression-based post-editing. *Computing Research Repository*, abs/2211.06196.

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.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. 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.

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.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

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.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. 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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

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.

Darsh Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8791–8798.

Matthew Snover, Bonnie Dorr, Richard Shwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas*.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. 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 3298–3309, Online. Association for Computational Linguistics.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and evaluating faithfulness in dialogue summarization. *Computing Research Repository*, arXiv:2210.11777.

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.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

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.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

A Details of Annotation

The annotators were told that the collected data would be used for academic study. In total, 10 people participated in the annotation. Two people read the annotation guidelines and then abandoned further annotation. One person annotated the small part used for testing and then gave up on further annotation. The other seven qualified participants who continued to annotate are from Asia. Three of them are female and four of them are male. One annotated three batches, another annotated two batches, and the others annotated one batch each. The screenshot of the annotation interface is shown in Figure 4 in Appendix E. Considering the space for corrected summaries is relatively narrow, we provide an excel file for annotators to help them write the corrected summaries (shown in Figure 5 in Appendix E). They can copy what they write in the excel file and paste it into the interface. They decide whether to use the excel file according to their needs. We use what they submit in the interface as the final result.

We provide the same definition of error categories for annotators as Pagnoni et al. (2021), but with different examples because the original examples are news summaries. They are shown in Table 10, Table 13, Table 14, Table 11, Table 15, Table 12, and Table 16 in Appendix A.

Entity Error (EntE)
Dialogue
Ola: Hey running late **Ola**: I should be free by 8 **Kurt**: Sure no prob, call me
Original Summary
Ola will be late. Kurt will call him by 8.
Corrected Summary
Ola will be late. He will call Kurt.

Table 10: An example of Entity Error.

Coreference Error (CorefE)
Dialogue
Ola: Hey running late **Ola**: I should be free by 8 **Kurt**: Sure no prob, call me
Original Summary
Ola will be late. Kurt will call him by 8.
Corrected Summary
Ola will be late. He will call Kurt.

Table 11: An example of Coreference Error.

Out of Article Error (OutE)
Dialogue
Dave: Hey, is Nicky still at your place? Her phone is off **Sam**: She just left **Dave**: Thanks!
Original Summary
Nicky just left her phone at Dave's place .
Corrected Summary
Nicky just left Dave's place .

Table 12: An example of Out of Article Error.

B Details of the use of factuality metrics

For **FactCC**[4] and **DAE** [5], We follow the way Pagnoni et al. (2021) used it. The summary is split into sentences by NLTK [6]. Each sentence is classified as CORRECT or INCORRECT. The factual score of a summary is represented as the ratio of factually correct sentences.

For **QuestEval** [7], we use the reference-less mode. For **BARTScore** [8], we use the $s \rightarrow h$ mode and the checkpoint trained by the authors on Parabank2.

C Details of retraining infilling models

We retrain the infilling model on summaries generated by MV-BART (Chen and Yang, 2020). The original approach uses the source document to train the infilling model and then makes predictions on the reference summary, which is to enhance the diversity of the pseudo data. However, we find that most of the subjects and objects extracted from the source dialogues are first- and second-person pronouns, such as "I" and "you", which are too different from the summaries from the third-person perspective. In order to adapt this approach to dialogue summarization, instead of using source documents, we use summaries generated by a model as training data for the infilling model.

D Model Settings and Training Details

Many FEC methods involve the construction of pseudo data. When it comes to data augmentation based on reference summaries and source documents, we use the training and validation sets from the summarization datasets SAMSum and DialogSum rather than our annotated data.

For different data augmentation approaches (**rule**, **infill**, and **infill-r**), we uniformly concatenate the summary to be corrected and the source document as input, and fine-tune some pre-trained models with the corrected summary as output for the above approaches. We conduct separate experiments using BART [9], PEGASUS [10] (Zhang et al., 2020) , T5 [11] (Raffel et al., 2022). For all training modes, we fine-tune the pre-trained language models for 20 epochs with a batch size of 32, and use the loss on the validation set as the criterion for

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

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

[6] version 3.7, https://www.nltk.org/

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

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

[9] using checkpoint from https://huggingface.co/facebook/bart-large

[10] using checkpoint from https://huggingface.co/sshleifer/distill-pegasus-cnn-16-4

[11] using checkpoint from https://huggingface.co/t5-base

Predicate Error (PredE)
Dialogue
Will: hey babe, what do you want for dinner tonight? **Emma**: gah, don't even worry about it tonight **Will**: what do you mean? everything ok? **Emma**: not really, but it's ok, don't worry about cooking though, I'm not hungry **Will**: Well what time will you be home? **Emma**: soon, hopefully **Will**: you sure? Maybe you want me to pick you up? **Emma**: no no it's alright. I'll be home soon, i'll tell you when I get home. **Will**: Alright, love you. **Emma**: love you too.
Original Summary
Emma doesn't want to cook dinner tonight. She will tell Will when she gets home.
Corrected Summary
Emma is not hungry tonight. She will tell Will when she gets home.

Table 13: An example of Predicate Error.

Circumstance Error (CircE)
Dialogue
Lenny: Babe, can you help me with something? **Bob**: Sure, what's up? **Lenny**: Which one should I pick? **Bob**: Send me photos **Lenny**: <file_photo> **Lenny**: <file_photo> **Lenny**: <file_photo> **Bob**: I like the first ones best **Lenny**: But I already have purple trousers. Does it make sense to have two pairs? **Bob**: I have four black pairs :D :D **Lenny**: yeah, but shouldn't I pick a different color? **Bob**: what matters is what you'll give you the most outfit options **Lenny**: So I guess I'll buy the first or the third pair then **Bob**: Pick the best quality then **Lenny**: ur right, thx | **Bob**: no prob :)
Original Summary
Lenny will buy the first or the third pair of purple trousers for Bob.
Corrected Summary
Lenny will buy the first or the third pair of purple trousers.

Table 14: An example of Circumstance Error.

Discourse Link Error (LinkE)
Dialogue
The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.
Original Summary
To produce the vaccine, scientists have to show successful human trials, then sequence the DNA of the virus.
Corrected Summary
To produce the vaccine, scientists have to show successful human trials, after sequence the DNA of the virus.

Table 15: An example of Discourse Link Error. This example is taken from Pagnoni et al. (2021), and we add a corrected summary.

Grammatical Error (GramE)
Dialogue
Everett: Ralph asked me if i could give him your phone number, is that cool? **Amy**: who's ralph? **Everett**: my friend, i introduced him to you at the pub last week, tall, brown hair, weird laugh... **Amy**: oh i remember him now, is he a psycho? **Everett**: no **Amy**: ok, he can have my number
Original Summary
Everett will give him him phone number .
Corrected Summary
Everett will give Ralph Amy's phone number .

Table 16: An example of Grammatical Error.

saving the best checkpoint. The learning rate is set to 3e-5. Hyperparameters for training the infilling models are kept at their default values.

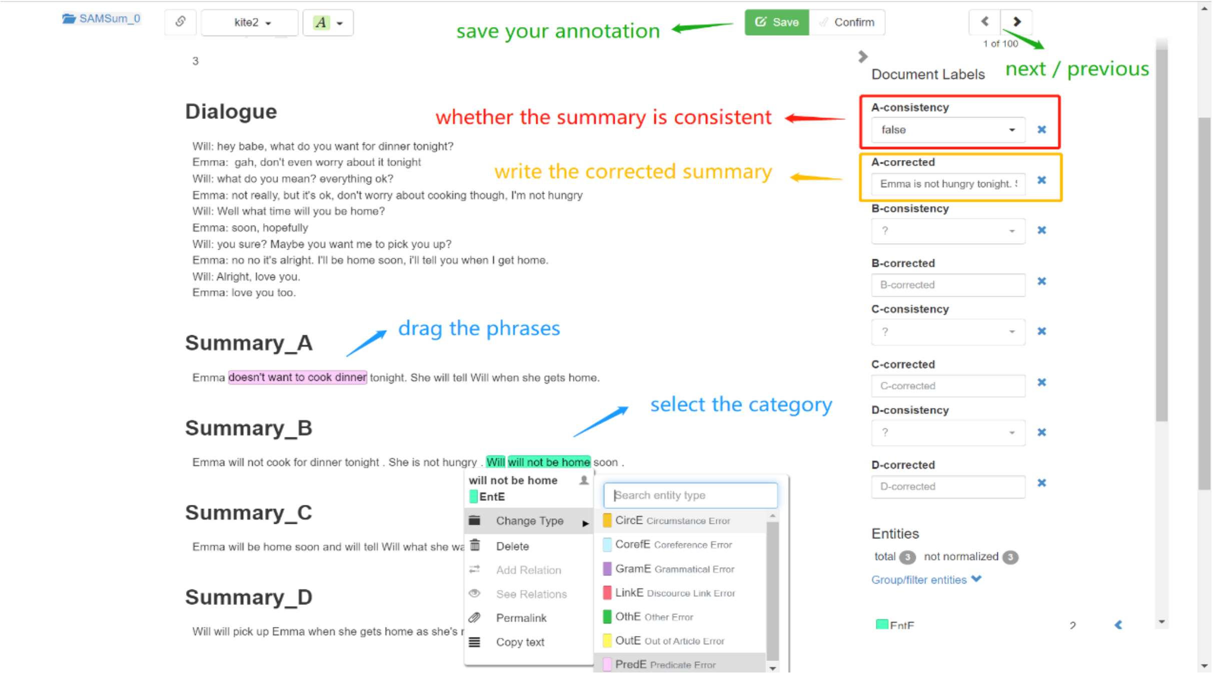
When constructing pseudo data, **rule** generates 40451 and 35174 items on the training sets of SAMSum and DialogSum, and 2259 and 1369 items on the validation set of SAMSum and DialogSum. both **infill** and **infill-r** generate more pseudo data than **rule**. We randomly sample the pseudo data generated from **infill** and **infill-r** to ensure that the number of pseudo-data is the same as **rule**.

For CCGS, we re-train the classifier according to the original approach. To reflect its effectiveness more comprehensively, in addition to BART, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are also used as pre-trained models for the classifier. Hyperparameters are kept at their default values.

For FactPegasus, we use three Spacy models (Version 2.2.4) to pre-process the text separately: `en_core_web_sm`, `en_core_web_md`, `en_core_web_lg`.

We use GeForce GTX 1080 Ti with 12GB memory for training and inference. Each single training session is less than 12 hours.

E Additional Figures and Tables



save your annotation

next / previous

whether the summary is consistent

write the corrected summary

drag the phrases

select the category

Dialogue

Will: hey babe, what do you want for dinner tonight?
Emma: gah, don't even worry about it tonight
Will: what do you mean? everything ok?
Emma: not really, but it's ok, don't worry about cooking though, I'm not hungry
Will: Well what time will you be home?
Emma: soon, hopefully
Will: you sure? Maybe you want me to pick you up?
Emma: no no it's alright. I'll be home soon, i'll tell you when I get home.
Will: Alright, love you.
Emma: love you too.

Summary_A

Emma doesn't want to cook dinner tonight. She will tell Will when she gets home.

Summary_B

Emma will not cook for dinner tonight . She is not hungry . Will will not be home soon .

will not be home
EntE

Summary_C

Emma will be home soon and will tell Will what she wa

Summary_D

Will will pick up Emma when she gets home as she's r

Document Labels

A-consistency: false

A-corrected: Emma is not hungry tonight. !

B-consistency: ?

B-corrected: B-corrected

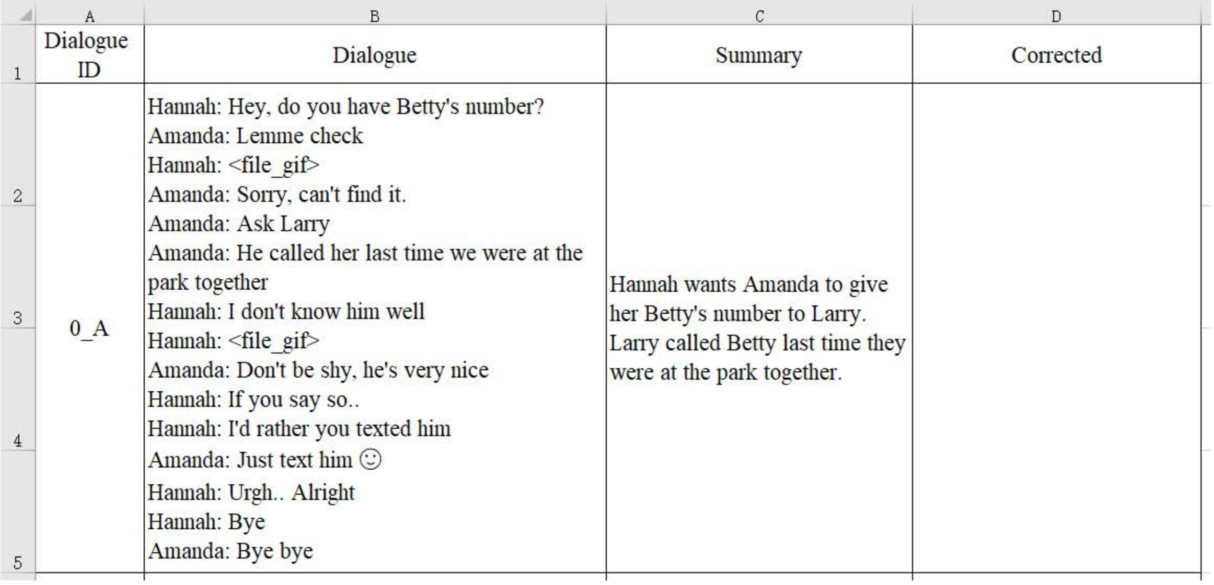
C-consistency: ?

C-corrected: C-corrected

D-consistency: ?

D-corrected: D-corrected

Figure 4: Annotation Interface.



Dialogue ID	Dialogue	Summary	Corrected
0_A	Hannah: Hey, do you have Betty's number? Amanda: Lemme check Hannah: <file_gif> Amanda: Sorry, can't find it. Amanda: Ask Larry Amanda: He called her last time we were at the park together Hannah: I don't know him well Hannah: <file_gif> Amanda: Don't be shy, he's very nice Hannah: If you say so.. Hannah: I'd rather you texted him Amanda: Just text him ☺ Hannah: Urgh.. Alright Hannah: Bye Amanda: Bye bye	Hannah wants Amanda to give her Betty's number to Larry. Larry called Betty last time they were at the park together.	

Figure 5: The excel file for annotation. Annotators decide whether to use it according to their needs.

SAMSum					
BART as the pre-trained model					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
FactCC	0.2399	0.2365	0.2408	0.2220	**0.2798**
DAE	0.1776	0.1748	**0.1783**	0.1763	0.1687
QuestEval	0.4803	0.4760	0.4798	**0.4863**	0.4722
BARTScore	-2.5505	-2.6273	-2.5510	**-2.3863**	-2.4609
PEGASUS as the pre-trained model					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
FactCC	0.2349	0.2340	**0.2373**	0.2358	0.2342
DAE	**0.1837**	0.1725	0.1796	0.1812	0.1392
QuestEval	0.4794	0.4748	0.4790	**0.4836**	0.4758
BARTScore	-2.5618	-2.5502	-2.5385	**-2.4945**	-2.4963
T5 as the pre-trained model					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
FactCC	0.2380	0.2374	0.2395	0.2447	**0.2513**
DAE	0.1800	0.1777	0.1812	0.1952	**0.1999**
QuestEval	0.4819	0.4777	0.4824	0.4814	**0.4851**
BARTScore	-2.5373	-2.5528	-2.5350	-2.4418	**-2.4274**

DialogSum					
BART as the pre-trained model					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
FactCC	0.1381	0.1298	0.1410	**0.2396**	0.2238
DAE	0.0754	0.0958	0.0883	**0.1094**	0.1050
QuestEval	0.3757	**0.3775**	0.3764	0.3687	0.3647
BARTScore	-2.7102	-2.7467	-2.7283	**-2.2739**	-2.4208
PEGASUS as the pre-trained model					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
FactCC	0.1366	0.1287	0.1348	0.1787	**0.2142**
DAE	0.0890	0.0912	0.0967	**0.1029**	0.0854
QuestEval	0.3758	0.3774	**0.3782**	0.3756	0.3563
BARTScore	-2.5409	-2.6986	-2.6993	**-2.2585**	-2.4360
T5 as the pre-trained model					
Type	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
FactCC	0.1479	0.1289	0.1322	0.1708	**0.1736**
DAE	0.0877	0.0921	0.0933	**0.1079**	0.1033
QuestEval	0.3759	0.3774	**0.3780**	0.3759	0.3723
BARTScore	-2.5735	-2.7000	-2.6973	**-2.2300**	-2.2905

Table 17: Performance (factuality metrics) of different training modes on SAMSum and DialogSum. The best results under the same pre-trained model are bolded. The data augmentation approach is set to **rule**. The pseudo data for the two datasets are constructed separately.

BART as the pre-trained model															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$
M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.74	9.30	**2.08**	1.77	4.65	2.02
R	6.38	1.83	4.26	16.00	2.44	7.58	26.47	5.49	**15.00**	2.03	6.10	2.34	1.25	3.66	1.44
U	2.50	1.82	7.04	6.25	5.45	6.07	15.62	9.09	**13.66**	3.40	9.09	3.89	3.98	14.55	4.66
Total	7.27	1.53	4.15	7.78	2.67	5.63	20.29	5.34	**13.01**	2.18	7.25	2.54	2.02	6.11	2.33
PEGASUS as pre-trained models															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$
M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.28	2.33	1.41	1.47	2.33	**1.59**
R	22.58	4.27	12.15	2.63	0.61	1.58	21.95	5.49	**13.72**	2.31	4.88	2.58	4.20	8.54	4.68
U	33.33	1.82	**7.46**	4.17	3.64	4.05	25.00	1.82	7.04	1.00	3.64	1.17	3.65	12.73	4.25
Total	20.00	3.05	9.48	2.75	1.15	2.15	20.00	3.82	**10.82**	1.76	4.20	1.99	3.71	8.40	4.18
T5 as the pre-trained model															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$
M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.76	2.33	**0.88**
R	16.67	4.27	10.54	0.00	0.00	0.00	25.00	6.71	**16.18**	3.23	5.49	3.52	4.07	10.98	4.66
U	50.00	1.82	7.94	50.00	5.45	18.99	57.14	7.27	**24.10**	5.42	16.36	6.26	7.09	16.36	7.99
Total	17.02	3.05	8.89	21.43	1.15	4.72	27.78	5.73	**15.69**	3.36	6.87	3.74	4.00	10.69	4.57

Table 18: Performance (FERRANTI: form-based categories) of different training modes on SAMSum. The best $F_{0.5}$ scores under the same pre-trained model are bolded. The data augmentation approach is set to **rule**.

BART as the pre-trained model															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
M	0	4	43	0	17	43	0	3	43	4	226	39	2	111	41
R	3	44	161	4	21	160	9	25	155	10	483	154	6	474	158
U	1	3	54	3	45	52	5	27	50	5	142	50	8	193	47
Total	4	51	258	7	83	255	14	55	248	19	851	243	16	778	246
PEGASUS as pre-trained models															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
M	0	6	43	0	23	43	0	5	43	1	77	42	1	67	42
R	7	24	157	1	37	163	9	32	155	8	339	156	14	319	150
U	1	2	54	2	46	53	1	3	54	2	198	53	7	185	48
Total	8	32	254	3	106	259	10	40	252	11	614	251	22	571	240
T5 as the pre-trained model															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
M	0	3	43	0	2	43	0	3	43	0	91	43	1	130	42
R	7	35	157	0	6	164	11	33	153	9	270	155	18	424	146
U	1	1	54	3	3	52	4	3	51	9	157	46	9	118	46
Total	8	39	254	3	11	259	15	39	247	18	518	244	28	672	234

Table 19: Performance (FERRANTI: form-based categories, TP, FP, FN) of different training modes on SAMSum. The data augmentation approach is set to **rule**.

BART as the pre-trained model															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$
M	11.11	1.82	**5.49**	0.00	0.00	0.00	0.00	0.00	0.00	0.66	1.82	0.75	2.13	3.64	2.32
R	4.32	1.96	**3.48**	3.57	0.56	1.72	3.70	0.56	1.74	1.43	2.79	1.58	1.20	2.51	1.34
U	20.00	4.65	**12.05**	3.95	6.98	4.32	4.55	4.65	4.57	2.52	13.95	3.02	2.03	11.63	2.43
Total	5.52	2.19	**4.24**	3.50	1.10	2.43	3.92	0.88	2.31	1.56	3.73	1.76	1.47	3.51	1.66
PEGASUS as pre-trained models															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$
M	66.67	3.64	**14.93**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	1.82	0.78
R	18.97	3.07	**9.32**	14.63	1.68	5.75	17.32	1.12	4.44	1.88	3.91	2.10	1.95	4.47	2.19
U	25.00	4.65	**13.33**	4.35	2.33	3.70	0.00	0.00	0.00	2.37	13.95	2.84	1.18	6.98	1.41
Total	21.74	3.29	**10.25**	9.09	1.54	4.58	13.79	0.88	3.50	1.74	4.39	1.98	1.64	4.39	1.87
T5 as the pre-trained model															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$	**P**	**R**	$F_{0.5}$
M	40.00	3.64	**13.33**	0.00	0.00	0.00	0.00	0.00	0.00	1.26	3.64	1.45	3.17	3.64	3.26
R	12.35	2.79	7.33	33.33	0.28	1.35	43.75	1.96	**8.29**	2.21	4.47	2.46	4.27	4.19	4.26
U	16.67	2.33	7.46	50.00	4.65	16.95	66.67	4.65	**18.18**	2.80	16.28	3.36	3.67	9.30	4.18
Total	14.13	2.85	7.89	25.00	0.66	2.98	42.86	1.97	**8.33**	2.20	5.48	2.50	4.02	4.61	4.12

Table 20: Performance (FERRANTI: form-based categories) of different training modes on DialogSum. The best $F_{0.5}$ scores under the same pre-trained model are bolded. The data augmentation approach is set to **rule**.

BART as the pre-trained model															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
M	1	8	54	0	11	55	0	4	55	1	151	54	2	92	53
R	7	155	351	2	54	356	2	52	356	10	691	348	9	741	349
U	2	8	41	3	73	40	2	42	41	6	232	37	5	241	38
Total	10	171	446	5	138	451	4	98	452	17	1074	439	16	1074	440
PEGASUS as pre-trained models															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
M	2	1	53	0	13	55	0	2	55	0	154	55	1	145	54
R	11	47	347	6	35	352	4	19	354	14	729	344	16	806	342
U	2	6	41	1	22	42	0	4	43	6	247	37	3	252	40
Total	15	54	441	7	70	449	4	25	452	20	1130	436	20	1203	436
T5 as the pre-trained model															
	Pseudo			**Real**			**Pseudo+Real**			**RefS**			**Pseudo+RefS**		
Type	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
M	2	3	53	0	5	55	0	2	55	2	157	53	2	61	53
R	10	71	348	1	2	357	7	9	351	16	709	342	15	336	343
U	1	5	42	2	2	41	2	1	41	7	243	36	4	105	39
Total	13	79	443	3	9	453	9	12	447	25	1109	431	21	502	435

Table 21: Performance (FERRANTI: form-based categories, TP, FP, FN) of different training modes on DialogSum. The data augmentation approach is set to **rule**.

	BART as the pre-trained model					PEGASUS as the pre-trained model					T5 as the pre-trained model				
	Pseudo	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
Ent:ObjE	2.02	9.09	**10.59**	4.79	3.29	14.71	2.40	**16.83**	3.91	8.17	14.34	0.00	**16.20**	4.46	6.20
Ent:AttrE	0.00	0.00	0.00	**5.81**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:ModE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:TensE	-	-	-	0.00	0.00	-	-	-	0.00	0.00	-	-	-	0.00	0.00
Pred:NegE	0.00	0.00	0.00	0.00	**21.74**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**33.33**
Pred:VerbE	4.76	3.46	**12.12**	0.50	1.10	0.00	**1.62**	0.00	0.34	2.21	0.00	**14.29**	13.76	2.94	1.88
CircE	0.00	0.00	0.00	5.05	**6.67**	0.00	0.00	0.00	**3.73**	0.00	0.00	0.00	0.00	0.00	**3.16**
CorefE	7.04	10.64	**25.32**	7.54	4.48	0.00	5.05	5.75	**6.61**	**6.61**	7.04	0.00	**24.27**	10.70	10.97
LinkE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NumE	31.25	0.00	**41.67**	0.00	0.00	**41.67**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**15.62**
OthE	-	-	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	4.15	5.63	**13.01**	2.54	2.33	9.48	2.15	**10.82**	1.99	4.18	8.89	4.72	**15.69**	3.74	4.57

Table 22: Performance (FERRANTI: content-based categories) of different training modes on SAMSum. The values are all $F_{0.5}$ scores. The best results under the same pre-trained model are bolded. The data augmentation approach is set to **rule**.

	BART as the pre-trained model					PEGASUS as the pre-trained model					T5 as the pre-trained model				
	Pseudo	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**	**Pseudo**	**Real**	**Pseudo+Real**	**RefS**	**Pseudo+RefS**
Ent:ObjE	**5.17**	2.54	3.90	3.89	4.68	**15.59**	10.99	8.16	5.60	6.05	11.79	4.88	**16.33**	6.51	8.32
Ent:AttrE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**5.05**	0.00
Pred:ModE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**35.71**	0.00	0.00	0.00	0.00	**16.67**	0.00
Pred:TensE	-	-	-	0.00	-	-	0.00	-	0.00	-	-	-	-	0.00	-
Pred:NegE	-	-	-	0.00	-	-	-	-	-	-	-	-	-	0.00	-
Pred:VerbE	0.00	**2.77**	0.00	1.12	0.74	**2.07**	1.20	0.00	0.06	0.14	0.00	**2.15**	**2.15**	1.09	0.80
CircE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CorefE	**12.20**	0.00	11.11	0.00	0.00	0.00	0.00	0.00	**4.27**	**4.27**	0.00	0.00	0.00	0.00	**9.90**
LinkE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NumE	-	0.00	-	0.00	0.00	0.00	-	-	-	-	-	-	-	0.00	-
OthE	**45.45**	0.00	0.00	0.00	26.32	**45.45**	0.00	0.00	0.00	12.82	**45.45**	0.00	0.00	0.00	11.63
Total	**4.24**	2.43	2.31	1.76	1.66	**10.25**	4.58	3.50	1.98	1.87	7.89	2.98	**8.33**	2.50	4.12

Table 23: Performance (FERRANTI: content-based categories) of different training modes on DialogSum. The values are all $F_{0.5}$ scores. The best results under the same pre-trained model are bolded. The data augmentation approach is set to **rule**.

	BART as pre-trained models						PEGASUS as pre-trained models						T5 as pre-trained models					
	Pseudo			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**		
	rule	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**Infill**	**infill-r**	**rule**	**Infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**
FactCC	0.2399	0.2370	0.2426	0.2408	**0.2432**	0.2370	0.2349	0.2362	0.2344	0.2373	0.2323	**0.2386**	0.2380	0.2374	**0.2415**	0.2395	0.2411	0.2395
DAE	0.1776	0.1754	0.1757	**0.1783**	0.1779	0.1737	0.1837	0.1766	0.1805	0.1796	**0.1866**	0.1809	0.1800	0.1800	0.1768	0.1812	**0.1829**	0.1812
QuestEval	0.4803	**0.4808**	0.4774	0.4798	0.4785	0.4778	**0.4794**	0.4793	0.4793	0.4790	0.4789	0.4790	0.4819	0.4784	0.4797	**0.4824**	0.4808	0.4796
BARTScore	-2.5505	-2.5475	-2.5517	-2.5510	**-2.5462**	-2.5581	-2.5618	-2.5581	-2.5644	**-2.5385**	-2.5452	-2.5444	-2.5373	-2.5484	-2.5463	**-2.5350**	-2.5464	-2.5475

Table 24: Performance (factuality metrics) of different data augmentation approaches on SAMSum. The best results under the same pre-trained model are bolded.

	BART as pre-trained models						PEGASUS as pre-trained models						T5 as pre-trained models					
	Pseudo			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**		
	rule	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**Infill**	**infill-r**	**rule**	**Infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**
FactCC	0.1381	0.1397	0.1347	**0.1410**	0.1298	0.1327	0.1366	0.1297	0.1305	0.1348	**0.1379**	0.1364	**0.1479**	0.1309	0.1309	0.1322	0.1310	0.1297
DAE	0.0754	0.0838	0.0858	0.0883	0.0954	**0.0996**	0.0890	0.0879	0.0883	**0.0967**	0.0958	0.0946	0.0877	0.0921	0.0921	**0.0933**	0.0896	0.0921
QuestEval	0.3757	**0.3774**	0.3765	0.3764	0.3765	0.3771	0.3758	**0.3795**	0.3783	0.3782	0.3788	0.3786	0.3759	0.3780	0.3774	0.3780	0.3777	**0.3783**
BARTScore	-2.7102	**-2.6945**	-2.7290	-2.7283	-2.7231	-2.7106	**-2.5409**	-2.6548	-2.6398	-2.6993	-2.6295	-2.6284	**-2.5735**	-2.6996	-2.6982	-2.6973	-2.6974	-2.6948

Table 25: Performance (factuality metrics) of different data augmentation approaches on DialogSum. The best results under the same pre-trained model are bolded.

		BART as the pre-trained model						PEGASUS as the pre-trained model						T5 as the pre-trained model					
		Pseudo			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**		
		rule	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**Infill**	**infill-r**	**rule**	**Infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**
Ent:ObjE	**P**	2.78	22.22	7.41	15.15	13.79	15.38	24.00	9.09	18.75	26.92	25.00	23.08	20.00	20.00	10.00	25.00	10.00	14.29
	R	0.96	3.85	1.92	4.81	3.85	3.85	5.77	1.92	2.88	6.73	3.85	2.88	6.73	0.96	0.96	6.73	0.96	2.56
	$\mathrm{F}_{0.5}$	2.02	**11.36**	4.72	10.59	9.09	9.62	14.71	5.21	8.93	**16.83**	11.90	9.62	14.34	4.03	3.47	**16.20**	3.47	7.46
Ent:AttrE	**P**	100.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00	0.00	100.00	100.00	100.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\mathrm{F}_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:ModE	**P**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\mathrm{F}_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:TensE	**P**	-	-	0.00	0.00	-	-	-	0.00	0.00	-	0.00	-	-	-	-	-	-	-
	R	-	-	100.00	100.00	-	-	-	100.00	100.00	-	100.00	-	-	-	-	-	-	-
	$\mathrm{F}_{0.5}$	-	-	0.00	0.00	-	-	-	0.00	0.00	-	0.00	-	-	-	-	-	-	-
Pred:NegE	**P**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\mathrm{F}_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:VerbE	**P**	16.67	3.70	3.85	19.05	14.29	16.67	0.00	7.69	5.26	0.00	33.33	14.29	0.00	0.00	0.00	42.86	27.27	50.00
	R	1.23	1.23	1.23	4.94	2.47	4.94	0.00	1.23	1.23	0.00	2.47	1.23	0.00	0.00	0.00	3.70	3.70	3.70
	$\mathrm{F}_{0.5}$	4.76	2.65	2.70	**12.12**	7.30	11.30	0.00	3.76	3.18	0.00	**9.52**	4.59	0.00	0.00	0.00	13.76	12.00	**14.29**
CircE	**P**	0.00	0.00	0.00	0.00	100.0	0.00	0.00	0.00	100.0	0.00	100.0	100.0	100.0	0.00	0.00	0.00	0.00	100.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\mathrm{F}_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CorefE	**P**	12.50	0.00	0.00	40.00	15.38	12.50	0.00	0.00	0.00	8.33	11.11	25.00	12.50	0.00	0.00	31.25	9.09	14.29
	R	2.560	0.00	0.00	10.26	5.13	2.56	0.00	0.00	0.00	2.56	2.56	7.69	2.56	0.00	0.00	12.82	2.56	2.56
	$\mathrm{F}_{0.5}$	7.04	0.00	0.00	**25.32**	10.99	7.04	0.00	0.00	0.00	5.75	6.67	**17.24**	7.04	0.00	0.00	**24.27**	6.02	7.46
LinkE	**P**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\mathrm{F}_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NumE	**P**	33.33	100.00	100.00	50.00	0.00	100.00	40.00	0.00	0.00	50.00	0.00	0.00	0.00	100.00	100.00	0.00	100.00	100.00
	R	25.00	0.00	0.00	25.00	0.00	0.00	50.00	0.00	0.00	50.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\mathrm{F}_{0.5}$	31.25	0.00	0.00	**41.67**	0.00	0.00	41.67	0.00	0.00	**50.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OthE	**P**	-	0.00	-	-	-	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R	-	100.00	-	-	-	-	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	$\mathrm{F}_{0.5}$	-	0.00	-	-	-	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	**P**	7.27	9.26	4.76	20.29	13.79	15.00	20.00	6.38	8.51	20.00	18.92	19.44	17.02	7.14	3.03	27.78	14.71	20.83
	R	1.53	1.91	1.15	5.34	3.05	3.44	3.05	1.15	1.53	3.82	2.67	2.67	3.05	0.38	0.38	5.73	1.91	1.91
	$\mathrm{F}_{0.5}$	4.15	5.23	2.92	**13.01**	8.10	8.96	9.48	3.33	4.44	**10.82**	8.54	8.62	8.89	1.57	1.27	**15.69**	6.28	6.98

Table 26: Performance (FERRANTI: content-based categories) of different data augmentation approaches on SAMSum. The best $\mathrm{F}_{0.5}$ scores under the same pre-trained model are bolded.

		BART as the pre-trained model						PEGASUS as the pre-trained model						T5 as the pre-trained model					
		Pseudo			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**		
		rule	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**
Ent:ObjE	**TP**	1	4	2	5	4	4	6	2	3	7	4	3	7	1	1	7	1	1
	FP	35	14	25	28	25	22	19	20	13	19	12	10	28	4	9	21	9	9
	FN	103	100	102	99	100	100	98	102	101	97	100	101	97	103	103	97	103	103
Ent:AttrE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	0	1	1	0	1	1	1	3	2	1	2	1	0	0	1	0	0	0
	FN	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
Pred:ModE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pred:TensE	**TP**	-	-	0	0	-	-	-	0	0	-	0	-	-	-	-	-	-	-
	FP	-	-	1	1	-	-	-	2	2	-	1	-	-	-	-	-	-	-
	FN	-	-	0	0	-	-	-	0	0	-	0	-	-	-	-	-	-	-
Pred:NegE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FN	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Pred:VerbE	**TP**	1	1	1	4	2	4	0	1	1	0	2	1	0	0	0	3	3	3
	FP	5	26	25	17	12	20	3	2	18	3	4	6	2	5	13	4	8	3
	FN	80	80	80	77	79	77	81	80	80	81	79	80	81	81	81	78	78	78
CircE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	2	1	3	2	0	1	2	1	0	3	0	0	0	1	5	1	1	0
	FN	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	4	14
CorefE	**TP**	1	0	0	4	2	1	0	0	0	1	1	3	1	0	0	5	1	1
	FP	7	6	5	6	11	7	3	4	5	11	8	9	7	2	3	11	10	6
	FN	38	39	39	35	37	38	39	39	39	38	38	36	38	39	39	34	38	38
LinkE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FN	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
NumE	**TP**	1	0	0	1	0	0	2	0	0	2	0	0	0	0	0	0	0	0
	FP	2	0	0	1	1	0	3	1	2	2	2	2	1	0	0	1	0	0
	FN	3	4	4	3	4	4	2	4	4	2	4	4	4	4	4	4	4	4
OthE	**TP**	-	0	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0
	FP	-	1	-	-	-	-	1	1	1	1	1	1	1	1	1	1	1	1
	FN	-	0	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0
Total	**TP**	4	5	3	14	8	9	8	3	4	10	7	7	8	1	1	15	5	5
	FP	51	49	60	55	50	51	32	44	43	40	30	29	39	13	32	39	29	19
	FN	258	257	259	248	254	253	254	259	258	252	255	255	254	261	261	247	257	257

Table 27: Performance (FERRANTI: content-based categories, TP, FP, FN) of different data augmentation approaches on SAMSum.

	BART as the pre-trained model						PEGASUS as the pre-trained model						T5 as the pre-trained model					
	Pseudo			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**		
	rule	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**
Ent:ObjE	5.17	**5.82**	4.46	3.90	4.23	5.07	**15.59**	4.22	4.37	8.16	14.87	11.67	11.79	2.49	2.39	**16.33**	13.41	11.67
Ent:AttrE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:ModE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:TensE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.00	-	-	-
Pred:NegE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pred:VerbE	0.00	**2.71**	2.40	0.00	1.31	1.20	**2.07**	1.89	1.81	0.00	0.00	0.00	0.00	0.00	1.81	2.15	**4.15**	2.07
CircE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CorefE	**12.20**	0.00	0.00	11.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LinkE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NumE	-	-	-	-	-	-	0.00	0.00	0.00	-	0.00	0.00	-	-	-	-	-	-
OthE	**45.45**	0.00	9.09	0.00	0.00	0.00	**45.45**	0.00	0.00	0.00	0.00	0.00	**45.45**	0.00	0.00	0.00	0.00	0.00
Total	4.24	**4.28**	3.60	2.31	2.63	3.00	**10.25**	2.22	2.40	3.50	6.58	5.24	7.89	0.97	1.80	**8.33**	7.65	5.99

Table 28: Performance (FERRANTI: content-based categories) of different data augmentation approaches on DialogSum. The values are all $F_{0.5}$ scores. The best results under the same pre-trained model are bolded. The pseudo-data corpus is DialogSum. The pseudo-data corpus is DialogSum. Please see Table 29 and Table 30 in Appendix E for precision, recall, or TP, etc.

		BART as the pre-trained model						PEGASUS as the pre-trained model						T5 as the pre-trained model					
		Pseudo			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**		
		rule	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**Infill**	**infill-r**	**rule**	**Infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**
Ent:ObjE	**P**	5.30	6.48	5.10	5.56	5.26	6.17	20.97	11.76	13.33	21.05	32.00	27.27	14.12	12.50	10.00	42.11	30.43	27.27
	R	4.73	4.14	2.96	1.78	2.37	2.96	7.69	1.18	1.18	2.37	4.73	3.55	7.10	0.59	0.59	4.73	4.14	3.55
	$F_{0.5}$	5.17	**5.82**	4.46	3.90	4.23	5.07	**15.59**	4.22	4.37	8.16	14.87	11.67	11.79	2.49	2.39	**16.33**	13.41	11.67
Ent:AttrE	**P**	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00	100.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$F_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:ModE	**P**	100.00	0.00	100.00	100.00	100.00	100.00	100.00	0.00	0.00	100.00	100.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$F_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pred:TensE	**P**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.00	-	-	-
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100.00	-	-	-
	$F_{0.5}$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.00	-	-	-
Pred:NegE	**P**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	$F_{0.5}$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pred:VerbE	**P**	0.00	5.71	4.26	0.00	2.63	2.13	33.33	11.11	8.33	0.00	0.00	0.00	0.00	0.00	8.33	100.00	66.67	33.33
	R	0.00	0.87	0.87	0.00	0.44	0.44	0.44	0.44	0.44	0.00	0.00	0.00	0.00	0.00	0.44	0.44	0.87	0.44
	$F_{0.5}$	0.00	**2.71**	2.40	0.00	1.31	1.20	**2.07**	1.89	1.81	0.00	0.00	0.00	0.00	0.00	1.81	2.15	**4.15**	2.07
CircE	**P**	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00	0.00	0.00	0.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$F_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CorefE	**P**	33.33	0.00	0.00	25.00	0.00	100.00	100.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	100.00
	R	3.45	0.00	0.00	3.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$F_{0.5}$	**12.20**	0.00	0.00	11.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LinkE	**P**	0.00	100.00	100.00	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$F_{0.5}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NumE	**P**	-	-	-	-	-	-	0.00	0.00	0.00	-	0.00	0.00	-	-	-	-	-	-
	R	-	-	-	-	-	-	100.00	100.00	100.00	-	100.00	100.00	-	-	-	-	-	-
	$F_{0.5}$	-	-	-	-	-	-	0.00	0.00	0.00	-	0.00	0.00	-	-	-	-	-	-
OthE	**P**	50.00	100.00	7.69	0.00	0.00	0.00	50.00	100.00	100.00	100.00	0.00	0.00	50.00	100.00	100.00	100.00	0.00	0.00
	R	33.33	0.00	33.33	0.00	0.00	0.00	33.33	0.00	0.00	0.00	0.00	0.00	33.33	0.00	0.00	0.00	0.00	0.00
	$F_{0.5}$	**45.45**	0.00	9.09	0.00	0.00	0.00	**45.45**	0.00	0.00	0.00	0.00	0.00	**45.45**	0.00	0.00	0.00	0.00	0.00
Total	**P**	5.52	6.04	4.88	3.92	4.03	4.41	21.74	5.45	7.14	13.79	21.05	20.69	14.13	6.67	8.00	42.86	27.27	21.88
	R	2.19	1.97	1.75	0.88	1.10	1.32	3.29	0.66	0.66	0.88	1.75	1.32	2.85	0.22	0.44	1.97	1.97	1.54
	$F_{0.5}$	4.24	**4.28**	3.60	2.31	2.63	3.00	**10.25**	2.22	2.40	3.50	6.58	5.24	7.89	0.97	1.80	**8.33**	7.65	5.99

Table 29: Performance (FERRANTI: content-based categories) of different data augmentation approaches on DialogSum. The best $F_{0.5}$ scores under the same pre-trained model are bolded.

		BART as the pre-trained model						PEGASUS as the pre-trained model						T5 as the pre-trained model					
		Pseudo			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**			**Pseudo**			**Pseudo+Real**		
		rule	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**	**rule**	**infill**	**infill-r**
Ent:ObjE	**TP**	8	7	5	3	4	5	13	2	2	4	8	6	12	1	1	8	7	6
	FP	143	101	93	51	72	76	49	15	13	15	17	16	73	7	9	11	16	16
	FN	161	162	164	166	165	164	156	167	167	165	161	163	157	168	168	161	162	163
Ent:AttrE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	1	2	2	5	2	2	0	5	4	0	0	0	0	0	0	0	2	0
	FN	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Pred:ModE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	0	2	0	0	0	0	0	1	2	0	0	1	0	0	0	0	0	0
	FN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Pred:TensE	**TP**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-
	FP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
	FN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-
Pred:NegE	**TP**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	FP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	FN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pred:VerbE	**TP**	0	2	2	0	1	1	2	1	1	0	0	0	0	0	1	1	2	1
	FP	16	33	45	29	37	46	2	8	11	6	3	3	3	6	11	0	1	2
	FN	229	227	227	229	228	228	228	228	228	229	229	229	229	229	228	228	227	228
CircE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	7	1	1	2	3	4	0	0	1	1	2	1	0	0	1	0	0	4
	FN	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
CorefE	**TP**	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	2	1	3	3	2	0	0	3	1	3	1	0	2	1	1	1	3	0
	FN	28	29	29	28	29	29	29	29	29	29	29	29	29	29	29	29	29	29
LinkE	**TP**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	FP	1	0	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	FN	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
NumE	**TP**	-	-	-	-	-	-	0	0	0	0	0	0	-	-	-	-	-	-
	FP	-	-	-	-	-	-	2	20	7	0	6	1	-	-	-	-	-	-
	FN	-	-	-	-	-	-	0	0	0	3	0	0	-	-	-	-	-	-
OthE	**TP**	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	FP	1	0	12	7	1	1	1	0	0	0	1	1	1	0	0	0	2	3
	FN	2	3	2	3	3	3	2	3	3	3	3	3	2	3	3	3	3	3
Total	**TP**	10	9	8	4	5	6	15	3	3	4	8	6	13	1	2	9	9	7
	FP	171	140	156	98	119	130	54	52	39	25	30	23	79	14	23	12	24	25
	FN	446	447	448	452	451	450	441	453	453	452	448	450	443	455	454	447	447	449

Table 30: Performance (FERRANTI: content-based categories, TP, FP, FN) of different data augmentation approaches on DialogSum.

SAMSum							**DialogSum**						
BART as the pre-trained model							BART as the pre-trained model						
	Pseudo			**Pseudo+Real**				**Pseudo**			**Pseudo+Real**		
	rule	**infill**	**inill-r**	**rule**	**infill**	**inill-r**		**rule**	**infill**	**inill-r**	**rule**	**infill**	**inill-r**
M	0.00	0.00	0.00	0.00	0.00	0.00	**M**	5.49	0.00	**10.10**	0.00	0.00	5.26
R	4.26	4.63	1.36	**15.00**	9.62	8.12	**R**	3.48	**3.60**	2.36	1.74	2.15	2.11
U	7.04	11.49	13.33	**13.66**	7.41	13.25	**U**	**12.05**	9.68	5.99	4.57	5.85	5.13
Total	4.15	5.23	2.92	**13.01**	8.10	8.96	**Total**	4.24	**4.28**	3.60	2.31	2.63	3.00
PEGASUS as the pre-trained model							PEGASUS as the pre-trained model						
	Pseudo			**Pseudo+Real**				**Pseudo**			**Pseudo+Real**		
	rule	**infill**	**inill-r**	**rule**	**infill**	**inill-r**		**rule**	**infill**	**inill-r**	**rule**	**infill**	**inill-r**
M	0.00	0.00	0.00	0.00	**7.46**	0.00	**M**	**14.93**	0.00	0.00	0.00	0.00	0.00
R	12.15	1.76	1.89	**13.72**	5.68	9.06	**R**	**9.32**	2.20	2.20	4.44	8.89	7.04
U	7.46	11.49	**15.79**	7.04	18.99	13.33	**U**	**13.33**	3.40	5.75	0.00	0.00	0.00
Total	9.48	3.33	4.44	**10.82**	8.54	8.62	**Total**	**10.25**	2.22	2.40	3.50	6.58	5.24
T5 as the pre-trained model							T5 as the pre-trained model						
	Pseudo			**Pseudo+Real**				**Pseudo**			**Pseudo+Real**		
	rule	**infill**	**inill-r**	**rule**	**infill**	**inill-r**		**rule**	**infill**	**inill-r**	**rule**	**infill**	**inill-r**
M	0.00	0.00	0.00	0.00	0.00	0.00	**M**	**13.33**	0.00	0.00	0.00	0.00	0.00
R	10.54	0.00	0.00	**16.18**	1.87	2.16	**R**	7.33	1.26	2.35	**8.29**	7.71	5.61
U	7.94	7.94	7.04	24.10	24.10	**26.67**	**U**	7.46	0.00	0.00	**18.18**	14.93	13.33
Total	8.89	1.57	1.27	**15.69**	6.28	6.98	**Total**	7.89	0.97	1.80	**8.33**	7.65	5.99

Table 31: Performance (FERRANTI: form-based categories) of different data augmentation approaches on SAMSum and DialogSum. The values are all $F_{0.5}$ scores. The best results under the same pre-trained model are bolded. The pseudo data for the two datasets are constructed separately.

SAMSum					**DialogSum**				
Pre-trained Models		BART	BERT	RoBERTa	Pre-trained Models		BART	BERT	RoBERTa
Ent:ObjE	**TP**	1	0	1	**Ent:ObjE**	**TP**	1	2	2
	FP	20	28	36		**FP**	25	20	12
	FN	103	104	103		**FN**	168	167	167
Ent:AttrE	**TP**	0	0	0	**Ent:AttrE**	**TP**	0	0	0
	FP	0	2	0		**FP**	1	2	1
	FN	6	6	6		**FN**	7	7	7
Pred:ModE	**TP**	0	0	0	**Pred:ModE**	**TP**	0	0	0
	FP	0	0	0		**FP**	0	0	0
	FN	1	1	1		**FN**	2	2	2
Pred:TensE	**TP**	-	-	-	**Pred:TensE**	**TP**	-	-	-
	FP	-	-	-		**FP**	-	-	-
	FN	-	-	-		**FN**	-	-	-
Pred:NegE	**TP**	0	0	0	**Pred:NegE**	**TP**	-	-	-
	FP	0	0	0		**FP**	-	-	-
	FN	7	7	7		**FN**	-	-	-
Pred:VerbE	**TP**	0	0	0	**Pred:VerbE**	**TP**	0	0	0
	FP	0	0	0		**FP**	1	0	1
	FN	81	81	81		**FN**	229	229	229
CircE	**TP**	0	0	0	**CircE**	**TP**	0	0	0
	FP	0	0	3		**FP**	1	1	1
	FN	14	14	14		**FN**	5	5	5
CorefE	**TP**	0	0	0	**CorefE**	**TP**	0	0	0
	FP	0	0	0		**FP**	0	0	0
	FN	39	39	39		**FN**	29	29	29
LinkE	**TP**	0	0	0	**LinkE**	**TP**	0	0	0
	FP	0	0	0		**FP**	1	0	0
	FN	6	6	6		**FN**	12	12	12
NumE	**TP**	0	0	0	**NumE**	**TP**	-	-	-
	FP	2	1	4		**FP**	-	-	-
	FN	4	4	4		**FN**	-	-	-
OthE	**TP**	-	-	-	**OthE**	**TP**	0	0	0
	FP	-	-	-		**FP**	0	0	0
	FN	-	-	-		**FN**	3	3	3
Total	**TP**	1	0	1	**Total**	**TP**	1	2	2
	FP	22	31	43		**FP**	29	23	15
	FN	261	262	261		**FN**	455	454	454

Table 32: Performance (FERRANTI: content-based categories, TP, FP, FN) of CCGS on SAMSum and DialogSum.

SAMSum					**DialogSum**				
Spacy Models		sm	md	lg	Spacy Models		sm	md	lg
Ent:ObjE	**TP**	3	2	5	**Ent:ObjE**	**TP**	1	2	1
	FP	164	119	117		**FP**	156	211	295
	FN	101	102	99		**FN**	168	167	168
Ent:AttrE	**TP**	0	0	0	**Ent:AttrE**	**TP**	0	0	0
	FP	5	4	5		**FP**	11	15	19
	FN	6	6	6		**FN**	7	7	7
Pred:ModE	**TP**	0	0	0	**Pred:ModE**	**TP**	0	0	0
	FP	0	0	0		**FP**	0	0	0
	FN	1	1	1		**FN**	2	2	2
Pred:TensE	**TP**	-	-	-	**Pred:TensE**	**TP**	-	-	-
	FP	-	-	-		**FP**	-	-	-
	FN	-	-	-		**FN**	-	-	-
Pred:NegE	**TP**	0	0	0	**Pred:NegE**	**TP**	-	-	-
	FP	0	0	0		**FP**	-	-	-
	FN	7	7	7		**FN**	-	-	-
Pred:VerbE	**TP**	0	0	0	**Pred:VerbE**	**TP**	0	0	0
	FP	9	1	4		**FP**	46	33	34
	FN	81	81	81		**FN**	229	229	229
CircE	**TP**	0	0	0	**CircE**	**TP**	0	0	0
	FP	10	8	7		**FP**	6	6	9
	FN	14	14	14		**FN**	5	5	5
CorefE	**TP**	0	0	0	**CorefE**	**TP**	0	0	0
	FP	3	4	3		**FP**	0	0	0
	FN	39	39	9		**FN**	29	29	29
LinkE	**TP**	0	0	0	**LinkE**	**TP**	0	1	0
	FP	0	0	0		**FP**	0	10	0
	FN	6	6	6		**FN**	12	11	12
NumE	**TP**	0	0	0	**NumE**	**TP**	-	-	-
	FP	5	2	2		**FP**	-	-	-
	FN	4	4	4		**FN**	-	-	-
OthE	**TP**	-	0	0	**OthE**	**TP**	0	0	0
	FP	-	1	1		**FP**	3	21	7
	FN	-	0	0		**FN**	3	3	3
Total	**TP**	3	2	5	**Total**	**TP**	1	3	1
	FP	196	136	139		**FP**	222	296	364
	FN	259	260	257		**FN**	455	453	455

Table 33: Performance (FERRANTI: content-based categories, TP, FP, FN) of FactPegasus on SAMSum and DialogSum.

SAMSum				**DialogSum**			
Pre-trained Models	BART	BERT	RoBERTa	Pre-trained Models	BART	BERT	RoBERTa
FactCC	0.2325	0.2290	**0.2348**	**FactCC**	0.1273	**0.1281**	**0.1281**
DAE	0.1750	**0.1756**	0.1733	**DAE**	0.0867	**0.0892**	**0.0892**
QuestEval	0.4793	0.4794	**0.4807**	**QuestEval**	0.3793	**0.3802**	0.3799
BARTScore	**-2.7799**	-2.7888	-2.7879	**BARTScore**	-2.9237	-2.9175	**-2.9111**

Table 34: Performance (factuality metrics) of CCGS on SAMSum and DialogSum.

SAMSum				**DialogSum**			
Spacy Models	sm	md	lg	Spacy Models	sm	md	lg
FactCC	**0.2380**	0.2273	0.2277	**FactCC**	**0.2282**	0.2195	0.2278
DAE	0.1550	**0.1625**	0.1604	**DAE**	**0.1042**	0.0819	0.0865
QuestEval	0.4671	0.4689	**0.4726**	**QuestEval**	0.3822	**0.3846**	0.3844
BARTScore	-2.9895	**-2.9477**	-2.9650	**BARTScore**	**-3.0689**	-3.1092	-3.1079

Table 35: Performance (factuality metrics) of FactPegasus on SAMSum and DialogSum.

SAMSum					**DialogSum**				
BART as the pre-trained model					BART as the pre-trained model				
	Pseudo		**Pseudo+Real**			**Pseudo**		**Pseudo+Real**	
corpus	**SAMSum**	**Mix**	**SAMSum**	**Mix**	**corpus**	**DialogSum**	**Mix**	**DialogSum**	**Mix**
M	0.00	0.00	0.00	0.00	**M**	5.49	**5.75**	0.00	0.00
R	4.26	**14.95**	**15.00**	11.63	**R**	3.48	**3.80**	1.74	**2.51**
U	**7.04**	5.26	**13.66**	7.41	**U**	**12.05**	3.50	4.57	**7.69**
Total	4.15	**11.58**	**13.01**	9.36	**Total**	**4.24**	3.91	2.31	**3.47**
PEGASUS as the pre-trained model					PEGASUS as the pre-trained model				
	Pseudo		**Pseudo+Real**			**Pseudo**		**Pseudo+Real**	
corpus	**SAMSum**	**Mix**	**SAMSum**	**Mix**	**corpus**	**DialogSum**	**Mix**	**DialogSum**	**Mix**
M	0.00	0.00	0.00	0.00	**M**	14.93	14.93	0.00	0.00
R	12.15	**14.29**	13.72	**18.07**	**R**	9.32	**9.97**	4.44	**6.49**
U	7.46	7.46	**7.04**	6.33	**U**	13.33	**14.93**	0.00	0.00
Total	9.48	**11.19**	10.82	**13.83**	**Total**	10.25	**10.87**	3.50	**5.14**
T5 as the pre-trained model					T5 as the pre-trained model				
	Pseudo		**Pseudo+Real**			**Pseudo**		**Pseudo+Real**	
corpus	**SAMSum**	**Mix**	**SAMSum**	**Mix**	**corpus**	**DialogSum**	**Mix**	**DialogSum**	**Mix**
M	0.00	0.00	0.00	0.00	**M**	13.33	13.33	0.00	0.00
R	**10.54**	14.53	**16.18**	15.28	**R**	7.33	**9.12**	8.29	**9.13**
U	7.94	7.94	**24.10**	14.93	**U**	7.46	7.46	**18.18**	8.47
Total	8.89	**11.90**	**15.69**	13.49	**Total**	7.89	**9.38**	**8.33**	8.04

Table 36: Performance (FERRANTI: form-based categories) of different pseudo-data corpus on SAMSum and DialogSum. **Mix** means to mix the pseudo data constructed from SAMSum and DialogSum together. The values are all $F_{0.5}$ scores. The better of the two pseudo-data corpus results is bolded. The data augmentation approach is set to **rule**.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the "Limitations" Section.

☑ A2. Did you discuss any potential risks of your work?
Yes, in the "Ethics Statement" Section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, we summarize the main claims and our contributions in the abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes. We use many scientific artifacts in Sections 3, 4, 5, 6. We provide artifacts in Section 3 and Section 5.

☑ B1. Did you cite the creators of artifacts you used?
Yes. We cite the creators in the corresponding sections or appendixes. An URL is provided if possible.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Yes, in the "Ethics Statement" Section.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, in the "Ethics Statement" Section.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Yes, in the "Ethics Statement" Section.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes. We provide the relevant information of the dataset in Section 3 and the information of the toolkit in Section 5 and Appendix.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, in Section 6 and Appendix D.

C ☑ Did you run computational experiments?

Yes, in Section 4 and Section 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes, in Appendix D.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, in Section 6, Appendix B and D.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes, in Section 7.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes, in Section 5, Appendix B and D.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Yes, in Section 3.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Yes, in Section 3 and Appendix A.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Yes, in Section 3 and the "Ethics Statement" Section.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Yes, in Appendix A.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. There is no formal ethics committee in our institution, but our plan was discussed internally. Our data collection adheres to the relevant code of ethics.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Yes, in Appendix A.