Prompt scoring system for dialogue summarization using GPT-3

George P. Prodan and Elena Pelican

Abstract—Recent results in language processing show that language models are capable of performing several natural language tasks without the need of supervised learning. A challenging task for pre-trained language models is dialogue summarization. One way of generating summaries is engineering prompt templates for few-shot training. However, a static approach of creating prompts leads to unreliable outcomes between different classes of dialogues. Focusing on the dialogues structure properties we propose a scoring system to improve the few-shot training performances. We build tuned prompts composed by the highest scored dialogue samples. Our evaluation based on ROUGE scores and human evaluation shows that there is an improvement for the experiments in which we use the score system. All experiments are performed within the framework of the GPT-3 API. We use different engines for comparison. Moreover, the human evaluation we conducted showed that the number of failures decreased by 11% after applying our scoring system.

Index Terms—natural language processing, pre-trained language models, abstractive summarization, prompt tuning.

I. INTRODUCTION

LANGUAGE models have evolved significantly in recent years. Whereas task specific models showed that they can attain very good results only in one direction [1] [2], language models proved that they can handle a variety of NLP tasks without supervised learning. Colossal models such GPT-3 are already used by thousands of developers. Few-shot learning without weights updating [3] is one reason for that, as it makes possible a fast development of applications in several directions (classification, semantic search, content generation, summarization and so forth). GPT series are based on the Transformer architecture [4], which relies on self-attention mechanisms [5] [6].

Dialogue summarization is a challenging problem that can be tackled using state-of-the-art NLP technologies. In this study we analyze the possibility to use such models for abstractive summarization of conversations. The aim is to perform generic and informative single-document summarization [7]. This can be done using GPT-3 in a few-shot setting, meaning that one constructs a prompt consisting of one or more summarized dialogues along with the input dialogue, which is given to the model for completion. The fact that this can be done only by using a few training samples, without changing any of the model weights, has several advantages over the classic fine-tuning: (i) there is no need in creating another model, so we save memory space and time and (ii) even if

there is a lack of corpora for fine-tuning purposes, we can still obtain reliable results based only on prompt tuning.

Online communication platforms and mobile chat applications can implement functionalities based on dialogue summarization. Multiple notifications can be replaced by a few summaries providing a better user experience. There is a massive activity on chat applications nowadays, hence such a feature would match the needs of the users. However, one should take into account that processing an enormous volume of messages using models like GPT-3 will demand large computational costs. This leads to the problem of cost minimization. Another objective of this work is to study the behaviour of summarization performance when we are using less computational resources.

A. Related work

As far as we know, there have not been any solutions explored in this direction (i.e. dialogue summarization prompt tuning). Anyway, improving the few-shot training performances is a problem of interest as there can be instabilities in the GPT-3 performances due to the way prompt is chosen [8]. A solution based on contextual calibration was proposed for tasks such as text classification, fact retrieval or information extraction [9].

There have been investigated fine-tuning approaches which also prove that prompt-based tuning increases the language models performances [10] [11]. It is known that GPT series of models are multitask learners [12], but different tasks demand different prompt tuning approaches. Our focus is to find the best way of tuning the prompts for dialogue summarization. The research in dialogue summarization has started to gain popularity in the last two years [13]. There is an impressive increase of data-sets for the summarization of the chat conversations [14] [15] [16] and several models have been developed [17]–[26].

The main known problem of the dialogue summarization is when the summary provides wrong references [13] [27], distorting the information from the original dialogue, phenomenon also known as *hallucinations* [28], [29].

B. Solution details

We aim to improve the quality of the summaries generated in a few-shot training regime by choosing the best picks for the training samples. For that, we establish a simple but efficient scoring system which takes into account the dialogue content, size and number of active participants in the conversation

G. P. Prodan is with University of Bucharest and Ovidius University of Constanta, Romania.

E. Pelican is with Ovidius University of Constanta, Romania.

Manuscript received April 19, 2005; revised August 26, 2015.

(Section 3) in order to find feature similarities between dialogues. Two similar conversations will have a higher score than two different ones.

Firstly, we calibrate the scoring system configuration in order to achieve the best performances (Section 4). We separate each component and prove its relevance to the system. The scoring system is initialised with a weight distribution based on the initial experiments. The final distribution is established after several variations.

Secondly, we vary the model temperature. We noticed an increase of the performance around 25% for low temperatures. We continued the experiments at a low temperature of 0.25. More precise determinations can be done for each parameter we consider in our evaluations. However, at the moment there is no proof to state that the scoring system configuration we use will behave at its best for any other data-set. Further work will be needed to analyze that.

C. Evaluations

The performance evaluation is based on the ROUGE scores [30]. The established score system is evaluated with respect to the results obtained with random generated prompts (Section 5). We observe an improvement of the ROUGE scores for the summaries generated with the score system (SS) proposed by us. Apart from evaluating the score system, we also tested the summarization behaviour for different GPT engines. We noticed that SS improves the summarization performances for each engine. Lastly, we performed human evaluation and compared the results with those obtained with ROUGE-1 and ROUGE-L F1 metrics.

II. EXPERIMENTAL SETUP

We used the GPT-3 API¹ provided by OpenAI [3] to access the GPT-3 engines (curie, curie-instruct-beta, ada, babbage).

A. Datasets

The data-sets on which we rely in experiments are briefly described. They are large datasets that could be used in the training of the dialogue summarization:

- 1) **SAMSum Corpus** - SCd, *A Human-annotated Dialogue Dataset for Abstractive Summarization* [14] is a dataset published by Samsung researchers in 2019. It contains 16369 messenger conversations created by linguists.
- 2) **DialogSum** - DSd, *A Real-Life Scenario Dialogue Summarization Dataset* [15] is similar to SAMSum Corpus, but spoken conversations are summarized instead. There are 13460 summarized dialogues labeled manually.

We decided to use mainly the SAMSum Corpus dataset (SCd) to drive several experiments, because it is currently the largest provider of conversations similar to those discussed on online chats. These conversations are generally natural dialogues in English. The topics approached are diverse, such that we can find discussions between friends, colleagues, more formal discussions or even some inappropriate language. Another

interesting aspect about SAMSum is the fact that it was created by linguists, who are asked to reproduce common discussions which should reflect reality. The fact that the dialogue samples are not very long (as token count) and the variety of topics are the two main reasons for using SAMSum Corpus in our study.

B. Prompt generation

A preliminary processing takes place for the dialogues considered in the prompt shots selection - the selection pool (SP). We construct the term frequency matrix (TF) and compute the inverse document frequency coefficients (IDF). Also, we retrieve the token count and the number of persons participating in the conversation for each dialogue in SP.

The input dialogue which has to be summarized is also analysed. Consequently, the score system is assigning a score with respect to the input dialogue features for each entry in the selection pool. The dialogues having the largest k scores are picked to form the prompt. The dialogues, along with the reference summaries, are constructing the prompt ordered ascending by the score. The prompt includes the following components: the instruction hint for summarization (first row), the selected examples and the delimiters ("" in our case). We provide an example for a two-shot prompt in Table VII.

C. Implementation details

In our experiments SP consists in the training part of SCd and the tests are taken one by one using input dialogues from the testing part of SCd. The data reduction of the SP is performed before testing. A diagram illustrating the whole process can be visualized in Figure 1. For each SP dialogue we compute and save the features we mentioned earlier (TF, IDF, token count, attendance and ID). The score system is retrieving information about the SP dialogues only by searching through the SP features. Then, we use the IDs of the highest k scored dialogues in order to retrieve the content and the summary. Finally, we construct the prompt by following the order imposed by the scores and generate the summary using the GPT-3 engine. The prompt is engineered in such a way the summary to correspond with the prompt completion. Usually for testing, we repeated this procedure for 50, 100 or 200 runs of the model in the same configuration.

III. SCORING SYSTEM

Three aspects are considered in order to have the best picks of dialogues for the few-shot training of the model: the content, size and attendance. Each of them is characterised by a corresponding score and weight. We establish the weights based on empirical results. The quantities for which an optimal result is sought are the ROUGE coefficients [30], and, also, the scores given by human evaluations.

A. Content

The content evaluation is based on the TF-IDF approach [31]. Using the BERT tokenizer from transformers library [32],

¹<https://beta.openai.com/>

Fig. 1. Diagram illustrating the prompt generation process. The preliminary data reduction is performed only once before using the score system. Dataset has been stored in a JSON file during our experiments. Before handling the prompt to the model (GPT-3 engine) the dialogues are ordered ascending by the score. The model is completing the prompt and, therefore, we receive back the summary.

we determined the tokens distribution for each dialogue and computed the TF-IDF weights,

$$w_{td} = \frac{f_{td}}{f_{md}} \log \frac{N}{n_t}$$

where f_{td} is the token t frequency in dialogue d , f_{md} is the maximum frequency of a token in the dialogue d , N is the total number of dialogues and n_t is the number of dialogues in which we can find token t .

A score is given by the cosine similarity between an input dialogue and a dialogue from the training dataset. Let W_i and W_d be the TF-IDF weights vectors. Then, the content score, κ , will be the cosine similarity of the two vectors.

B. Size

We aim to find dialogues having a similar length to that of the input dialogue. We also can define variables on which we can rely for cost minimisation or faster queries. In these cases, the quality of the selected dialogues from the training data-set would not be the top priority.

Let $\mathcal{S} = A \times d_0$ be the set of dialogue pairs from which we search the training shots. For each selected pair (d, d_0) we compute a length similarity coefficient λ . This coefficient is given by a function f_λ which should obey the following conditions:

- $f_\lambda : \mathcal{S} \rightarrow [0, 1]$
- $|d| = |d_0| \Leftrightarrow f_\lambda(d, d_0) = 1, \forall d \in A$
- $\forall d_r, d_l$ such that $|d_r| - |d_0| = |d_0| - |d_l| > 0 \Rightarrow f_\lambda(d_l, d_0) \geq f_\lambda(d_r, d_0)$

The last condition increases the chances to choose the shorter conversation if there are multiple ones with similar content. There is no reason to increase the number of tokens in the

Fig. 2. SAMSum dataset statistics

query which will be handled to the model if the content score is the same. From the distribution of the token count (Figure 2), one can observe that the short conversations are much more and the probability to find content similarities are high. Thus, in most of the cases we can perform the search only through a pool of shorter conversations.

The maximum score is achieved when we pick a conversation of exactly the same token count. If such a sample is not found, then we try to find another one with a small error. Like in the case of a standard deviation, there can be defined an interval within the samples to be scored high. Indeed, we can choose a standard distribution for scoring, meaning that we got a curve similar to a symmetrical Gaussian,

$$f_\lambda(d, d_0) = C e^{-\frac{|d| - |d_0|}{\sigma}} \quad (1)$$

but which does not help if we are looking to minimize the costs. Shorter conversations help in creating a low token count prompt. Therefore, we will need an asymmetrical curve in order to boost the probability of picking short dialogues. In this work we tested the performance of the asymmetrical double sigmoid (*asd*) function in two configurations: narrow and broad. Different *asd* trends can be obtained by modifying the free parameters c_1 - c_6 ,

$$f_\lambda(d, d_0) = c_1 + \frac{c_2}{1 + \exp\left(-\frac{|d| - |d_0| - c_3 + c_4}{c_5}\right)} \times \left[1 - \frac{1}{1 + \exp\left(-\frac{|d| - |d_0| - c_3 - c_4}{c_6}\right)}\right] \quad (2)$$

The large number of free parameters allowed us to test different shapes and find a good set-up for our experiments.

C. Attendance

The number of participants in a dialogue also is a key factor. As also suggested in [33], better results are obtained when the summarization system takes into account the number of persons involved in the discussion. The conversations between two people are usually easier to summarize than those when there are more participants involved. Thus, we group the conversations of the training data-set in two classes: 2 participants class and 3+ participants class. The attendance score is simply

$$\delta_{id} = \begin{cases} 1 & \text{if the dialogues are in the same class} \\ 0 & \text{otherwise} \end{cases}$$

D. Weights

The final score is given by the weighted sum of the scores:

$$s = w_1 \kappa + w_2 \lambda + w_3 \delta \quad (3)$$

TABLE I
ROUGE SCORES AND AVERAGE TOKEN COUNT FOR DIFFERENT WEIGHT DISTRIBUTIONS

w_1	w_2	w_3	ROUGE-1	ROUGE-L	Avg Token Count
0.5	0.3	0.2	31.9 ± 11.4	22.9 ± 11.4	251.56
0.2	0.6	0.2	30.8 ± 11.4	22.0 ± 9.2	258.84
0.6	0.2	0.2	29.2 ± 11.0	20.7 ± 9.1	252.2
0.4	0.4	0.2	28.4 ± 13.6	21.1 ± 11.6	256.54
0.6	0.3	0.1	30.4 ± 13.6	22.5 ± 11.3	251.08

In creating the tuned prompt, we place the scored dialogues in an ascending order. The last shot should be the one scored highest as it has more impact on the completion returned by the model.

IV. CALIBRATING THE SCORE SYSTEM

A. Weights distribution

Several weights are tested whereas keeping the same GPT-3 configuration. By GPT-3 configuration, we refer only to the engine, temperature and number of shots. The results for 50 tests of GPT-3 curie-instruct-beta model at a temperature of 1 are presented in Table I. Standard deviation is provided: score $\pm$ std. We notice a slight increase in the average token count if the size weight (w_2) is larger than content weight (w_1). This confirms that the probability to find content similarities is higher for shorter conversations. Once again, this is due to the nature of the data-set. To establish if this data-set reflects a real scenario or not is beyond the scopes of this work and requires further investigations.

B. Size similarity function

The function which computes the size coefficient has a significant contribution in cost minimisation. We test three functions: the normal distribution with $\sigma = 20$ tokens, meaning a FWHM of 47.2 tokens (gauss), an asymmetrical double sigmoid having a FWHM of 61.8 tokens (ads narrow), and another asymmetrical double sigmoid having a larger FWHM of 85.6 tokens (ads broad). We find that a broader *ads* function helps for cost minimisation. Also, the performance is not affected significantly. There are very small differences between the ROUGE-1 scores:

- *normal distribution* ($\sigma = 20$ tokens): 229.63 average token count, 41.76 ROUGE-1 F1 average score
- *ads narrow*: 220.76 average token count, 41.84 ROUGE-1 F1 average score
- *ads broad*: 215.96 average token count, 41.9 ROUGE-1 F1 average score

C. Selection pool

The selection pool remains unchanged during the experiments. However, to investigate if the score system results depend on the SP we perform a series of separate experiments including a significant amount of foreign data. The results of these experiments are discussed in the next section. Software applications can be developed using the score system

TABLE II
RUNNING THE MODEL AT DIFFERENT TEMPERATURES

Temperature	ROUGE-1	ROUGE-L
1.00	31.2 ± 12.9	22.8 ± 10.8
0.75	37.0 ± 12.0	28.1 ± 9.0
0.50	38.5 ± 13.4	30.5 ± 12.5
0.25	43.4 ± 13.3	34.6 ± 12.8
0.00	40.7 ± 13.0	32.9 ± 11.2

we propose. To do so, one must consider a way of defining the SP. Different categories of users may use different SP. Depending on the users behaviour, SP will also have to change in time. Consequently, a solution based on dynamic SP can be implemented. Users feedback also can be used in upgrading the SP in two ways: (i) by providing better summaries and (ii) by setting up several preferences. Methods for gathering data to improve or form new SPs could be a continuation of this work.

V. RESULTS AND DISCUSSIONS

In our experiments we use mostly the curie-instruct-beta engine of GPT-3 due to its improved performances with respect to curie engine and cost reasons. To settle the temperature, we vary the temperature using the same score system configuration with $w_1 = 0.5$, $w_2 = 0.3$ and $w_3 = 0.2$ (5:3:2), which is highlighted in Table I. We run the curie-instruct-beta model on 200 two-shot prompts (standard deviation is provided: score $\pm$ std) and we obtain better ROUGE scores for lower temperatures as it is presented in Table II. Therefore, we have conducted the experiments at a low temperature setting (0.25).

A. Performance evaluation of the score system

We experimented further by running at different numbers of prompt tuning shots. As expected, the performance increases as more tuning samples are provided to the prompt. The results at the temperature of 0.25 for ROUGE-1 and ROUGE-L scores can be seen in Figure 3. Multiple sessions of 100 tests have been made in order to evaluate the fluctuations of the ROUGE scores. In the boxplot from Figure 4 we consider a baseline of random generated prompts using samples from the same SP. An improvement is slightly observed in the case of the prompts tuned using the proposed score system (5:3:2 weights distribution, ads broad for size function). If we assume a normal distribution for the ROUGE scores, there are no visible changes of the standard deviations.

B. Monitoring the costs

We showed in the previous section that we can diminish the costs (i.e. token count) by modifying the coefficients of the size scoring function f_λ - which in our case is the asymmetrical double sigmoid. Therefore, we can monitor the costs only by defining the desired settings for the *ads* function. Obviously, there are many other options for the scoring function. If we are looking only to minimize the costs, then we can choose a function for which $f(d, d_0) = 0$ for any d

Fig. 3. Score system performance evaluations (ROUGE-1 and ROUGE-L orange lines) with respect to random generated prompts baseline (blue line). Score system: 5:3:2 ads broad.

Fig. 4. The boxplot of score system ROUGE-1 performances. Score system: 5:3:2 ads broad.

and d_0 such that $|d| > |d_0|$. Moreover, in order to increase the scoring for shorter conversations, we can give up on the condition that the score is maximum when the dialogues have equal token counts. We study the behaviour of two functions: the piecewise function,

$$f(d, d_0) = \begin{cases} 0 & \text{if } |d| > |d_0| \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

and the following function, that can improve the chance in picking similar size conversations depending on the value of the free parameter a ,

$$f(x) = \begin{cases} 1 + ax, & \text{if } x \leq 0 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

We run 100 tests for these functions using SS on 5:3:2 configuration:

- *Piecewise function*: 119.77 average token count
- *Equation 5* ($a = 0.005$): 204.64 average token count

Using a smaller model would also reduce the costs. We evaluated the performances for ada, babbage, curie and curie-instruct-beta engines of GPT-3 with and without using SS. As expected, the results (Figure 5) show that performance decreases significantly when using a smaller model.

Fig. 5. Performance evaluation by ROUGE-1 scores for different GPT-3 engines. Score system (SS): 5:3:2 ads broad.

C. Selection pool variations

We insert data from the DialogSum (DSd) data-set into the current SP, which has been created using the training part of SCd. The new SP consists of 54% samples from SCd (14732), 46% from DSd (12460). We ran the model again for 100 tests with curie-instruct-beta engine at temperature of 0.25 and 5:3:2 ads-broad score system. Also, we run the model only on tests from DSd. The results are shown in Table III. As expected, in the case of one-shot training, the performance is slightly better for a larger SP. However, the SCd only two-shots training scores higher than for the other SPs. Also, there is a large difference in the score provided by SS between DSd experiments and the others. This difference is actually a measure of the discrepancies between SCd and DSd.

D. Human evaluation

For human evaluation (HE) we considered four criteria following [34]: Coherence - it is related to the overall quality of the sentences and how well is the summary structured, Consistency - it measures the factual information transfer, Fluency - it scores the quality of individual sentences (i.e. grammar, formatting and so forth) and Relevance - it shows how important is the selected information, an excess of information is penalised. We asked the annotators to rate different summaries for 100 dialogues on a scale from 1 to 5. The evaluation was blind and it included the reference summary from SAMSum dataset and the summaries generated by GPT-3's curie-instruct-beta (two shots and 5:3:2 low temperature configuration) with and without applying the score system. The results are presented in Table V. The average scores of 3 x 100 dialogue summaries are provided together with the number of dialogues rated between unit intervals. A significant increase due to the use of the scoring system can be observed. For Relevance we can spot the largest difference in the average score. However, an improvement is visible for each criterion. The number of failures (i.e. poor summaries rated under 2.5) decreases with 11% of the total amount of dialogues after applying the scoring system.

TABLE III

THE ROUGE SCORES, AVERAGE SCORES PROVIDED BY OUR SCORE SYSTEM AND TOKEN COUNT FOR DIFFERENT SELECTION POOLS AND NUMBER OF SHOTS USED IN THE TRAINING.

Selection pool	Training	ROUGE-1	ROUGE-L	Avg Score	Token count
SCd	One-Shot	39.0 $\pm$ 13.1	31.5 $\pm$ 13.1	0.645	109.40
SCd + DSd	One-Shot	40.8 $\pm$ 13.1	32.3 $\pm$ 11.6	0.6529	110.55
DSd	One-Shot	40.4 $\pm$ 13.9	32.1 $\pm$ 13.2	0.4635	118.59
SCd	Two-Shot	41.9 $\pm$ 14.3	33.4 $\pm$ 13.5	0.6255	215.96
SCd + DSd	Two-Shot	41.5 $\pm$ 14.4	33.1 $\pm$ 12.3	0.6335	216.34
DSd	Two-Shot	39.1 $\pm$ 15.7	31.6 $\pm$ 15.0	0.4377	242.62

We calculated the Pearson coefficient to evaluate the correlation between human judgment and ROUGE scores (Table VI). We see that the ROUGE scores are not reliable in this case. There are many examples, as the one provided in Table IV, when very good summaries are rated with low ROUGE scores.

VI. CONCLUSION

In this study we investigated possible prompt-based improvements of language models to perform abstractive summarization. We show that choosing the right dialogues can increase the quality of the summaries and reduce the number of failures by 11 %. We test the scoring system for several GPT-3 engines and we obtain better results for each engine when applying the scoring system. Also, the scoring system we proposed for selecting the best picks to create prompts also can control the computational costs by using different size similarity functions. We proved that content similarities between dialogues are also valuable for tuning the prompt, and as a result computational resources can be saved.

We evaluate the scoring system using ROUGE metrics and, also, by conducting human evaluation. It seems that in our experiment small variations in the average ROUGE score correspond to larger discrepancies in the scores given by the annotators. However, both evaluations show that applying the score system increases the quality of the summaries.

By studying the selection pool behaviour, we identified another research direction that can represent a future work. The way one gathers the dialogue samples in the selection pool can lead to different results. It may be necessary to engineer dynamic data-sets of dialogue samples depending on the end users behaviour and individual feedback. Thus, investigating prompt tuning methods that do not require fine-tuning is an advantage as we do not need to use space for distinct modules of a fine-tuned version of the model for each user.

APPENDIX A

TWO SHOT PROMPT EXAMPLE

In Table VII we show one example of a two-shot training to illustrate the prompt structure for summary generation.

ACKNOWLEDGMENT

The authors would like to thank OpenAI for allowing us to access GPT-3 API for academic purposes. We also thank the annotators for their contribution in evaluating the summaries.

REFERENCES

- [1] J. Chai and A. Li, "Deep learning in natural language processing: A state-of-the-art survey," in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2019, pp. 1–6.
- [2] P. Kłosowski, "Deep learning for natural language processing and language modelling," in *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2018, pp. 223–228.
- [3] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>
- [4] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.
- [5] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
- [6] Q. Guo *et al.*, "Low-rank and locality constrained self-attention for sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, p. 2213–2222, Dec. 2019. [Online]. Available: <https://doi.org/10.1109/TASLP.2019.2944078>
- [7] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Systems with Applications*, vol. 121, pp. 49–65, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417418307735>
- [8] Y. Lu *et al.*, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," *arXiv e-prints*, p. arXiv:2104.08786, 2021.
- [9] T. Z. Zhao *et al.*, "Calibrate Before Use: Improving Few-Shot Performance of Language Models," *arXiv e-prints*, p. arXiv:2102.09690, Feb. 2021.
- [10] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *ACL/IJCNLP*, 2021.
- [11] S. Hu *et al.*, "Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification," *arXiv e-prints*, p. arXiv:2108.02035, Aug. 2021.
- [12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.
- [13] X. Feng, X. Feng, and B. Qin, "A survey on dialogue summarization: Recent advances and new frontiers," *arXiv e-prints*, p. arXiv:1711.01731, 2021.
- [14] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "Samsam corpus: A human-annotated dialogue dataset for abstractive summarization," *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 2019. [Online]. Available: <http://dx.doi.org/10.18653/v1/D19-5409>
- [15] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "Dialogsum: A real-life scenario dialogue summarization dataset," in *FINDINGS*, 2021.
- [16] L. Mehnaz *et al.*, "GupShup: An Annotated Corpus for Abstractive Summarization of Open-Domain Code-Switched Conversations," *arXiv e-prints*, p. arXiv:2104.08578, Apr. 2021.
- [17] L. Zhao, W. Xu, and J. Guo, "Improving abstractive dialogue summarization with graph structures and topic words," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 437–449. [Online]. Available: <https://aclanthology.org/2020.coling-main.39>
- [18] X. Feng *et al.*, "Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks," in *Chinese Computational Linguistics*. Cham: Springer International Publishing, 2021, pp. 127–142.

TABLE IV
EXAMPLE OF SUMMARY SCORED LOW BY ROUGE METRICS, BUT EVALUATED AS VERY GOOD BY HUMAN JUDGMENT

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.			
	Source	Summary	R1	RL
	GPT-3 2-shot + SS	Emma is not feeling well and doesn't want to cook. Will is trying to make sure she's ok.	25.80	19.35
	SAMSum	Emma will be home soon and she will let Will know.	100	100
	GPT-3 2-shot	Emma is not hungry and is not sure when she will be home. Will is worried about her.	48.27	34.48

TABLE V
HUMAN EVALUATION RESULTS

Statistics	SAMSum	GPT-3 + SS	GPT-3 random
Coherence	4.23	3.87	3.58
Consistency	4.11	3.53	3.22
Fluency	4.16	3.86	3.61
Relevance	3.97	3.29	2.80
Average Score	4.12	3.64	3.30
Between 4-5	68	47	37
Between 3-4	26	32	34
Between 2-3	15	25	32
Between 1-2	4	10	19
Failures (1 - 2.5)	9	22	33

TABLE VI
PEARSON CORRELATION COEFFICIENTS BETWEEN HUMAN EVALUATION RESULTS AND ROUGE METRICS FOR TWO SET-UPS: GPT-3 WITH SCORING SYSTEM AND WITHOUT IT (RANDOM PROMPTS)

Set-up	Criterion	R1	RL
GPT-3 + SS	Coherence	0.3324	0.3122
	Consistency	0.2860	0.2719
	Fluency	0.2328	0.2454
	Relevance	0.3665	0.3284
	Average	0.3478	0.3296
GPT-3 random	Coherence	0.1423	0.2052
	Consistency	0.0884	0.1415
	Fluency	0.1853	0.2332
	Relevance	0.1693	0.2499
	Average	0.1683	0.2391

- [19] —, “Language model as an annotator: Exploring dialogpt for dialogue summarization,” *ACL 2021*, May 2021.
- [20] W. Chien-Sheng *et al.*, “Controllable abstractive dialogue summarization with sketch supervision,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5108–5122. [Online]. Available: <https://aclanthology.org/2021.findings-acl.454>
- [21] N. Shashi *et al.*, “Planning with Learned Entity Prompts for Abstractive Summarization,” *arXiv e-prints*, p. arXiv:2104.07606, Apr. 2021.
- [22] J. Chen and D. Yang, “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*. Online: Association for Computational Linguistics, Jun. 2021, pp. 1380–1391. [Online]. Available: <https://aclanthology.org/2021.naacl-main.109>
- [23] L. Mike *et al.*, “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*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: <https://aclanthology.org/2020.acl-main.703>
- [24] Z. Jingqing *et al.*, “PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 11 328–11 339. [Online]. Available: <https://proceedings.mlr.press/v119/zhang20ae.html>
- [25] L. Dong *et al.*, “Unified language model pre-training for natural language understanding and generation,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach *et al.*, Eds., 2019, pp. 13 042–13 054. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html>
- [26] Y. Zhang *et al.*, “DIALOGPT : Large-scale generative pre-training for conversational response generation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 270–278. [Online]. Available: <https://aclanthology.org/2020.acl-demos.30>
- [27] J. Chen and D. Yang, “Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization,” in *EMNLP*, 2020.
- [28] H. Yichong *et al.*, “The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey,” *arXiv e-prints*, p. arXiv:2104.14839, Apr. 2021.
- [29] Z. Zheng *et al.*, “Reducing Quantity Hallucinations in Abstractive Summarization,” *arXiv e-prints*, p. arXiv:2009.13312, Sep. 2020.
- [30] C.-Y. Lin, “Rouge: a package for automatic evaluation of summaries,” in *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, July 2004*, pp. 74–81. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries/>
- [31] C. Sammut and G. I. Webb, Eds., *TF-IDF*. Boston, MA: Springer US, 2010, pp. 986–987.
- [32] T. Wolf *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://aclanthology.org/2020.emnlp-demos.6>
- [33] M. H. Bokaei *et al.*, “Summarizing meeting transcripts based on functional segmentation,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 10, p. 1831–1841, Oct. 2016. [Online]. Available: <https://doi.org/10.1109/TASLP.2016.2585859>
- [34] A. R. Fabbri *et al.*, “SummEval: Re-evaluating Summarization Evaluation,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 04 2021. [Online]. Available: https://doi.org/10.1162/tacl_a_00373

TABLE VII
EXAMPLE OF A TWO-SHOT TRAINING TO ILLUSTRATE THE PROMPT STRUCTURE

Item	Prompt
Hint	Summarize conversations
Delimiter	""
Shot 1	<p>Jackson: file-gif Jackson: file-gif Jackson: file-gif Jackson: file-gif Madison: LOL dude stop it, youre being such a little troll right now Jackson: file-gif Jackson: file-gif Jackson: file-gif Madison: STOP lol Jackson: file-gif Jackson: :)</p>
Delimiter	""
Summary 1	Summary: Jackson is sending Madison a lot of gifs.
Delimiter	""
Shot 2	<p>Paul: Hey Rosie ;3 Rosemary: Hey Paolooo Paul: i had so much fun last time we met Paul: i miss you already Rosemary: miss you too Rosemary: file-gif Paul: file-gif Paul: file-gif Paul: yesterday i quit my job Rosemary: wait what? :(Paul: yeah im pissed Paul: they fired most of my friends Paul: so i quit Rosemary: :(Paul: dont worry ive been saving money for long time Paul: i'll find sth soon :ç Rosemary: good luck :O</p>
Delimiter	""
Summary 2	Summary: Paul quit his job yesterday because most of his friends got laid off. He has money saved and will find a new job soon.
Delimiter	""
Dialogue to summarize	<p>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</p>
Delimiter	""
Summary to complete	Summary:

George Pantelimon Prodan received the B.S. degree in Computer Science from the Ovidius University of Constanta, Romania and B.S. degree in Physics from the University of Bucharest, Romania. He is currently pursuing his M.S. studies in Physics of Data at the Department of Physics and Astronomy, University of Padua, Italy. His research interests: image and language processing and their general applications in life data and astronomy.

Elena Pelican is Assoc. Prof. of Optimization, Machine Learning, and Data Mining at the Department of Mathematics and Computer Science, Ovidius University of Constanta, Romania. She earns a PhD from Politehnica University of Bucharest, Romania in 2007, and she is the holder of “2009 Young Researcher of the Year” awarded by the Romanian Industrial and Applied Mathematics Society. Her current research interests are pattern recognition and data mining with applications in natural language processing and computer vision.