Записки научных
семинаров ПОМИ
Том 529, 2023 г.

D. Kosenko, D. Zharikova

KRGP: KNOWLEDGE-BASED RESPONSE GENERATION WITH PERSONA

Abstract. To create a personalized response, a generative model must take into account personal information about the user, question asked, and domain knowledge. Therefore, it is necessary to learn how to extract relevant information that will help the generative model to compose a response to the user. In this work, we propose to split the process into three stages: selection of relevant sentences from the textual knowledge base, selection of the most suitable sentences of the textual persona description taking into account the extracted knowledge, and response generation based on the knowledge and persona. We use the sentence Transformer and adapt the algorithm from the CLIP paper to obtain contextual sentence embeddings to extract the most relevant text spans from the knowledge base. We have found that the focal loss shows better results in tasks of binary classification of a persona using the FoCus imbalanced dataset as an example. We have also shown that text2text Transformer BART performs well in the tasks of conditional response generation in a dialog. This system achieved a state-of-the-art result at the leaderboard of The 1st Workshop on Customized Chat Grounding Persona[1]. The code for this work is available at `https://github.com/dmitrymailk/deeppavlov_focus`.

§1. Introduction

Creating personalities for chatbots makes them more attractive to the target audience. By incorporating brand-specific personality traits and characteristics, we can create a chatbot with a strong personality that will drive customer engagement and satisfaction. In addition, having a persona helps manage the language and tone of the chatbot, which can help businesses communicate more effectively with their customers [1,5,6]. Besides, knowledge-based chatbots can be used for customer support or information retrieval [2, 26]. Combining both persona and knowledge may help to create a chatbot that could provide users with information relevant to them, and at the same time do it in a form that is unique to a brand.

Key words and phrases: knowledge grounding and persona-based generation and knowledge-based generation and dialog systems.

In this work, we propose the KRGP (Knowledge-based Response Generation with Persona) system, which uses external textual knowledge about the object and the textual user's persona description. This model takes into account the context of the use of knowledge about the world, and also evaluates the appropriateness of involving personal information about the user in order to generate a response.

We propose to use three models: Knowledge Extractor, Persona Extractor, and Generation Model. Knowledge Extractor selects the most relevant knowledge about the world based on the whole user's persona and dialog context. Persona Extractor then evaluates the appropriateness of using certain facts about the user from the persona description. And finally, dialog context, extracted persona and extracted knowledge sentences are passed to the Generation Model to generate a human-like response to the user.

Section 2 introduces the most common approaches to persona-based and knowledge-based generation. The utilized dataset is described in Section 3. The detailed description of the proposed system is given in Section 4. The setup of the experiments and the results are presented in Sections 5 and 6.

§2. Related Work

2.1. Knowledge Grounded Conversation. To allow the model to organically add information to the cue and create believable answers, researchers create datasets with open domain dialogs [2, 7, 19]and dialogs with personal recommendations [3]. Researchers then use pre-trained models such as T5 [14] or BART [8] and build dialog systems in end-to-end mode [9, 21]. These models can generate text in response to user questions or comments using information from the knowledge base and context from previous messages in the conversation. Through the use of pre-trained models and large amounts of data, these conversational systems can be quite accurate and natural in their responses.

2.2. Persona Grounded Conversation. To make responses more natural and personalized to the user, the PersonaChat [24] dataset was created; the dataset contains not only dialogs but also personal information about each participant in the conversation. The authors of [24] proposed to use the Profile Memory Network, which takes into account the history of the dialog and information about each participant. However, these models tend to be too self-centered, presenting information about themselves and not taking into account the interests of the interlocutor.

In COSPLAY [23], the authors solve this problem using the Concept Set framework. However, in real-life applications we may not always have a description of the persona of the interlocutor, which requires us to create her profile from available data. The authors pay attention to this problem in the following work [25].

§3. Dataset

The 1st Workshop on Customized Chat Grounding Persona introduced a new dataset FoCus [5] which for each sentence contains information about the utilized persona and general knowledge. This dataset emulates a dialog between a machine and human. FoCus contains 12,484 dialogs, 5,152 Wikipedia facts, and 32,855 user persona sentences. FoCus has an official division into train and validation subsets. In this paper, the sample is a pair of user requests and responses with relevant personal information and knowledge of the outside world. A sample from the dataset is presented in Figure 1.

Each dialog contains a list of sentences with general knowledge and a list of sentences with a user's persona description. On each step of the dialog, different sentences from the general knowledge and persona can be utilized. The dataset provides the sentences utilized for response generation on each step of the dialog.

§4. Methodology

The process of conditional response generation to the user differs depending on the stage, e.g., training or inference. During the training process, we utilize the user's persona and general knowledge sentences originally indicated in the dataset. In the inference process, since we do not know which general knowledge and persona sentences should be utilized, we use the Knowledge Extractor and Persona Extractor, respectively. Both Extractors select sentences from given lists corresponding to each dialog.

4.1. Model Components. KRGP consists of three parts: Knowledge Extractor, Persona Extractor and Generation Model. The Knowledge Extractor is responsible for extracting the most relevant sentences from the list of knowledge which is originally a list of facts from Wikipedia. The Persona Extractor determines how appropriate it is to use user's persona description to fit it seamlessly into the response. The Generation Model

```
{
  # list of facts about the object
  "knowledge": List[str] = [
     “Nazareth House is a heritage-listed benevolent institution at 272...”,
     “In 1982 Nazareth House ceased its function as a care facility for children...”,
     ...
  ],
  # 5 persona sentences
  "persona": List[str]  = [
       "I would like to visit the Nazareth House again.",
       "I have curiosity about the Description of this place.",
       ...
  ],
  # dialogue, 5-7 iterations
  "utterance": [
    {
      # question and answer
      "dialogue": List[str] = [
         “Can you describe this house to me?”,
         “You have curiosity about the description of Nazareth House and I will tell you. Nazareth House is...”
      ],
      # labels of the used person
      "persona_grounding": List[bool] = [true, false, ..., true],
      # copy of sentences from persona
      "persona_candidate": List[str],
      # 10 knowledge candidates, 1 correct, 9 incorrect
      "knowledge_candidates": List[str],
      # correct knowledge index
      "knowledge_answer_index": int,
    },
  ],
}
```

Figure 1. Sample from the FoCus [5] dataset. List of persona sentences is the same for the entire dialog. List of knowledge sentences can vary through the dialog.

uses all the information selected in the previous steps to generate a response to the current dialog context.

Depending on the mode of use, the logic of the algorithm may differ. In Figure 2 we demonstrate a training data processing pipeline. During training all three components utilize the ground truth selection of the knowledge and persona.

During the inference stage, we first utilize the Knowledge Extractor to select the most relevant knowledge sentences among the given list, then pass those selected knowledge sentences to the Persona Extractor that determines which persona sentences may be used for the response generation. At the end, selected knowledge and persona sentences are transferred to

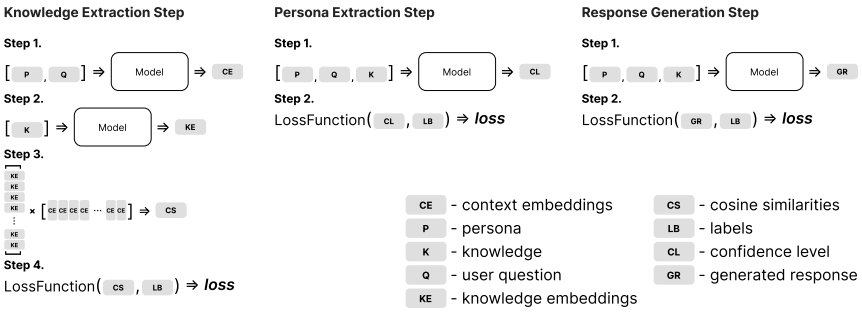

Figure 2. Training pipeline. Knowledge Extraction consists of four steps: (1) computing context embedding from persona description and the user question; (2) computing embeddings for knowledge candidates; (3) computing cosine similarity for the knowledge candidates' embeddings and context embedding; (4) computing the loss with original knowledge candidates' labels. Persona Extraction consists of two steps: (1) computing the confidence score based on the persona, user question and knowledge sentence; (2) computing the loss with the original persona sentences' labels. Response Generation consists of two steps: (1) generating the response based on the persona, knowledge and user question; (2) computing the loss with the original one.

the response Generation Model to generate a response according to the current context.

4.1.1. *Knowledge Extractor.* The dataset contains a list of general knowledge sentences for each dialog which may be used for response generation. On each step of the dialog exactly one knowledge sentence is used, which allows us to reformulate the knowledge extraction problem to finding the knowledge sentence most similar to the dialog context.

In order to build a model that would correctly respond to the context, the original dataset is altered as follows. Each training sample is a pair of two strings, similarly to CLIP [17]. The first one is context, which is a concatenation of the whole user's persona description and the last user

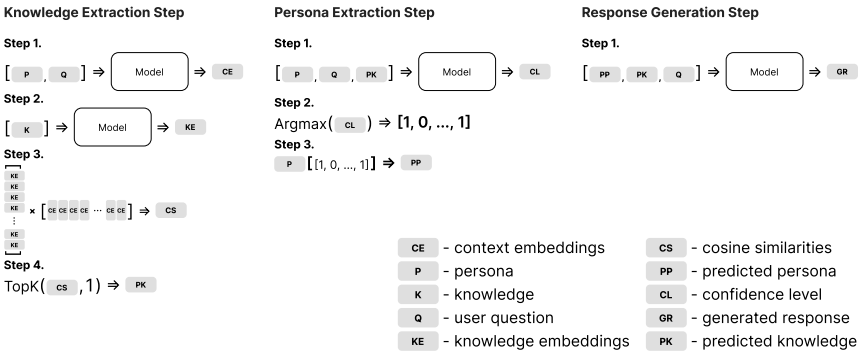

Figure 3. Inference pipeline. Knowledge Extraction consists of four steps: (1) computing the context embedding from persona description and the user question; (2) computing embeddings for knowledge candidates; (3) computing cosine similarity for the knowledge candidates' embeddings and context embedding; (4) select the predicted knowledge candidate that maximizes the cosine similarity. Persona Extraction consists of three steps: (1) computing confidence score based on the persona, user question and knowledge sentence; (2,3) selecting predicted persona sentence with the highest confidence level. Finally, the Response Generation Model generates a response based on the predicted persona, predicted knowledge, and user question.

utterance in the dialog context. The second one is the knowledge sentence which may be utilized for the response generation on the current step.

We want to find the most relevant knowledge sentence among the given list, so we build a model that would assign the highest score to the ground truth pair of context and knowledge. In order to train a ranking model, we used the symmetric cross-entropy loss, similar to the loss utilized by the CLIP model. The cosine distance between vector representations of texts is utilized as a metric of relevance. The algorithm is presented in Algorithm 1, and the scale factor is a hyperparameter significantly affecting the convergence.

Algorithm 1 Algorithm for Training Knowledge Extractor.

$context \leftarrow$ `combination of the persona and last user question`
$knowledge \leftarrow$ `knowledge facts from dataset`
$L \leftarrow$ `length of knowledge`
$scale\_factor \leftarrow 20$
$source\_sentences \leftarrow lm\_model(context)$
$knowledge\_sentences \leftarrow lm\_model(knowledge)$
$labels \leftarrow range(L)$
$scores \leftarrow scale\_factor \times source\_sentences \times knowledge\_sentences$
$loss\_f \leftarrow CrossEntropyLoss()$
$loss \leftarrow \frac{1}{2} \times (loss\_f(scores, labels) + loss\_f(scores^T, labels))$

4.1.2. *Persona Extractor.* In order to make the system's responses look more natural, we propose a model that determines the appropriateness of using facts about the user, e.g., sentences from persona description. On each step any number of persona sentences can be utilized. Therefore, the problem was reformulated as binary classification where sentences belong to class 1 if they should be used for response generation and to class 0 in other cases. On the training stage, the input of the Persona Extractor is the last user utterance and the ground truth utilized knowledge. On the inference stage, we use the knowledge sentence selected by the Knowledge Extractor.

On average, any information about a person in the original dataset was used only in 13% of the cases, so we are dealing with an imbalanced dataset. To solve this problem, we used focal loss [12] as well as a special method of selecting training examples. The strategy is as follows: since the initial dataset contains a predominant number of negative samples, we decided to take one positive and one negative sample from every step of the dialog. If there was no positive example at some step, only one negative sample was added to the training examples. The Focal loss function for two classes was utilized for training with hyperparameters γ equal to 2.0 and α equal to 0.5.

4.1.3. *Generation Model.* We used the original BART architecture to train the response generation model. As an input to the model, we provide concatenated selected persona sentences, selected knowledge sentence, and the last user utterance. As the target sequence, we use the ground truth system's response from the dataset. During the training stage, the ground

truth knowledge and persona sentences were fed to the model, while during the inference stage, the sentences selected by Knowledge and Persona Extractors were used.

§5. Experiments

5.1. Experiment Details. We compared two models from the Hugging Face [22] library for Knowledge and Persona Extractors: small DeBERTaV3 [4] (`microsoft/deberta-v3-small`) with 44M parameters and base MPNet [18] (`sentence-transformers/all-mpnet-base-v2`) with 110M parameters. The model DeBERTaV3 was selected as a state of the art model for the SuperGLUE task but demonstrated low performance on our task. MPNet was picked up as a state of the art model for the sentence embedding task. Persona and Knowledge Extractors were trained for five and four epochs, which took 3.1 and 3.2 hours correspondingly.

The base BART model [8] (`facebook/bart-base-model`) with 140M parameters was used to create the response Generation Model without any architectural changes. For the generative model, batch size of four was used with dialog history of one last utterance. All models were trained on an RTX-3060 12GB. The response Generation Model was trained for four epochs, which took 2.5 hours.

To optimize all three models, we used the AdamW optimizer [13] with a learning rate of $6.25e-5$.

Baselines. As the baselines, we chose the original solution by the dataset's authors [5], the INFO [10] and Proto-Gen [20] models.

5.2. Evaluation. To evaluate the quality of our models, we use the same metrics as in the original FoCus paper. Text generation is evaluated using chrF++ [16], BLEU [15], and ROUGE-L [11]. To assess the quality of extraction, accuracy metric is used for both Knowledge and Persona Extractors.

§6. Results

6.1. Knowledge Extractor. The comparison of different models for Knowledge Extractor is presented in Table 1. Since the task was formulated as a ranking problem, it could be solved by computing the cosine similarity of the context and knowledge embeddings. We tried two different embedding models: small DeBERTaV3 and base MPNet. Our experiments show that `sentence-transformers/all-mpnet-base-v2` performs significantly

Table 1. Knowledge Extractor training results.

Model	**Training**	**Knowledge Accuracy**	**Data Subset**
TF-IDF		61.23	valid
debertav3-small		7.98	valid
debertav3-small	Finetuned	17.68	valid
all-mpnet-base-v2		93.43	valid
all-mpnet-base-v2		93.38	test
all-mpnet-base-v2	Finetuned	98.90	valid
all-mpnet-base-v2	Finetuned	**98.84**	test

Table 2. Persona Extractor training results. CE loss denotes cross-entropy loss.

Model	**Loss**	**Persona Accuracy**	**Data Subset**
predict only zeros (never use persona)	-	86.69	valid
debertav3-small	CE loss	92.12	valid
debertav3-small	Weighted CE loss	91.74	valid
debertav3-small	Weighted CE loss	91.49	test
debertav3-small	Focal loss	92.12	valid
all-mpnet-base-v2	CE loss	92.44	valid
all-mpnet-base-v2	Focal loss	**92.45**	valid
all-mpnet-base-v2	Focal loss	**92.27**	test

better than a similar `microsoft/deberta-v3-small` model. The MPNet model also performs well in unsupervised mode.

6.2. Persona Extractor. The comparison of different models for Persona Extractor is presented in Table 2. The imbalance of the persona selection classes demonstrates that the officially proposed accuracy metric is not the best choice for the task. With accuracy metric, the score is high even if the persona extractor does not predict any persona sentence utilization. According to the results of experiments, Focal loss improves the model based on the `sentence-transformers/all-mpnet-base-v2` which also outperforms extractors based on `debertav3-small`.

Table 3. Table with system comparisons. KRGP denotes our proposed system.

System	Persona Accuracy	Knowledge Accuracy	BLEU	CharF++	ROUGE-L
FOCUS [5]	86.85	65.06	10.87	27.90	30.98
INFO [10]	82.70	**99.24**	31.46	53.29	53.06
Proto-Gen [20]	85.02	85.18	19.85	42.32	38.84
KRGP	**92.27**	98.84	**32.47**	**53.90**	**54.31**

6.3. Response Generation Model. The comparison of our system to the baselines is presented in Table 3. The performance of the model was evaluated on a dataset of dialog responses. We found out that our KRGP approach outperformed other considered methods, including the model from the original paper, and achieved state-of-the-art results in terms of both extractor's accuracy and text evaluation metrics.

6.4. Examples. In Figure 4 we present an example of the KRGP system's predictions. One the first step, we predict the knowledge sentence taking into account all persona sentences and user input. Second, we predict persona sentences (zero, one or several) based on the predicted knowledge and user input. Finally, the model conditioned on predicted knowledge and persona sentences and user input generates a response.

6.5. Effect of extractors on generation. Additionally, we conducted an ablation study to analyze the contribution of extractors in our system to the overall performance. The results of the ablation study are shown in Table 4.

We have analyzed the influence of extractors and searched for the upper bound for our generative model. According to Table 2, we can see that if we have perfect extractors (utilizing true sentences), for a given generative model, we can increase generation scores by an average of 1.71 points. However, Persona Extractor has a significant negative impact on the metrics on a valid subset, which could be explained by poor performance compared to Knowledge Extractor.

§7. Conclusion

In this work, we have presented a Knowledge-based Response Generation with Persona system which extracts most relevant facts from general

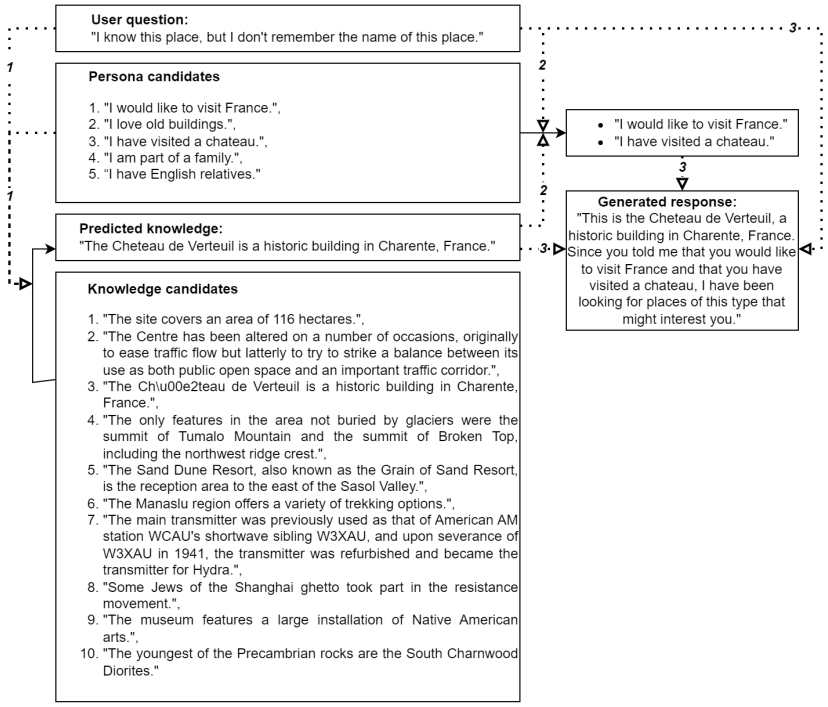

Figure 4. Example of the system predictions.

Table 4. Analysis of the effect of Knowledge and Persona Extractors on the Response Generation Model.

Persona	**Knowledge**	**BLEU**	**CharF++**	**ROUGE-L**	**Avg Generation**	**Data Subset**	**Diff. from the best avg**
true	true	36.02	57.25	57.35	50.20	valid	-
true	extracted	34.85	56.20	56.16	49.07	valid	1.13
extracted	true	34.37	55.75	55.75	48.62	valid	1.58
extracted	extracted	34.26	55.61	55.60	48.49	valid	1.71
extracted	extracted	32.47	53.90	54.31	46.89	test	3.30

knowledge, selects appropriate sentences from the persona description, and

generates a response to the current dialog context taking into account selected knowledge and persona. The approach is developed based on the FoCus dataset presented in the framework of The 1st Workshop on Customized Chat Grounding Persona. The system consists of three consequent steps of data processing and may be easily integrated into a modular dialog system. The proposed system contains the following steps: ranking of knowledge sentences to select the most suitable one, binary classification of pairs dialog context and persona sentence, and conditional response generation. The proposed KRGP system achieves state of the art results, outperforming both the baselines and, as of today, all competitors on the leaderboard. The KRGP system is based on models with 110M and 140M parameters, which makes it a production-ready approach.

§8. Limitations

Unfortunately, the proposed approach has a number of limitations that require additional research. First, the system does not construct user data, using ready-made examples instead. It is likely that if we change the format and domain of the user data, the model will not work like it did on the original dataset. Second, we make an assumption that the information necessary for response generation is found in only one section of the text. Such an assumption does not let us use several fragments of the text for further conditioning of the model, which might equip it with additional knowledge on each step and thus improve its performance.

References

1. M. Adam, M. Wessel, A. Benlian, *Ai-based chatbots in customer service and their effects on user compliance.* — Electronic Markets **31**, No. 2 (2021), 427–445.
2. E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, J. Weston, *Wizard of wikipedia: Knowledge-powered conversational agents*, arXiv preprint arXiv:1811.01241 (2018).
3. C. Gao, W. Lei, X. He, M. de Rijke, Tat-Seng Chua, *Advances and challenges in conversational recommender systems: A survey*, AI Open **2** (2021), 100–126.
4. P. He, J. Gao, W. Chen, *Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing*, 2021.
5. Y. Jang, J. Lim, Y. Hur, D. Oh, S. Son, Y. Lee, D. Shin, S. Kim, H. Lim, *Call for customized conversation: Customized conversation grounding persona and knowledge*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 10803–10812.
6. H. Jiang, Y. Cheng, J. Yang, S. Gao, *Ai-powered chatbot communication with customers: Dialogic interactions, satisfaction, engagement, and customer behavior*, Computers in Human Behavior **134** (2022), 107329.

7. V. Konovalov, O. Melamud, R. Artstein, I. Dagan, *Collecting Better Training Data using Biased Agent Policies in Negotiation Dialogues*, Proceedings of WOCHAT, the Second Workshop on Chatbots and Conversational Agent Technologies (Los Angeles), Zerotype, September 2016.
8. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, arXiv preprint arXiv:1910.13461 (2019).
9. Y. Li, S. A. Hayati, W. Shi, Z. Yu, *Deux: an attribute-guided framework for sociable recommendation dialog systems*, arXiv preprint arXiv:2105.00825 (2021).
10. J. Lim, M. Kang, Y. Hur, S. Jung, J. Kim, Y. Jang, D. Lee, H. Ji, D. Shin, S. Kim, et al., *You truly understand what i need: Intellectual and friendly dialogue agents grounding knowledge and persona*, arXiv preprint arXiv:2301.02401 (2023).
11. C.-Y. Lin, *ROUGE: A package for automatic evaluation of summaries*, Text Summarization Branches Out (Barcelona, Spain), Association for Computational Linguistics, July 2004, pp. 74–81.
12. T.-Yi Lin, P. Goyal, R. Girshick, K. He, P. Dollár, *Focal loss for dense object detection*, Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
13. I. Loshchilov, F. Hutter, *Decoupled weight decay regularization*, arXiv preprint arXiv:1711.05101 (2017).
14. K. K. Pal, K. Kashihara, U. Anantheswaran, K. C. Kuznia, S. Jagtap, C. Baral, *Exploring the limits of transfer learning with unified model in the cybersecurity domain*, arXiv preprint arXiv:2302.10346 (2023).
15. K. Papineni, S. Roukos, T. Ward, W. jing Zhu, *Bleu: a method for automatic evaluation of machine translation*, Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
16. M. Popović, *chrF: character n-gram F-score for automatic MT evaluation*, Proceedings of the Tenth Workshop on Statistical Machine Translation (Lisbon, Portugal), Association for Computational Linguistics, September 2015, pp. 392–395.
17. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., *Learning transferable visual models from natural language supervision*, International conference on machine learning, PMLR, 2021, pp. 8748–8763.
18. N. Reimers, I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 11 2019.
19. S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, et al., *Recipes for building an open-domain chatbot*, arXiv preprint arXiv:2004.13637 (2020).
20. S. Saha, S. Das, R. K. Srihari, *Proto-gen: An end-to-end neural generator for persona and knowledge grounded response generation*, Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge, 2022, pp. 9–14.
21. K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, *Retrieval augmentation reduces hallucination in conversation*, arXiv preprint arXiv:2104.07567 (2021).

22. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., *Huggingface's transformers: State-of-the-art natural language processing*, arXiv preprint arXiv:1910.03771 (2019).
23. C. Xu, P. Li, W. Wang, H. Yang, S. Wang, C. Xiao, *Cosplay: Concept set guided personalized dialogue generation across both party personas*, Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 201–211.
24. S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, *Personalizing dialogue agents: I have a dog, do you have pets too?*, arXiv preprint arXiv:1801.07243 (2018).
25. H. Zhong, Z. Dou, Y. Zhu, H. Qian, J.-R. Wen, *Less is more: Learning to refine dialogue history for personalized dialogue generation*, arXiv preprint arXiv:2204.08128 (2022).
26. K. Zhou, S. Prabhumoye, A. W. Black, *A dataset for document grounded conversations*, arXiv preprint arXiv:1809.07358 (2018).

Moscow Institute of Physics
and Technology,
Moscow, Russia
E-mail: dimweb.tech@mail.ru

Moscow Institute of Physics
and Technology,
Moscow, Russia
E-mail: dilyara.rimovna@gmail.com

Поступило 6 сентября 2023 г.