Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 35-52
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Personalized Dialogue Generation Model Based on BERT and Hierarchical Copy Mechanism

Zijian Liu, Yan Peng*, Shifeng Ni

School of Computer Science and Engineering, Sichuan University of Science and Engineering, Yibin, China
Email: *2634932795@qq.com

How to cite this paper: Liu, Z.J., Peng, Y. and Ni, S.F. (2022) Personalized Dialogue Generation Model Based on BERT and Hierarchical Copy Mechanism. *Journal of Computer and Communications*, **10**, 35-52.
https://doi.org/10.4236/jcc.2022.107003

Received: June 15, 2022
Accepted: July 24, 2022
Published: July 27, 2022

Abstract

Despite the great advances in generative dialogue systems, existing dialogue generation models are still unsatisfactory in maintaining persona consistency. In order to make the dialogue generation model generate more persona-consistent responses, this paper proposes a model named BERT-HCM (Personalized Dialogue Generation Model Based on BERT and Hierarchical Copy Mechanism). The model uses an encoder based on BERT initialization to encode persona information and dialogue queries and subsequently uses a Transformer decoder incorporating a hierarchical copy mechanism to dynamically copy the input-side content to guide the model in generating responses. The experimental results show that the proposed model improves on both automatic and human evaluation metrics compared to the baseline model and is able to generate more fluent, relevant and persona-consistent responses.

Keywords

Personalized Dialogue Generation, BERT, Hierarchical Copy Mechanism, Persona Consistency

1. Introduction

Building a conversational system that makes it impossible for humans to distinguish between truth and fiction is a long-term goal of artificial intelligence. With the development of deep learning technology and big data technology, generative dialogue systems have made great progress. Today's dialogue systems have made great progress in generating complete and fluent sentences, but the distance to passing the Turing test [1] is still far away, and one of the important reasons is the lack of persona consistency in dialogue generation. A dialogue system with incoherent personas will greatly degrade the user interaction expe-

rience, so it is important to build a dialogue system with consistent personas to gain the trust of users and establish long-term connections with them.

More recently, Zhang *et al.* [2] introduced the Persona-Chat and promoted widespread interest in researching personalized dialogue models [3] [4] [5] [6]. This dataset defines personas as several profile sentences, requiring the model to generate responses consistent with several persona sentences. Many attempts have been made to control the model to generate more persona-consistent responses [7]-[12].

However, although these attempts have more or less improved different aspects of persona-based dialogue generation, today's persona-based dialogue models still present significant challenges in generating responses with persona consistency. As shown in **Figure 1**, the strong baseline Transfertransfo fine-tuned on the persona dialogue dataset still produces inconsistent responses. By observing the data and analyzing it, we found that some of the content in the gold responses of the persona-based dialogue generation task overlaps with the persona information and dialogue queries. It indicates that we can directly use the words from the persona information and dialogue queries to reply in some cases to help improve the persona consistency of the responses. On the other hand, the model's insufficient understanding of the input content at the semantic level is also a reason for the inconsistent model responses. Previous studies usually encode the input by RNN or GPT model [13], RNN has the limitations of long-term memory forgetting and weak feature extraction, and the GPT model itself is trained with the goal of generation rather than understanding, so both are difficult to understand the input adequately.

To solve the above problems, this paper proposes a model named Personalized Dialogue Generation Model Based on BERT and Hierarchical Copy Mechanism (BERT-HCM), which facilitates the model to generate relevant and persona-consistent responses by improving the semantic understanding capability

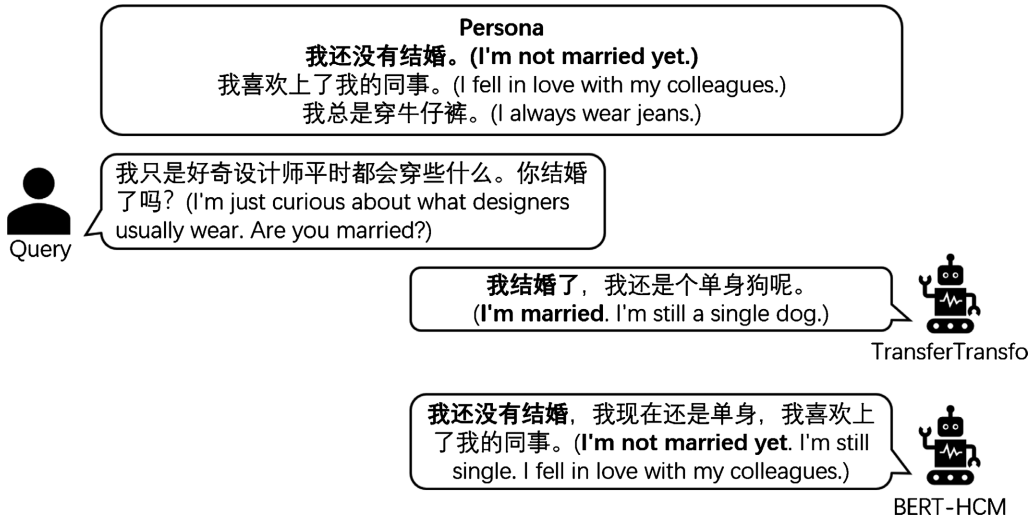

Figure 1. The baseline model Transfertransfo fine-tuned by GPT still generates inconsistent responses and proposed model can generate consistent responses.

of the encoder and directly copying the persona information and the dialogue query. Specifically, BERT-HCM is initialized and fine-tuned at the encoder side using the BERT model [14], which is a bi-directional pre-trained language model based on Transformer, and the natural language understanding ability learned in the pre-training stage of the BERT model can help improve the semantic understanding ability of the encoder as well as speed up the model training. On the decoder side, this paper innovatively adds a hierarchical copy mechanism to the original Transformer decoder to improve the utilization of input information by dynamically copying the content of the input. Extensive experiments on both Chinese and English persona-based dialogue datasets demonstrate that our model can generate more fluent, relevant and persona-consistent responses than other baselines.

2. Related Work

2.1. Personalized Dialogue Generation

Personalized dialogue generation has become an important research area in recent years. At first, researchers used implicit modeling to represent different persona [15] [16], but this implicit modeling approach suffers from the disadvantages of lack of interpretability and poor controllability [17]. Explicit modeling is one way to solve this problem, Zhang *et al.* [2] collected a persona-based dataset by asking randomly paired staff and assigning them random persona description information for chatting, and modeled it using RNN-based Seq2Seq. And then, a large number of studies based on explicit modeling emerged, Zheng *et al.* [3] proposed RNN-based models for persona-aware attention and persona-aware bias. Qian *et al.* [4] proposed incorporating a persona information detector to determine whether to reply using persona information. Song *et al.* [5] proposed the use of memory networks combined with CVAE to increase the diversity of persona-based dialogues. Yavuz *et al.* [6] designed the DeepCopy model, which uses a copy mechanism to integrate persona information. The above models are all RNN-based models. However, as Transformer-based pre-trained models achieved good results in various NLP tasks, researchers started to experiment with Transformer-based models for persona-based dialogue modeling. Wolf *et al.* [7] achieved first place in the automatic evaluation of the ConvAI2 competition [18] by fine-tuning the GPT model with direct splicing of persona information and dialogue queries, while the first place in the human evaluation in the same competition went to the GPT-based multi-input Transformer model proposed by Golovanov *et al.* [8]. In addition, Zheng *et al.* [9] added an attention routing on the Decoder side to dynamically exploit the persona information based on the work of [8]. Liu *et al.* [10] proposed a persona-aware robot that uses reinforcement learning to align the persona information of both parties during the conversation, improving the quality of conversation generation. Wang *et al.* [11] used persona sentence reconstruction and calibration networks to improve the persona consistency of

responses. Cao *et al.* [12] proposed the use of data augmentation and Curriculum learning to enhance the performance of persona dialogues without relying on the model.

2.2. Pre-Training Language Model

In recent years, pre-trained language models have made great progress. With the release of models such as GPT [13] and BERT [14] and the significant enhancements brought by them on several downstream tasks, pre-trained language models have become an indispensable and fundamental paradigm in the field of natural language processing. According to the different pre-training tasks, we can simply classify pre-trained language models into two different models, Autoregressive and Autoencoder. Autoregressive models have the same task as language models, predicting the next word based on the above or predicting the previous word based on the below, e.g., GPT model is obtained using Transformer's decoder as the basis, trained with the classical language model task, and belong to autoregressive models. Autoencoder models (e.g., BERT), also commonly referred to as Denoising Autoencoder models, which are pre-trained with the goal of predicting randomly masked words in the input based on context, have the advantage of simultaneously exploiting the contextual information of the predicted words and are therefore more suitable for natural language understanding tasks. In the area of dialogue generation, Chen *et al.* [19] set up several pre-training tasks for dialogue scenarios and proposed DialogVED to achieve excellent results on several dialogue datasets, and [20] proposed the DialogGPT model for dialogue tasks, which builds on the GPT-2 model and uses a large-scale dialogue corpus to train models that can produce more diverse and relevant responses. For persona-based dialogues, the performance of [7] and [8] demonstrates that pre-trained models can effectively improve the quality of persona-based dialogues responses.

2.3. Copy Mechanism

The copying mechanism can be divided into soft-gated copy and hard-gated copy. Gu *et al.* [21] first proposed CopyNet in 2016, and added the copy mode to distinguish the generate mode of the original Seq2Seq model. The two modes are in the form of soft gating co-directs dialogue generation, alleviating the OOV problem. See *et al.* [22] proposed a pointer generation network for text summarization and added pointer probabilities on the basis of CopyNet to fusion the two-mode probabilities in a weighted manner. [6] designed the DeepCopy model, which utilizes a hierarchical pointer generation network to integrate persona information to guide dialogue generation. [23] designed a hard-gated copy to choose whether the generated word originates from a predefined vocabulary or points to the dialogue history.

Different from the above work based on the pre-training language model and copy mechanism, this paper attempts to improve the quality and persona con-

sistency of the responses generated by the model by enhancing the model's semantic understanding of the input information and directly copying the input information. We propose a novel hierarchical copy mechanism and organically combine it with the Transformer model based on the BERT encoder.

3. BERT-HCM Model

3.1. Task Definition

The task of this paper is to learn a dialogue generation model based on the persona context, so that the response generated by the model is consistent with the information of the given persona description. Specifically, the input of the model consists of persona information and dialogue query. $P = \{P_1, P_2, \cdots, P_k\}$ denotes the persona description information, and the persona description of each sample consists of k sentences. $Q = q_1, q_2, \cdots, q_n$ denotes the dialogue query, with each dialogue query consisting of n words. $G = g_1, g_2, \cdots, g_n$ denotes the gold response, each gold response consists of n words. R denotes the generated responses of the model, $R = r_1, r_2, \cdots, r_n$, with each generated response consisting of n words. Our goal is to learn a model M such that the generated response R is as close as possible to the gold response G by modeling the generated response $R = M(P,Q)$.

3.2. Overview

The model proposed in this paper is a Transformer-based model with an encoder-decoder structure. The Transformer model has a strong capability of long-range information capture and parallel computation, so it is very suitable for modeling dialogue models, and the specific model structure is shown in **Figure 2**. Specifically, the input persona information of the model P and the dialogue query Q is spliced through the word embedding layer (including word embedding, position embedding and segment embedding) to obtain the word embedding vector, and then the BERT encoder is used to encode the word embedding vector to obtain the hidden state $H_{[P,Q]}$. At the decoder side, the gold response input G is similarly input to the hierarchical copy decoder for decoding through the embedding layer shared with the encoder, and the final response R is obtained.

3.3. Word Embedding Layer

The word embedding layer is directly initialized using the word embedding layer of BERT, including three parts: word embedding, position embedding and segment embedding, and the three parts are fused by vector summation, as follows:

$$e_i = e_{wi} + e_{pi} + e_{si} \tag{1}$$

where e_i denotes the input representation of the i-th word, e_{wi} denotes the word embedding of the i-th word, e_{pi} denotes the position embedding of the i-th word, and e_{si} denotes the segment embedding of the i-th word. The input representa-

tion is shown in Figure 3. In addition, the encoder and decoder of the model in this paper share the same word embedding layer.

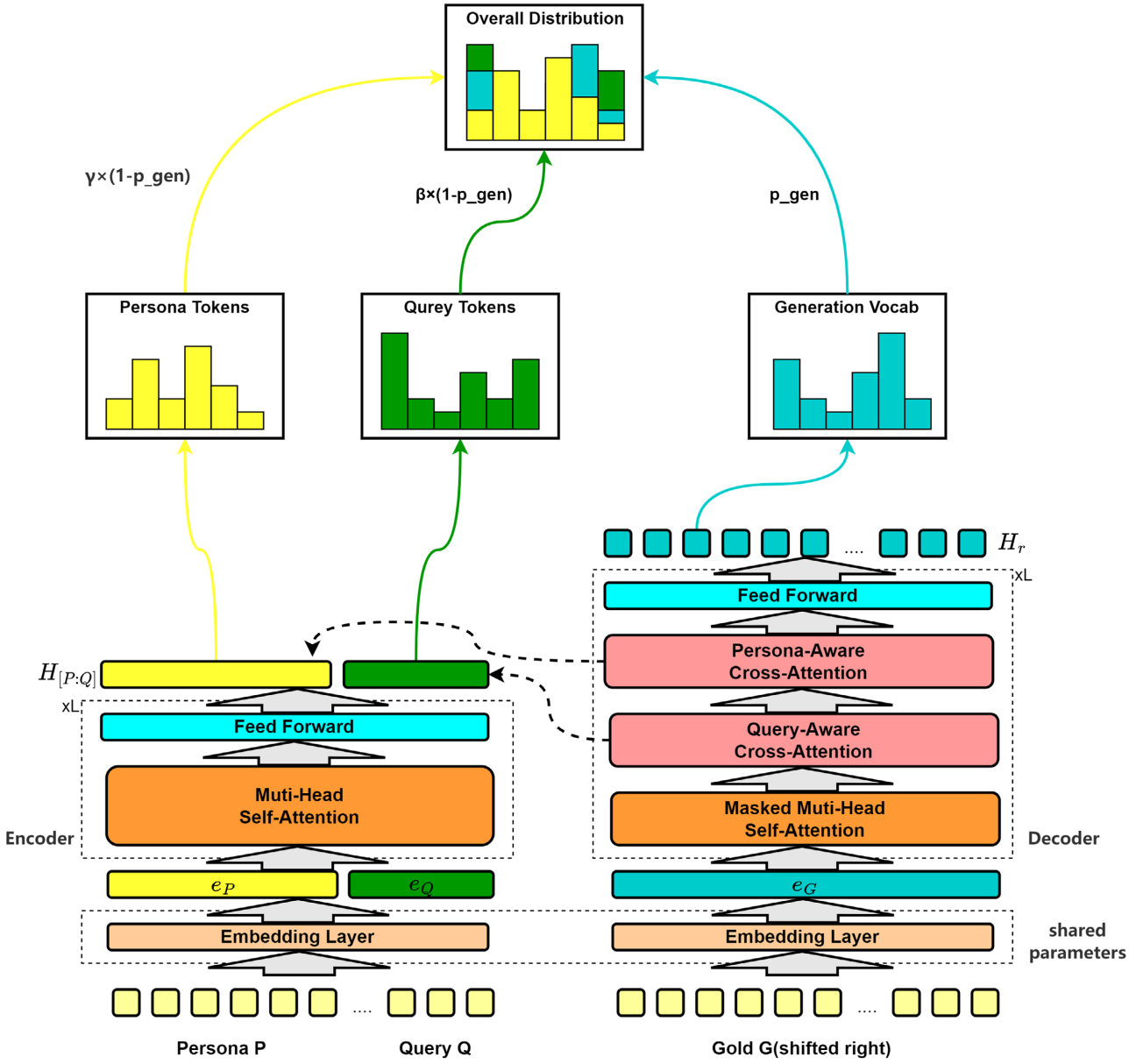

Figure 2. The overall architecture of the BERT-HCM model.

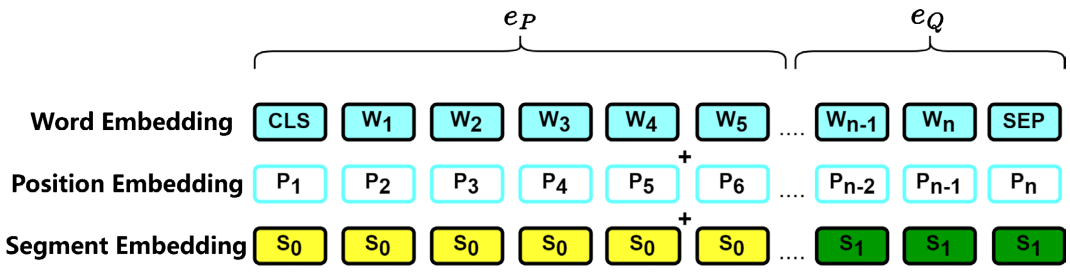

Figure 3. The representation of the dialogue input. The input embedding for each token is the sum of a word embedding, a positional embedding, and a segment embedding.

3.4. BERT Encoder

BERT [14] (Bidirectional Encoder Representation from Transformers) is a pre-trained language model proposed by Google AI Research Institute in 2018, using the Encoder structure of Transformer with hundreds of millions of learnable parameters. Unlike general autoregressive one-way pre-trained language models, BERT is a bi-directional self-coding model with a pre-training phase targeting masked language model (MLM) and next sentence prediction (NSP). Benefiting from different pre-training tasks, BERT has a very strong natural language understanding ability and is widely used for tasks related to natural language understanding. In order to enable the model to have a stronger semantic understanding and help the model to generate higher quality responses, the encoder part of this paper uses a standard BERT model. The output of the word embedding is subjected to the multi-layer bidirectional encoding of BERT to obtain a high-dimensional hidden state. In our model, the input consists of a persona information P and a dialogue query Q. The format of the input is shown in Equation (2).

$$e_{input} = [CLS], e_{p_1}^{(1)}, e_{p_1}^{(2)}, \cdots, e_{p_n}^{(m)}, [SEP], [CLS], e_q^{(1)}, \cdots, e_q^{(m)} [SEP] \tag{2}$$

where e represents the word embedding representation that has gone through the word embedding layer, n represents the number of sentences with persona information, and m represents the length of the sentence. After that, e_{input} will pass through multiple layers of multi-headed attention layers and feedforward neural network layers on the BERT encoder to obtain the hidden vector $H_{[P:Q]}^{L}$. Each layer can be represented as:

$$h_{[P:Q]}^{l+1} = FNN\left(MHA\left(h_{[P:Q]}^{l}, h_{[P:Q]}^{l}, h_{[P:Q]}^{l}\right)\right) \tag{3}$$

where MHA refers to the multi-head attention layer, FNN refers to the fully connected feedforward neural network layer, $h_{[P:Q]}^{0} = e_{input}$, l stands for the current number of layers, and L stands for the total number of layers.

3.5. Hierarchical Copy Decoder

To make the model use the input information more efficiently and accurately, we improve the pointer generation network based on [22], and innovatively propose a Transformer decoder that incorporates the hierarchical copy mechanism. This method not only organically combines the copy mechanism with the Transformer but also makes full use of the hierarchical guidance of the attention score to make the copy more accurate. The so-called hierarchical copy mechanism can be divided into two layers: the first layer calculates how much weight each copy mode and generate mode accounts for, and the second layer calculates how much weight each of the two different copy targets accounts for under the condition that the weight of copy in the first layer is known, and finally uses the weighted sum of these calculated weights to obtain the final vocabulary distribution. The specific structure is shown in the decoder part of **Figure 2**, in which

two cross-attention layers are stacked in a vertical manner, namely the persona-aware cross-attention layer and the query-aware cross-attention layer. The final generation result is obtained by a hierarchical soft fusion method similar to the gate mechanism in three parts: the vocabulary distribution generated by the Transformer decoder, the vocabulary distribution generated by the persona-aware cross-attention layer, and the vocabulary distribution generated by the persona-aware cross-attention layer. The input is similar to the encoder, except that the input here uses an upper triangular mask to ensure that the prediction only depends on the known output. The specific input format is shown in Equation (4).

$$e_{target} = \left[CLS\right], e_g^{(1)}, e_g^{(2)}, \cdots, e_g^{(m)} \tag{4}$$

where e represents the word embedding representation that has passed through the word embedding layer, and m represents the length of the sentence. Then, e_{target} will feed to the Transformer decoder based on the hierarchical copy mechanism, first passing through a multi-head self-attention layer with a mask, and then passing through the query-aware cross-attention layer and the persona-aware cross-attention layer in turn, cycling L times to obtain the hidden vector H_r^L. Equations (5) to (8) are the decoding process of the lth layer.

$$h_g^{l+1} = MMHA\left(h_r^l, h_r^l, h_r^l\right) \tag{5}$$

$$h_q^{l+1}, a_q^{l+1} = MHA\left(h_g^{l+1}, H_Q^L, H_Q^L\right) \tag{6}$$

$$h_p^{l+1}, a_p^{l+1} = MHA\left(h_q^{l+1}, H_P^L, H_P^L\right) \tag{7}$$

$$h_r^{l+1} = FNN\left(h_p^{l+1}\right) \tag{8}$$

where *MMHA* refers to multi-head attention layer with mask, *MHA* refers to multi-head attention layer, *FNN* refers to fully connected feedforward neural network layer, $h_r^0 = e_{target}$, l stands for the current number of layers, and L stands for the total number of layers.

Next, three vocabulary distributions are calculated, namely the vocabulary distribution P_{decode} generated by the Transformer decoder, the vocabulary distribution P_{query} generated according to the query-aware cross-attention layer and the vocabulary distribution $P_{persona}$ generated according to the persona-aware cross-attention layer. The calculation process is shown in Equations (9) to (11).

$$p_{decode} = softmax\left(H_r^L \cdot W_A + b_A\right) \tag{9}$$

$$p_{query} = a_q^L \cdot one\_hot_{qvocab} \tag{10}$$

$$p_{persona} = a_p^L \cdot one\_hot_{pvocab} \tag{11}$$

where $W_A \in \mathbb{R}^{h_{dim} \times |vocab|}$ denotes the learnable weight matrix, b_A denotes the bias, a denotes the attention score calculated by Equation (6) and Equation (7), and one_hot_{vocab} denotes the unique hot encoding of the input corresponding vocabulary. Equations (12) and (13) show the linear combination of C_{query} and $C_{persona}$ based on the decoder's attention distribution on the dialogue query and persona

information:

$$C_{query} = a_q^L \cdot H_Q^L \tag{12}$$

$$C_{persona} = a_p^L \cdot H_P^L \tag{13}$$

Finally, the word probability distribution of the final responses is obtained by means of a hierarchical soft fusion. The calculation process is shown in Equations (15) to (16).

$$p_{gen} = Sigmoid\left(cat\left(C_{query}, C_{persona}, H_r^L, e_{target}\right) \cdot W_B + b_B\right) \tag{14}$$

$$\beta, \gamma = softmax\left(cat\left(C_{query}, C_{persona}\right) \cdot W_C \cdot e_{target}\right) \tag{15}$$

$$p_{response} = p_{gen} \times p_{decode} + \beta \times \left(1 - p_{gen}\right) \times p_{query} + \gamma \times \left(1 - p_{gen}\right) \times p_{persona} \tag{16}$$

where $W_B \in \mathbb{R}^{4 \cdot h_{dim} \times 1}$ and $W_C \in \mathbb{R}^{h_{dim} \times h_{dim}}$ denotes the learnable weight matrix and b_B denotes the bias.

4. Experiment

4.1. Datasets

To verify the effectiveness of the model, this paper conducts experiments on two public character dialogue datasets respectively, including the Chinese Persona-Chat (CPC) dataset published by Baidu[1] and the Persona-Chat (PC) dataset published by Zhang *et al.* [2]. Both datasets have similar structures, and are Persona-Dense multi-turn conversation datasets. We sliced and processed data into a single-turn conversation with one question and one answer, and each sample consists of three parts, namely, persona information (Persona), conversation query (Query), and gold response (Gold), where the persona information is the persona information of the responding party, which generally consists of three or four sentences. It is worth noting that, since the original CPC dataset contains a large number of dialogue responses irrelevant to persona information, in order to make the data more relevant to persona information, a large number of dialogues irrelevant to persona information are filtered out using hand-written scripts. The distribution of sample sizes in the final two datasets is shown in Table 1.

4.2. Baseline Models

1) S2SA [24]. The encoder-decoder model is based on RNN and Attention mechanism, which was originally used in the field of machine translation. It is a classical Seq2Seq model and is widely used in comparison experiments for

Table 1. Number of data distribution.

Dataset	Train	Valid	Test
CPC	93462	10000	4117
PC	112499	10000	7801

[1]https://www.luge.ai/#/luge/dataDetail?id=38

dialogue generation. The number of RNN layers is set to 2, and the dimension size of both the hidden and word embedding layers is set to 512.

2) PGN [22]. The structure and parameter settings are the same as those of S2SA; the difference is that the decoder side adds a pointer network to directly copy the text from the input to guide the reply generation.

3) Transformer [25]. Based on the Encoder-Decoder model proposed by Google. It completely discards the RNN structure, replaces RNN with self-attention, and introduces a multi-head attention mechanism. In this paper, we use the same parameter settings for training as in the original paper.

4) Transfertransfo [7]. The first-place solution for automatic evaluation in the ConvAI2 competition, proposed by the Huggingface team. It was obtained using fine-tuning on pre-trained language model GPT on personalized dialogue data. Since there is no official Chinese version of GPT, this paper uses the Chinese GPT-2 model published by UER [26] for experiments.

5) LIC [8]. The first-place solution for human evaluation in the ConvAI2 competition, proposed by the Lost in Conversation team. It is a multi-input model that uses the encoder-decoder structure, using GPT initialization parameters similarly. For the Chinese dataset, the same Chinese GPT-2 model published by UER was used for experiments.

6) BERT2BERT [27]. A Seq2Seq model based on Transformer, both encoder and decoder are initialized using the official open-source BERT model.

4.3. Experimental Settings

In our experiments, the optimizer is AdamW, the loss function is NLLLoss, the batch size is 16, the dropout ratio is set to 0.1, and the partial learning rate of the initialization parameters using the pre-training model is 7e - 6, and the learning rates of other parts are 5e - 5. The learning rate adjustment strategy is Linear-Schedule, the number of preheating steps is 1/10 of the total steps, the number of epochs is 20, and each epoch is verified once, the best performing model in the verification set is saved as the final model. The decoding strategy is greedy decoding. For the vocabulary, all models in the Chinese dataset use the Chinese version of BERT vocabulary with the size of 21,128, while all models in the English dataset use the English BERT vocabulary except for the models initialized with GPT. All experiments are implemented on the PyTorch framework, using a single NVIDIA RTX 3090 24 GB GPU card for training.

4.4. Evaluation Metrics

1) BLEU [28]. It is used to generate the similarity of evaluation responses to the gold responses, which was originally used for the evaluation of machine translation and is now widely used in a variety of NLP tasks, with larger values being better. In this paper, four tuples of BLEU ($n = 1, 2, 3, 4$) are used.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{17}$$

where $BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$, $p_n = \dfrac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count_{clip}(n-gram')}$

2) F1 [18]. The accuracy of the responses is evaluated based on the calculation of precision and recall at the character level of the generated responses, and the larger the value, the better.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{18}$$

3) BERT-Score [29]. Word vector-based text generation evaluation metrics are encoded by pre-trained BERT to vectorize two sentences, and the degree of similarity between them is calculated by cosine similarity, and the higher the score, the better. In addition, to improve the readability of the data, we use rescaling BERT-Score, with a numerical range between −1 and 1.

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^{\mathrm{T}} \hat{x}_j \tag{19}$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in \hat{x}} x_i^{\mathrm{T}} \hat{x}_j \tag{20}$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{21}$$

4) Persona coverage (P-Cover). Inspired by [5] [11], we used Persona coverage to evaluate persona consistency at the word level, calculating the ratio of generated responses to the overlapping unigrams between the sentences of each persona. The calculation is defined as:

$$C_{per} = \max_{i \in [1,k]} \frac{S(R, P_i)}{n} \tag{22}$$

where k is the number of persona sentences in persona profile, n is the number of words in persona sentence, and $S(\cdot)$ is the set of shared words.

5) Persona embedding matching (P-EM). To evaluate the persona consistency at the sentence semantics level, we computed the cosine similarity of the generated response to each persona sentence after embedding using BERT. The calculation was defined as:

$$EM_{per} = \max_{i \in [1,k]} CS(BERT(R), BERT(P_i)) \tag{23}$$

where k is the number of persona sentences in persona profile, $CS(\cdot)$ denote the calculation of cosine similarity, and $BERT(\cdot)$ denotes embedding using BERT where the sentence vector is represented using the average of the word vectors.

6) Human Evaluation. Considering the limitations of automatic evaluation [30], which cannot reflect the real performance of the model well, this paper also conducts a manual evaluation. We randomly selected 100 test samples from the test data and assigned them to 3 reviewers for manual evaluation. They were evaluated from three aspects: fluency, relevance and persona consistency. Each aspect is divided into {0, 1, 2} three levels.

4.5. Results and Analysis

The experimental results of different models on the CPC dataset and PC dataset are presented in Table 2 to Table 5. It can be seen that the performance on the CPC dataset is numerically better than that on the PC dataset, probably because the CPC dataset was cleaned of data before use, while the PC dataset was not cleaned. Overall, the model proposed in this paper outperforms the baseline model in most of the major automatic evaluation metrics and human evaluation metrics.

Through the comparative analysis of the automatic evaluation results in Table 2 and Table 3, it is obvious that the traditional RNN-based model has a large gap with the Transformer-based model in various metrics, which indicates that the Transformer-based model is much better than the RNN-based model in terms of representation capability. In addition, the proposed model achieved competitive results in BLEU, F1 and BERT-Score metrics compared with other baseline models, which indicates that the proposed model can generate more fluent and relevant responses. In terms of persona consistency, the P-Cover and P-EM metrics achieve the best performance in both datasets, indicating that the proposed model effectively improves the generated response persona consistency. The above results all prove that the BERT encoder and hierarchical copy mechanism are effective in improving the response quality and persona consistency of the persona dialogue generation model. It is worth noting that the BERT-based

Table 2. Automatic evaluation results on CPC dataset.

Model	BLEU-1/2/3/4	F1	BERT-Score	P-Cover	P-EM
S2SA	28.80/9.32/2.50/0.56	31.80	0.103	25.49	72.67
PGN	31.00/12.13/4.86/1.96	33.17	0.119	26.52	73.40
Transformer	34.47/23.11/16.93/12.71	40.25	0.357	36.69	85.24
TransferTransfo	37.66/26.36/20.24/15.89	41.01	0.368	40.33	85.95
LIC	34.86/23.64/17.46/13.19	42.64	0.387	31.95	84.18
BERT2BERT	38.15/26.99/20.76/16.24	45.66	0.417	36.69	85.45
BERT-HCM	**39.59/28.69/22.59/18.08**	**46.01**	**0.423**	**40.58**	**86.25**

Table 3. Automatic evaluation results on PC dataset.

Model	BLEU-1/2/3/4	F1	BERT-Score	P-Cover	P-EM
S2SA	10.02/2.81/0.85/0.27	16.72	-0.196	9.03	50.39
PGN	13.01/2.62/0.58/0.11	15.17	-0.160	8.36	54.13
Transformer	19.18/8.57/4.71/2.52	19.16	0.111	13.31	70.57
TransferTransfo	16.77/7.60/3.95/2.08	17.47	0.154	12.87	71.49
LIC	18.07/7.77/3.88/1.97	18.74	**0.180**	12.28	69.68
BERT2BERT	20.66/9.90/5.91/3.40	20.43	0.118	13.41	70.99
BERT-HCM	**21.73/10.71/6.39/3.62**	**21.24**	0.122	**13.69**	**73.40**

model has a larger improvement in the relevant metrics compared to the GPT-based model, indicating that the addition of the BERT model does increase the semantic understanding of the model for the input.

Since the quality of RNN-based model generation is too low, only the Transformer-based model is manually evaluated and scored in this paper, as shown in Table 4 and Table 5. The results of the human evaluation are consistent with those presented in the automatic evaluation, and the model proposed in this paper outperforms the baseline model in all human evaluation metrics, especially in persona consistency. It indicates that the addition of the pre-trained model and the hierarchical copy mechanism effectively improves the quality of model responses and persona consistency.

4.6. Ablation Experiments

To verify the effectiveness of each part of the model proposed in this paper, ablation experiments are conducted in this section. Table 6 shows the results of the ablation experiments. w/o BERT indicates the replacement of the encoder initialized with BERT with the randomly initialized Transform encoder, and w/o HCM indicates the removal of the hierarchical copy mechanism. According to the results, it can be seen that the metrics of persona-based dialogue generation decrease when the BERT-initialized encoder and the hierarchical copy mechanism are not used, respectively. It indicates that both the BERT encoder and the hierarchical copy mechanism contribute to the quality of dialogue generation

Table 4. Human evaluation results of the CPC dataset.

Model	Fluency	Relevant	Persona Consistency
Transformer	1.782	1.402	1.076
TransferTransfo	1.805	1.451	1.178
LIC	1.834	1.458	1.157
BERT2BERT	1.833	1.438	1.162
BERT-HCM	**1.839**	**1.503**	**1.269**
Gold	1.892	1.636	1.474

Table 5. Human evaluation results of the PC dataset.

Model	Fluency	Relevant	Persona Consistency
Transformer	1.562	1.344	0.998
TransferTransfo	1.566	1.350	1.101
LIC	1.571	1.368	1.087
BERT2BERT	1.572	1.352	1.103
BERT-HCM	**1.588**	**1.372**	**1.133**
Gold	1.666	1.427	1.286

and the persona consistency of dialogue. The performance of both datasets is similar. For example, in the PC dataset, the BLEU-4 of the model decreases by 1.39%, BERT-Score decreases by 0.053, P-Cover and P-EM decreases by 0.52% and 3.7%, respectively, when the encoder initialized with BERT is not used. When the hierarchical copy mechanism was removed, BLEU-4 decreased by 0.31%, BERT-Score decreased by 0.002, P-Cover decreased by 0.07%, and P-EM decreased by 2.55%, respectively. It shows that the encoder based on BERT initialization contributes more to the model performance improvement, not only brings a greater improvement in quality and relevance of responses but also brings some improvement to the persona consistency. The hierarchical copy mechanism, although not as significant as the BERT encoder's contribution to the response quality, also brings a big improvement to the persona consistency of the model.

4.7. Case Study

Table 7 is a sample comparison of the results generated by the model in this

Table 6. Ablation experiment results on CPC and PC datasets.

Model	CPC				PC			
	BLEU-4	BERT-Score	P-Cover	P-EM	BLEU-4	BERT-Score	P-Cover	P-EM
BERT-HCM	18.08	0.423	40.58	86.25	3.62	0.122	13.41	73.40
w/o BERT	11.08	0.345	33.69	83.86	2.23	0.069	12.89	69.70
w/o HCM	16.94	0.421	38.25	85.53	3.31	0.120	13.34	70.85

Table 7. Responses generated from BERT-HCM and other baseline models.

Persona	我还没有结婚。我喜欢上了我的同事。我总是穿牛仔裤(I'm not married yet. I fell in love with my colleagues. I always wear jeans)
Query	我只是好奇设计师平时都会穿些什么。你结婚了吗？(I'm just curious about what designers usually wear. Are you married?)
Gold	我还没有结婚呢，不过我有喜欢的人。我很喜欢我的同事。(I'm not married yet, but I have someone I like. I like my colleagues very much.)
BERT-HCM	我还没有结婚，我现在还是单身，我喜欢上了我的同事。(I'm not married yet. I'm still single. I fell in love with my colleagues.)
BERT2BERT	我还没有结婚，我现在还在上学，我很爱我的男朋友。(I'm not married yet. I'm still at school. I love my boyfriend very much.)
LIC	我还没有结婚，我的朋友都说我是个女生。(I'm not married yet. My friends say I'm a girl.)
TransferTransfo	我结婚了，我还是个单身狗呢。(I'm married. I'm still a single dog.)
Transformer	我喜欢吃牛肉，我喜欢牛肉，我觉得牛肉很好吃。(I like beef. I like beef. I think beef is delicious.)
PGN	结了，都经婚我的作了，是为么人我欢。是为么人我不道。(When I am married, I have done it. Is it for me. I don't know why.)
S2SA	结了我经和一恋婚，是个生，是个孩，是最欢的情。是的(After I got married and fell in love, I was a student, a child and the happiest love. yes)

paper and the baseline model. It can be seen that the responses generated by the RNN-based models PGN and S2SA have great problems in readability, and the generated responses are basically unreadable. It confirmed the superiority of the Transformer-based model in dialogue generation modeling again. In addition, the model proposed in this paper is superior to several other baseline models in terms of the ability to understand dialogue queries and persona information and the consistency of generated responses. For example, the persona information shows "I like me my colleague" indicates that the background of the person is an office worker, but the BERT2BERT model replied that he was still in school, which conflicted with the persona information. Although there was no conflict in LIC, the response in the second half of the sentence was almost irrelevant to the dialogue query, and the response from the TransferTransfo model is inconsistent; the Transformer model mistakenly interprets jeans as food beef. In contrast, our model can generate a fluent, relevant and persona-consistent response, proving that the proposed model understands semantic information more fully and the responses have better persona consistency.

5. Conclusions

In this paper, we propose an encoder-decoder model based on Transformer to address persona-based conversation generation. Our model uses the pre-trained bidirectional language model BERT to initialize the encoder of the model to greatly enhance the model's language understanding ability. Moreover, we incorporate an innovative hierarchical copy mechanism on the decoder side, which allows the model to dynamically copy the content of the input side. The combination of both effectively improves the persona consistency of the generated responses. Experimental results show that the proposed model can generate more relevant, fluent and persona-consistent responses than baseline models.

The research in this paper has certain limitations. For example, this paper only considers a single-round dialogue. In reality, the dialogue often includes multiple rounds, and the continuity between the multiple rounds is relatively important. The follow-up plan is to further study the multi-round persona-based dialogue. In addition, the persona information of the dataset used in this paper is relatively limited. In the future, we will consider introducing more knowledge through topic models, common sense reasoning, etc., to improve the diversity of persona-based dialogues.

Fund

Supported by the Science and Technology Plan of Zigong Science and Technology Bureau (2018GYCX33), and Graduate Innovation Fund of Sichuan University of Science and Engineering (y2021096).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Turing, A.M. (2009) Computing Machinery and Intelligence. In: Epstein, R., Roberts, G. and Beber, G., Eds., *Parsing the Turing Test*, Springer, Dordrecht, 23-65. https://doi.org/10.1007/978-1-4020-6710-5_3

[2] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D. and Weston, J. (2018) Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too? *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Melbourne, 16-18 July 2018, 2204-2213. https://doi.org/10.18653/v1/P18-1205

[3] Zheng, Y., Chen, G., Huang, M., Liu, S. and Zhu, X. (2019) Personalized Dialogue Generation with Diversified Traits.

[4] Qian, Q., Huang, M., Zhao, H., Xu, J. and Zhu, X. (2018) Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation. *Proceedings of the* 27*th International Joint Conference on Artificial Intelligence*, Stockholm, 13-19 July 2018, 4279-4285. https://doi.org/10.24963/ijcai.2018/595

[5] Song, H., Zhang, W.-N., Cui, Y., Wang, D. and Liu, T. (2019) Exploiting Persona Information for Diverse Generation of Conversational Responses. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-19, Macao, 10-16 August 2019, 5190-5196. https://doi.org/10.24963/ijcai.2019/721

[6] Yavuz, S., Rastogi, A., Chao, G.-L. and Hakkani-Tur, D. (2019) DeepCopy: Grounded Response Generation with Hierarchical Pointer Networks. *Proceedings of the* 20*th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, 11-13 September 2019, 122-132. https://doi.org/10.18653/v1/W19-5917

[7] Wolf, T., Sanh, V., Chaumond, J. and Delangue, C. (2019) Transfertransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents.

[8] Golovanov, S., Kurbanov, R., Nikolenko, S., Truskovskyi, K., Tselousov, A. and Wolf, T. (2019) Large-Scale Transfer Learning for Natural Language Generation. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 29-31 July 2019, 6053-6058. https://doi.org/10.18653/v1/P19-1608

[9] Zheng, Y., Zhang, R., Huang, M. and Mao, X. (2020) A Pre-Training Based Personalized Dialogue Generation Model with Persona-Sparse Data. *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, *AAAI* 2020, New York, 7-12 February 2020, 9693-9700. https://doi.org/10.1609/aaai.v34i05.6518

[10] Liu, Q., Chen, Y., Chen, B., Lou, J.-G., Chen, Z., Zhou, B., *et al.* (2020) You Impress Me: Dialogue Generation via Mutual Persona Perception. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 1417-1427. https://doi.org/10.18653/v1/2020.acl-main.131

[11] Wang, W., Feng, S., Chen, L., Wang, D. and Zhang, Y. (2021) Learning to Improve Persona Consistency in Conversation Generation with Information Augmentation. *Knowledge-Based Systems*, **228**, Article ID: 107246. https://doi.org/10.1016/j.knosys.2021.107246

[12] Cao, Y., Bi, W., Fang, M., Shi, S. and Tao, D. (2022) A Model-Agnostic Data Manipulation Method for Persona-Based Dialogue Generation. *Proceedings of the* 60*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Dublin, 22-27 May 2022, 7984-8002. https://doi.org/10.18653/v1/2022.acl-long.550

[13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, **1**, 9.

[14] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *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*), Minneapolis, 3-5 June 2019, 4171-4186.

[15] Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J. and Dolan, B. (2016) A Persona-Based Neural Conversation Model. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Berlin, 8-10 August 2016, 994-1003. https://doi.org/10.18653/v1/P16-1094

[16] Olabiyi, O., Salimov, A.O., Khazane, A. and Mueller, E. (2019) Multi-Turn Dialogue Response Generation in an Adversarial Learning Framework. *Proceedings of the First Workshop on NLP for Conversational AI*, Florence, 1-2 August 2019, 121-132. https://doi.org/10.18653/v1/W19-4114

[17] Huang, M., Zhu, X. and Gao, J. (2020) Challenges in Building Intelligent Open-Domain Dialog Systems. *ACM Transactions on Information Systems* (*TOIS*), **38**, 1-32. https://doi.org/10.1145/3383123

[18] Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., *et al.* (2019) The Second Conversational Intelligence Challenge (ConvAI2). The Springer Series on Challenges in Machine Learning, Springer, Berlin, 187-208. https://doi.org/10.1007/978-3-030-29135-8_7

[19] Chen, W., Gong, Y., Wang, S., Yao, B., Qi, W., Wei, Z., *et al.* (2022) DialogVED: A Pre-Trained Latent Variable Encoder-Decoder Model for Dialog Response Generation. *Proceedings of the* 60*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Dublin, 22-27 May 2022, 4852-4864. https://doi.org/10.18653/v1/2022.acl-long.333

[20] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., *et al.* (2020) DIALOGPT: Large-Scale Generative Pre-Training for Conversational Response Generation. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, Online, 5-10 July 2020, 270-278. https://doi.org/10.18653/v1/2020.acl-demos.30

[21] Gu, J., Lu, Z., Li, H. and Li, V.O.K. (2016) Incorporating Copying Mechanism in Sequence-to-Sequence Learning. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Berlin, 8-10 August 2016, 1631-1640. https://doi.org/10.18653/v1/P16-1154

[22] See, A., Liu, P.J. and Manning, C.D. (2017) Get to the Point: Summarization with Pointer-Generator Networks. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Vancouver, 31 July-2 August 2017, 1073-1083. https://doi.org/10.18653/v1/P17-1099

[23] Wu, C.-s., Socher, R. and Xiong, C. (2019) Global-to-Local Memory Pointer Networks for Task-Oriented Dialogue. *Proceedings of the* 7*th International Conference on Learning Representations*, *ICLR* 2019, New Orleans, 6-9 May 2019.

[24] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate.

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al.* (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems* 30: *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 5998-6008.

[26] Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., *et al.* (2019) UER: An Open-Source Toolkit for Pre-Training Models. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*): *System Demonstrations*, Hong Kong, 3-7 November 2019, 241-246. https://doi.org/10.18653/v1/D19-3041

[27] Rothe, S., Narayan, S. and Severyn, A. (2020) Leveraging Pre-Trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, **8**, 264-280. https://doi.org/10.1162/tacl_a_00313

[28] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002) Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the* 40*th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, U 7-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[29] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y. (2020) BERTScore: Evaluating Text Generation with BERT. *Proceedings of the* 8*th International Conference on Learning Representations*, Addis Ababa, 26-30 April 2020.

[30] Novikova, J., Dušek, O., Cercas Curry, A. and Rieser, V. (2017) Why We Need New Evaluation Metrics for NLG. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, *EMNLP* 2017, Copenhagen, 9-11 September 2017, 2241-2252. https://doi.org/10.18653/v1/D17-1238