



# Prompt Guided Copy Mechanism for Conversational Question Answering

Yong Zhang, Zhitao Li, Jianzong Wang\*, Yiming Gao, Ning Cheng, Fengying Yu, Jing Xiao

Ping An Technology (Shenzhen) Co., Ltd., China

## Abstract

Conversational Question Answering (CQA) is a challenging task that aims to generate natural answers for conversational flow questions. In this paper, we propose a pluggable approach for extractive methods that introduces a novel prompt-guided copy mechanism to improve the fluency and appropriateness of the extracted answers. Our approach uses prompts to link questions to answers and employs attention to guide the copy mechanism to verify the naturalness of extracted answers, making necessary edits to ensure that the answers are fluent and appropriate. The three prompts, including a question-rationale relationship prompt, a question description prompt, and a conversation history prompt, enhance the copy mechanism's performance. Our experiments demonstrate that this approach effectively promotes the generation of natural answers and achieves good results in the CoQA challenge.

**Index Terms:** Prompt, Copy Mechanism, Conversation Question Answering

## 1. Introduction

Conversational Question Answering (CQA) aims to answer conversation flow questions given an understanding of the text. Unlike traditional single-turn question answering, CQA questions are conversational in nature. They are based on the conversation history, which may require history modeling to handle coreference and pragmatic reasoning. In addition, CQA answers are coherent and must consider the previous questions and answers in the conversation. This coherence means that pure rationale may need to be edited for the natural flow of the conversation.

Extractive methods focus on identifying the relevant rationale from the passage to answer the question, based on the connection between the question and the conversation history. These methods employ an extraction-based approach to generate answers, by predicting the start and end position of the rationale in the passage. HAE[1] considers the construction of historical answer markers in the embeddings. FlowQA[2] mimics the memory mechanism of humans by processing questions and information sequentially as the conversation progresses. GHR[3] introduces a dialogue-based attention mechanism to capture interactions between different conversation turns, and Excord[4] rewrites the questions based on the history to complement meaning before treating them as single-turn QA. However, these methods often ignore the relationship between the question and the passage rationale.

On the other hand, naturalness-based methods focus on generating fluent and coherent answers, using the sequence-

to-sequence structure. Answer-BART[5] uses an end-to-end model to process the question and passage, generating potential evidence and a natural answer. REAG[6] incorporates the evidence extraction task into the transformer model's encoder to improve the natural answer's confidence. Unlike the above methods, S-net[7] fuses the extraction and generation that it first uses the extraction model to collect the passage's most-important sub-text and then synthesizes them into the final answer by the generative model.

Although sequence-to-sequence methods have achieved significant success in generating natural and fluent answers in conversational question answering, they still suffer from the risk of producing hallucinated answers that are not related to the input. In contrast, extractive methods can directly capture answer spans from the associated text, which reduces the randomness of generative models and provides a strong connection to the input. However, extractive methods often do not take the conversation history and question type into account.

We propose a pluggable prompt-guided copy mechanism to improve the naturalness of extractive answers in conversational question answering. The mechanism verifies and potentially rewrites the extracted rationale span to meet the two principles of naturalness. Firstly, the answer should be fluent, given that the rationale text may be partial, redundant, and not always consistent with the question. Secondly, the answer should be appropriate for the question type, as the natural rationale text may need to be rephrased to suit the question type.

To improve the copy mechanism, we propose the use of three prompts to link conversational questions with rationale texts. These prompts include a prompt that establishes the relationship between the question and its rationale, a prompt that provides additional information about the question, and a prompt that draws on examples from past conversations. While prompts are commonly used in single-turn question answering[8, 9, 10], our approach applies them to multi-turn CQA tasks for more effective fine-tuning.

During fine-tuning, explicit text prompt clues are transformed into hidden embeddings and fully interacted under the self-attention mechanism to activate crucial words. These embeddings are then converted into guidance information for the copy mechanism to make better decisions in generating the answer. To leverage the information at different granularities constructed in the prompts, we propose letting the decoder pay attention to all representations generated by the encoders, providing a multi-view perspective.

Our paper makes the following contributions:

- We propose a pluggable approach for extractive methods that uses a prompt-guided copy mechanism to improve the fluency and appropriateness of extracted answers.

\*Corresponding author: Jianzong Wang, jzwang@188.com.

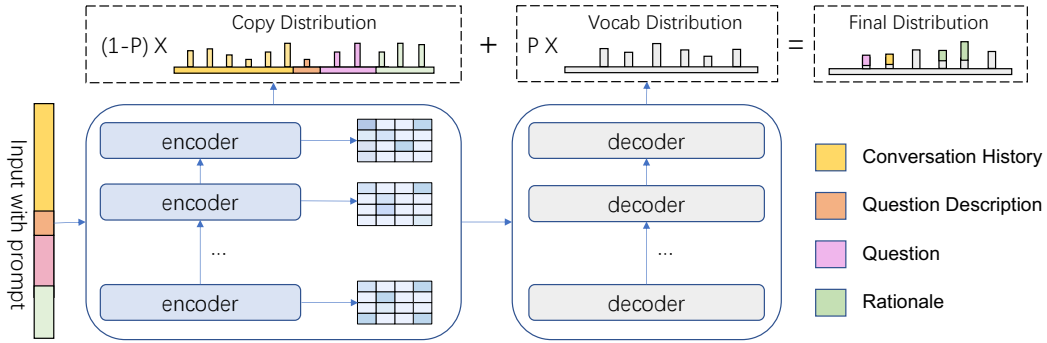


Figure 1: The architecture of our proposed model

- We propose the use of three prompts, including a question-rationale relationship prompt, a question description prompt, and a conversation history prompt, to enhance the effectiveness of the copy mechanism.
- We introduce a multi-view perspective attention mechanism that digests prompts to guide the copy mechanism and improve its performance.
- We demonstrate the effectiveness of the proposed approach in promoting the generation of natural answers and achieving good results in the CoQA[11] challenge.

## 2. Methodology

This section describes our prompt-guided copy mechanism method to enhance the naturalness of answers in conversational question answering. Figure 1 outlines the method’s architecture, highlighting the interplay between the question and the rationale span in final answer generation.

The method transforms raw inputs into enriched text through tailored prompts, which are converted into embeddings by the encoder. This enriched embedding guides the decoder’s copy mechanism to generate the final natural answer.

Formally, an “answer” is a consistent response in dialogues that follows the right reference, tense, and grammar, and matches the question format. The “rationale” is a small text piece from the reference article, which may be semantically incomplete and is not bound by the questions in the dialogue. We consider extracted answers as rationales. Given a question  $q_t$  and rationale  $r_t$  with history dialogue  $H_t = [(q_1, a_1), (q_2, a_2), \dots, (q_{t-1}, a_{t-1})]$ , where  $(q_i, a_i)$  indicates the question and answer in the  $i$ -th turn in conversation history, the goal is to verify the rationale for the  $q_t$  and give answer  $a_t$ .

### 2.1. Prompt Design

The manual-designed prompt is efficient in many tasks[8]. To enhance the information of input questions and rationales, we propose three prompts considering the connection between conversational questions and rationales, question categories, and contextual information. These prompts are detailed in Table 1.

Table 1: Three proposed prompts

Prompt	Text
Question Rationale	Question: q + Rationale: r
Question Description	Category Question: q + Rationale: r
Conversation History	Category Question: $q_{t-1}$ + Rationale: $r_{t-1}$ +Answer: $a_{t-1}$ + Category Question: q + Rationale: r + Answer:

#### 2.1.1. Question Rationale Prompt

The first one is the question rationale relationship prompt. We naturally use a special symbol  $[SEP]$  to isolate the question and rationale text, but it lacks prior information and requires fine-tuning to learn the connection. To better activate the pre-trained model’s knowledge, we add the words “Question” and “Rationale” before the corresponding text to directly establish the connection between the two input texts.

#### 2.1.2. Question Description Prompt

In the second prompt, we aim to address the impact of question categories on the paraphrasing of the rationale text. The mapping of rationale to the answer often relies on the question category[12]. The rationale may require different types of paraphrasing for various question categories, including but not limited to tense transition, pronoun substitution, and conjunction substitution. For instance, count questions require counting the enumerates in the text and providing the number as the answer. To capture the impact of question categories, we analyze the first word of the questions in the training set and select the most common interrogative words as question categories. We then append the corresponding interrogative word to “Question” as a question description prompt.

#### 2.1.3. Conversation History Prompt

To enhance the fluency and appropriateness of the generated answers, we propose using the conversation history as a prompt for input. The rationale text, extracted from the passage based on the comprehension of the question and conversation history, may be semantically incomplete and lack essential information. By incorporating the conversation history as a prompt, we can provide necessary background information during the pragmatic reasoning process of answer generation.

Additionally, question-answer pairs in the conversation history serve as answer examples to guide the model in generating style-aware answers for different categories of questions. This helps alleviate the burden of the model in learning all question-aware styles[9]. To implement this, we assemble the corresponding history conversation question-rationale-answer samples with text class labels “question” and “answer” and add them to the beginning of the input in history order.

## 2.2. Guided Copy Mechanism

We introduce a guided copy mechanism [13, 14, 15] that incorporates prompt-enhanced information under self-attention to verify the suitability of extracted rationales. If the rationale is not qualified, the mechanism performs an edit to ensure the answer is fluent and appropriate. By directly referencing the vocabulary from the question and rationale, the copy mechanism tackles issues of incorrect or inconsistent terminology, hence fostering consistent usage across multi-turn dialogues.

### 2.2.1. Multi View Copy Distribution

To fully capture the different levels of syntactic and semantic information, we propose that the last decoder should attend to all encoder layers [16, 17], a concept similar to but distinct from the transparent attention [18] method, which mainly improves gradient flow. The multi-head of the self-attention mechanism [19] would learn diverse aspects of the information, and we hope some heads will focus on the prompts. We employ the encoder-decoder attention in the last decoder layer as the copy distribution of the copy mechanism [20, 21, 22]. To allow the model to determine the importance of attention weight from different encoder layers, we assign a learnable weight to it. To denote the copy probability of the token in context, we use:

$$\alpha_{t,l} = \sum_i^N W_{a,i} \text{softmax} \left( \frac{W_s s_t (W_{h,i} h_{l,i})^T}{\sqrt{d_k}} \right) \quad (1)$$

$h_{l,i}$  represents the representation of the  $l$ -th token in context from the  $i$ -th encoder, while  $s_t$  represents the output representation of the  $t$ -th token from the decoder.  $W_{a,i}$  is a learnable parameter, and  $W_s$  and  $W_h$  are the query and key parameters, respectively, of the scaled dot-product attention. The final copy distribution for  $y_t$  is:

$$P_{copy}(y_t) = \sum_{i:x_i=y_t} \alpha_{t,l} \quad (2)$$

### 2.2.2. Verify and Edit

The generation coefficient  $p_{gen}$  is calculated from the last encoder’s input representation  $h_l$ , with  $a_l$  and  $s_t$ . It controls whether to generate a new word or copy it from the context.

$$p_{gen} = \text{sigmoid} \left( w_c \sum_l^N a_l h_l + w_s s_t + b \right) \quad (3)$$

The final generation probability of  $y_t$  is the combination of vocabulary distribution  $P_{vocab}$  and copy distribution  $P_{copy}$ :

$$P_{vocab}(y_t) = \text{softmax}(W_v s_t + b_t) \quad (4)$$

$$P(y_t) = p_{gen} P_{vocab}(y_t) + (1 - p_{gen}) P_{copy}(y_t) \quad (5)$$

## 3. Experiments

### 3.1. Datasets

Our experiments are conducted using the CoQA dataset. Specifically, we use the train set for training our model and evaluate it on the dev set. CoQA consists of English dialogues that include questions, answers, and rationale. For our model, we use the question and rationale as the input and the answer as the output label. We consider an answer to be extractive if it overlaps with the rationale, in which case the answer is directly chosen as the rationale. In the train set, which contains 135K samples, 66.8% of the answers are extractive, meaning they are directly taken from the text, while the remaining 33.2% require rewriting, which we classify as generative samples. The dev set consists of 5.5k extractive answers and 2.4k generative samples.

### 3.2. Model Setting

Our proposed method is evaluated using the T5 base model[23], which has 220 million parameters, 6 encoder layers, and 6 decoder layers with 8 heads. We construct our experiments based on the Huggingface Transformers framework[24]. The model is fine-tuned with prompts initialized from the T5 checkpoint using the Adam optimizer with a learning rate of  $2e-5$  and cross-entropy as the loss function. We train the model for 10 epochs using a batch size of 8. Our experiments were implemented on a V100 featuring 16GB of memory, and an estimated 1.5 hours was required for each epoch of training and evaluation.

### 3.3. Comparing Methods

We assess the performance of our T5-based Prompt-Guided Copy Mechanism (PGC-T5) against the following methods:

- Raw: Human annotation of the rationale as the answer
- Vanilla-T5: Vanilla T5-base model using [SEP] to concatenate the question and rationale.
- V-PGNet-T5: T5-based pointer generator, which uses the vanilla encoder-decoder attention as the copy distribution leveraging [SEP] to concatenate the question and rationale.
- P-T5: Vanilla T5-base model with the conversation history prompt.
- P-PGNet-T5: T5-based pointer generator, which uses the vanilla encoder-decoder as the copy distribution using the conversation history prompt.

### 3.4. Evaluation

Our proposed method is evaluated and compared with others using standard CoQA evaluation metrics of EM and F1 scores. Our focus is on the performance of both extractive and generative answers, taking into consideration the dataset distribution. We report three types of EM scores: O-EM for overall samples, G-EM for generative samples, and E-EM for extractive samples. Additionally, we report three types of F1 scores: O-F1 for overall samples, G-F1 for generative samples, and E-F1 for extractive samples.

### 3.5. Main Result

Results from our CoQA dataset experiments are captured in Table 2. We observe that only 66.3% of the human-annotated rationales are considered as answers. While all models perform similarly well in extracting answers, their performance in generative tasks varies.

The vanilla T5-base model significantly improves the quality of rationales, but at the cost of the extracted answer score. Incorporating the copy mechanism causes a slight drop in the generation score of PGNet-T5. However, by utilizing the design prompt to guide the copy mechanism, PGNet-T5 achieves higher generation scores while preserving extraction scores. Our experiments show that this approach outperforms P-T5, highlighting the effectiveness of the copy mechanism in handling lengthy prompts. Furthermore, PGC-T5, attending to all encoder layers, generates more fluent and appropriate answers.

Table 2: The results of compared models on the CoQA dev set.

Models	O-EM	O-F1	G-EM	G-F1	E-EM	E-F1
Raw	66.3	73.4	0.0	18.1	<b>95.5</b>	<b>97.8</b>
T5	83.6	89.8	59.1	72.9	94.4	97.3
V-PGNet-T5	83.3	89.6	58.3	72.2	94.4	97.3
P-T5	83.9	90.1	60.3	73.7	94.3	97.3
P-PGNet-T5	84.0	90.2	60.8	74.1	94.2	97.3
PGC-T5 (version3)	<b>84.2</b>	<b>90.4</b>	<b>61.3</b>	<b>74.8</b>	94.3	97.3

### 3.6. Ablation Study

We evaluate the effectiveness of our proposed prompt through a series of experiments and analyze the results. In Table 4, we conduct a linear upgrade of our prompt and evaluate it using the PGNet-T5 + Multi-view attention. The upgraded versions include the question rationale prompt (Version 1), question description (Version 2), and conversation history prompt (Version 3). Our experiments demonstrate that the model’s performance significantly improved with the incorporation of our prompt.

**The question rationale prompt** helps the model identify the input text type and focus on the rationale text.

**The question description prompt** outperforms the question rationale prompt in generative score while maintaining the

Table 3: Typical edit types in the dev set of CoQA

Type	Question	Extractive Answer	PGC-T5 Answer
Word Class	How many planets are there away from the Sun?	the fifth planet from the Sun	Five
Pronoun	When does he not pay attention to Dan?	when I 'm well	when he's well
Tense	What did she do first?	taking a nap	took a nap
Yes-NO	Did she return safely?	Gardner was nowhere to be found	no
Simplification	Who are the 2 main actors/actresses in the film?	Ryan Reynolds wonders if marrying his boss , Sandra Bullock	Ryan Reynolds and Sandra Bullock

Table 4: The ablation experiments of proposed methods

Models	O-EM	O-F1	G-EM	G-F1	E-EM	E-F1
Base model	83.2	89.5	57.8	71.8	94.4	97.3
+version 1	83.7	90.0	59.3	73.1	94.5	97.4
+version 2	83.8	90.0	59.4	73.3	<b>94.5</b>	<b>97.4</b>
+version 3	<b>84.2</b>	<b>90.4</b>	<b>61.3</b>	<b>74.8</b>	94.3	97.3

same extractive score. The question description prompt encourages the model to generate more fluent and appropriate answers, leading to significant improvements in various question categories as illustrated in Figure 2. Please note that we only show the top 10 categories by quantity in the figure due to space constraints.

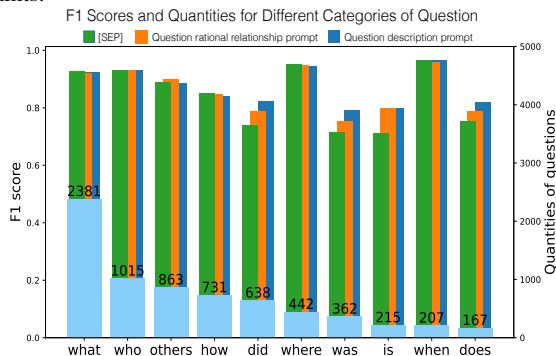


Figure 2: F1 scores and Quantities for question categories.

The conversation history prompt can significantly improve the overall score, with a generative score improvement of 1.5 F1 and 1.9 EM. This prompt can help the model better understand vague questions, handle coreference, and remove redundant answer text in the conversation history. For example, as shown in Table 5, this prompt can drop the redundant text “overdose of sedatives” from the final answer. Figure 3’s attention map provides insights into how the model directs its attention in response to the prompt. The left graph demonstrates the model’s ability to distinguish between repeated information and the answer based on the conversation history. The right graph highlights four distinct attentional components of the model, including its focus on the conversation history itself, the connection between the history and the current turn’s question, the repetition of answers in response to the current turn’s question and history, and the current turn’s input.

### 3.7. Human Evaluation

In our study, we carried out a human evaluation on the predicted development set, assessing the fluency and appropriateness of the generative samples. Table 3 shows five notable improvements in the methods concerning these aspects:

- **Word Class:** The model understands the required word class and refines potential words accordingly.
- **Pronoun:** The model correctly uses pronouns consistent

Table 5: Example of redundant text removal in CoQA Dev set

Turn	Text
1	Q: He died of what? A: an overdose
2	Q: Of what? A: sedatives
3	Q: And what else? R: overdose of sedatives and the surgical anesthetic propofol A: surgical anesthetic propofol A (version 1): overdose of sedatives and the surgical anesthetic propofol A (version 3): surgical anesthetic propofol

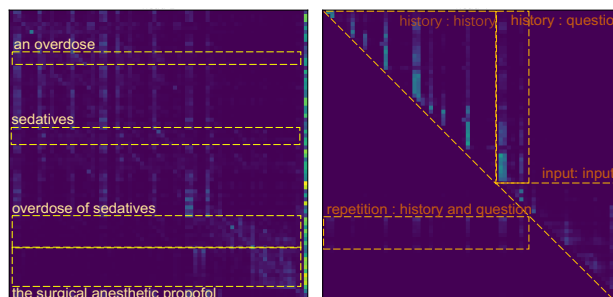


Figure 3: Visualizing Attention: Two heads of last encoder layer with the question.

- **Tense:** The model flexibly changes verb tense in the answer based on the tense of the question.
- **Yes-No:** The model recognizes yes-no questions and provides appropriate answers.
- **Simplification:** The model removes redundant text and improves the answer structure based on the question.

## 4. Conclusions

Our paper proposes a pluggable approach for extractive methods that uses a prompt-guided copy mechanism to improve the quality of extracted answers in Conversational Question Answering. The approach employs prompts to link questions to answers and attention to verify and make necessary edits for the extracted answers to be fluent and appropriate. Our experiments on the CoQA challenge show the effectiveness of our prompt modeling and multi-view attention-based copy mechanism. Further research is needed to verify its generalizability and performance under different scenarios.

## 5. Acknowledgement

This paper is supported by the Key Research and Development Program of Guangdong Province under grant No.2021B0101400003. Corresponding author is Jianzong Wang from Ping An Technology (Shenzhen) Co., Ltd (jzwang@188.com).

## 6. References

- [1] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, and M. Iyyer, “Bert with history answer embedding for conversational question answering,” in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 1133–1136.
- [2] H. Huang, E. Choi, and W. Yih, “Flowqa: Grasping flow in history for conversational machine comprehension,” in *7th International Conference on Learning Representations, ICLR 2019*.
- [3] J. Qian, B. Zou, M. Dong, X. Li, A. Aw, and Y. Hong, “Capturing conversational interaction for question answering via global history reasoning,” in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 2071–2078.
- [4] G. Kim, H. Kim, J. Park, and J. Kang, “Learn to resolve conversational dependency: A consistency training framework for conversational question answering,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6130–6141.
- [5] S. Peshterliev, B. Oguz, D. Chatterjee, H. Inan, and V. Bhardwaj, “Conversational answer generation and factuality for reading comprehension question-answering,” *arXiv preprint arXiv:2103.06500*, 2021.
- [6] C. Li, B. Bi, M. Yan, W. Wang, and S. Huang, “Addressing semantic drift in generative question answering with auxiliary extraction,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 942–947.
- [7] C. Tan, F. Wei, N. Yang, B. Du, W. Lv, and M. Zhou, “S-net: From answer extraction to answer synthesis for machine reading comprehension,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, 2018, pp. 5940–5947.
- [8] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, “Gpt understands, too,” *arXiv preprint arXiv:2103.10385*, 2021.
- [9] S. Wang, Y. Xu, Y. Fang, Y. Liu, S. Sun, R. Xu, C. Zhu, and M. Zeng, “Training data is more valuable than you think: A simple and effective method by retrieving from training data,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3170–3179.
- [10] W. Zhong, Y. Gao, N. Ding, Y. Qin, Z. Liu, M. Zhou, J. Wang, J. Yin, and N. Duan, “ProQA: Structural prompt-based pre-training for unified question answering,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 4230–4243.
- [11] S. Reddy, D. Chen, and C. D. Manning, “Coqa: A conversational question answering challenge,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.
- [12] Y. Su, Y. Wang, S. Baker, D. Cai, X. Liu, A. Korhonen, and N. Collier, “Prototype-to-style: Dialogue generation with style-aware editing on retrieval memory,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2152–2161, 2021.
- [13] O. Vinyals, M. Fortunato, and N. Jaitly, “Pointer networks,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, 2015, pp. 2692–2700.
- [14] A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1073–1083.
- [15] S. Liu, D. Wang, X. Li, M. Huang, and M. Ding, “A copy-augmented generative model for open-domain question answering,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022, pp. 435–441.
- [16] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does bert look at? an analysis of bert’s attention,” in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 276–286.
- [17] G. Jawahar, B. Sagot, and D. Seddah, “What does BERT learn about the structure of language?” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3651–3657.
- [18] A. Bapna, M. Chen, O. Firat, Y. Cao, and Y. Wu, “Training deeper neural machine translation models with transparent attention,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3028–3033.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [20] S. Xu, H. Li, P. Yuan, Y. Wu, X. He, and B. Zhou, “Self-attention guided copy mechanism for abstractive summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1355–1362.
- [21] B. Bi, C. Li, C. Wu, M. Yan, W. Wang, S. Huang, F. Huang, and L. Si, “Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8681–8691.
- [22] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung, “Transferable multi-domain state generator for task-oriented dialogue systems,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 808–819.
- [23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.