Auto-ACE: An Automatic Answer Correctness Evaluation Method for Conversational Question Answering

Zhixin Bai[1*], **Bingbing Wang**[2*], **Bin Liang**[3†], **Ruifeng Xu**[2,4†]
[1] Harbin Institute of Technology, Harbin, China
[2] Harbin Institute of Technology, Shenzhen, China
[3] The Chinese University of Hong Kong, Hong Kong, China
[4] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
{baizhixin,bingbing.wang}@stu.hit.edu.cn,
bin.liang@cuhk.edu.hk, xuruifeng@hit.edu.cn

Abstract

Conversational question answering aims to respond to questions based on relevant contexts and previous question-answer history. Existing studies typically use ground-truth answers in history, leading to the inconsistency between the training and inference phases. However, in real-world scenarios, progress in question answering can only be made using predicted answers. Since not all predicted answers are correct, indiscriminately using all predicted answers for training introduces noise into the model. To tackle these challenges, we propose an automatic answer correctness evaluation method named **Auto-ACE**. Specifically, we first construct an Att-BERT model which employs attention weight to the BERT model, so as to bridge the relation between the current question and the question-answer pair in history. Furthermore, to reduce the interference of the irrelevant information in the predicted answer, A-Scorer, an answer scorer is designed to evaluate the confidence of the predicted answer. We conduct a series of experiments on QuAC and CoQA datasets, and the results demonstrate the effectiveness and practicality of our proposed Auto-ACE framework.

1 Introduction

Conversational Question Answering (ConvQA) involves responding to a sequence of questions within a conversation, while considering the relevant context provided (Qu et al., 2020; Pearce et al., 2023; Reddy et al., 2019). Different from the traditional extractive question-answering tasks which conduct one-turn dialog, as shown in Figure 1, ConvQA is expected to resolve such implicit information from the conversational history in a multi-turn way.

With the rise of virtual assistants and chatbots, ConvQA has recently garnered increased interest. Hence, numerous works have been conducted for further study. Raposo et al. (2022) proposed a conversational question answering system specifically designed for the Search-Oriented Conversational AI (SCAI) shared task, and provided a detailed analysis of its question rewriting module. Qu et al. (2019b) introduced a positional history answer embedding method to encode conversation history with positional information using BERT (Devlin et al., 2018). They also designed a history attention mechanism (HAM) for each question-answer pair and utilized multi-task learning to predict the final answer. Nevertheless, despite their successes, these works on ConvQA rely on the ground-truth answer, overlooking the fact that real-world progress can only be achieved using predicted answers.

Existing researchers found a way to tackle this limitation by using the predicted label (Mandya et al., 2020; Christmann et al., 2022). This method can partially trade off the balance between training and inference. However, if the predicted answer is incorrect, it will introduce noisy samples into the model, thereby affecting performance. For example, as shown in Figure 1, $\{Q_2, A_2, Q_1, A_1\}$ are

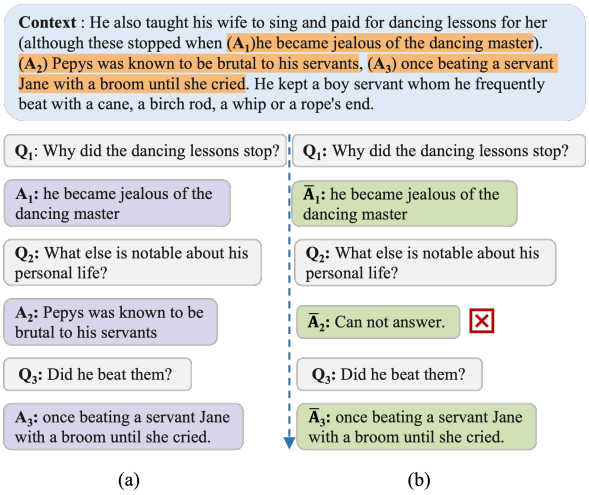

Figure 1: Examples of using (a) ground-truth answers and (b) predicted answers.

*These authors contributed equally to this work.
†Corresponding author.

Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pages 80–87

input as the conversation history of Q_3 into the Question Answering (QA) model to perform inference for A_3. Figure 1(a) indicates that ground-truth answers are used for inference, which significantly differs from the real-world inference scenarios. Figure 1(b) indicates the use of all predicted answers for inference. Although this way is more practical, it introduces noise into the model to some extent when the predicted answers are incorrect. Therefore, we propose an answer scorer model that can automatically assign attention weights to predicted answers and incorporate them into the QA model's inference phase. The most similar work to ours is the work of Jeong et al. (2023), which requires an initial round of training and prediction to obtain the predicted answers along with their confidences and uncertainties before the official training. This additional training and prediction step increases the overall training time and computational resource consumption.

In this paper, we propose an automatic answer correctness evaluation method named Auto-ACE, which comprises an Att-BERT and an A-Scorer method to maximize the use of effective information from the predicted answers. To be more specific, we first construct an Att-BERT model which employs attention weight to the BERT model, so as to bridge the relation between the current question and the question-answer pair in history. Furthermore, to reduce the disturbance of the irrelevant content in the predicted answer, A-Scorer, an answer scorer is designed to evaluate the confidence of the predicted answer. During the training phase, Att-BERT and A-Scorer are trained, while in the inference period, A-Scorer evaluates each question-answering pair in the history to obtain the correctness of the predicted answer. Numerous experiments on QuAC and CoQA datasets demonstrate the effectiveness and practicality of our proposed Auto-ACE framework[1].

The main contributions of our work can be summarized as follows:

- We propose an Auto-ACE framework to establish the connection between the current question and historical question-answer pairs, balancing the process between training and inference phases.
- To bridge the relation between the current question and the historical question-answer pair, Att-BERT is designed. Moreover, we devised the A-Scorer, which is trained with Att-BERT during the training phase and evaluates the correctness of the predicted answer during the inference phase to mitigate the impact of erroneous predicted answers and maximize the utilization of historical conversation information.
- Our Auto-ACE framework achieves excellent performances on QuAC and CoQA datasets, which shows our approach is effective.

[1] https://github.com/baibaizhixin/Auto-ACE

2 Related Works

2.1 Conversational Question Answering

ConvQA is an extension of the QA task which aims to train a model that can answer the question by means of understanding the context of the given context and the previous conversational questions and answers. In the work by Nishida and Tomita (2019), BERT is utilized to encode contexts independently conditioned with each question and answer within a multi-turn context. This process enables the method to predict answers based on the context representations encoded with BERT. Qu et al. (2019a) presented a distinct method termed history answer embedding, which incorporates conversation history into a ConvQA model built on BERT. Query rewriting became a popular technique for ConvQA. Vakulenko et al. (2021) addressed question ambiguity by rewriting them, ensuring they can be effectively processed by existing QA models as standalone questions, independent of the conversation context. Wu et al. (2022) introduced a query rewriting model tailored for converting conversational questions within a context into standalone queries. This model is trained using a novel reward function, optimized directly for retrieval via reinforcement learning. Although the above studies attain excellent performance in ConvQA, they ignore the unbalance between training and inference phases due to the utilization of ground truth or predicted answers.

2.2 Score-based Methods

Score-based methods have gained significant attention in various Natural Language Processing (NLP) tasks due to their capability to enable models to selectively focus on relevant parts of the input sequence. For example, Osama et al. (2020) introduced the Score-Based Ambiguity Detector and Resolver method. This system uses Stanford

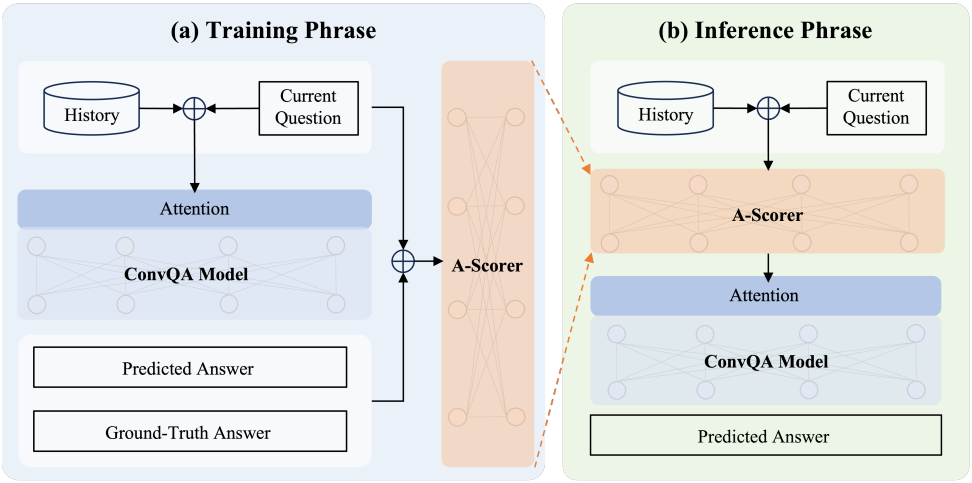

Figure 2: A figure

CoreNLP to generate possible parse trees for each sentence in a given textual requirement. It then analyzes these parse trees through four filtering pipelines to detect syntactic ambiguities and suggest multiple possible interpretations, effectively resolving the ambiguities. Several attempts have been made to enable self-attention to learn dependencies between words in a sentence and capture the sentence's inner structure (Tan et al., 2018; Cao et al., 2018). Liu et al. (2021) devised an attention score-based word rank approach, incorporating a word sequence encoder and a word-level attention layer. Despite the extensive work on score-based methods in various natural language tasks, their application in ConvQA remains under-explored. This is particularly important when historical questions and answers contain implicit information, making the predicted answer unusable directly.

3 Methodology

This section begins with a concise introduction to the ConvQA task, then we describe how the proposed Auto-ACE framework can bridge the gap between training and real-world inference scenarios by incorporating predicted answers into model training, as demonstrated in Figure 2. In addition, we discuss the calculation of attention weights and the overall training pipeline.

3.1 Conversational Question Answering

We first provide a general description of the ConvQA task. For the i-th turn of the conversation, a question Q_i and its corresponding context C are given, as well as a conversation history H_i composed of previous questions and answers: $H_i = \{Q_{i-1}, A_{i-1}, ..., Q_1, A_1\}$. Then, the goal of ConvQA is to correctly extract the answer A_i from C, along with Q_i and H_i, as shown below:

$$\begin{aligned} P(A_i) &= P(A_i \mid C, Q_i, \mathcal{H}_i) \\ &= M_\theta(C, Q_i, Q_{i-1}, A_{i-1}, \ldots, Q_1, A_1) \end{aligned} \tag{1}$$

where M_θ is the ConvQA model.

In previous work, some of them assumed that the ground-truth answers $\{A_{i-1}, A_{i-2}, ..., A_1\}$ are available in the inference phase, as shown in Equation 1. However, this setup is far from reality because progress in the real world can only be made using the predicted answer. If the training process always uses ground-truth answers, it will lead to the model not performing well in real-world inference scenarios. Another part of them recognized this and tried to select whether to include the predicted answer of a certain historical turn in the conversation history by setting a threshold. However, it often requires an additional step of training to calculate the confidence of all predicted answers and determine the value of the threshold, which increases the overall training time and computational resource consumption. Therefore, we modify the formulation in Equation 1 to bridge the gap between the training and the inference phase, which we will describe in the following section.

3.2 Training with Predicted Answers

As delineated in Section 3.1, employing ground-truth answers during model training and predicted answers during inference is inadvisable. To align the model's training phase more closely with real-world inference scenario, a rational strategy entails utilizing the model's prior predictions as inputs to

the conversation history for subsequent turns of prediction, as follows:

$$P(A_i) = M_\theta\big(C, Q_i, Q_{i-1}, \overline{A}_{i-1}, \ldots, Q_1, \overline{A}_1\big) \quad (2)$$

where $\{\overline{A}_{i-1}, \overline{A}_{i-2}, ..., \overline{A}_1\}$ are the predicted answers.

Given that the accuracy of the model's predictive answers is not infallible, incorporating erroneous predictions into the conversation history may introduce superfluous noise, thereby potentially degrading the efficacy of the predictions. Therefore, we propose an Att-BERT model that applies attention weights to the BERT model, giving higher weights to answers with high confidence and lower weights to answers with low confidence, which allows for the use of predicted answers during training while minimizing the noise caused by incorrect predicted answers.

To be more specific, we assign attention weights to each turn's question Q_j and predicted answer $\overline{A}_j$ in the conversation history. The weight of the Q_j represents the degree of relevance to the current question, and on this basis, the weight of the $\overline{A}_j$ also represents the confidence of the predicted answer, as shown in the following.

$$\begin{aligned} P(A_i) = M_\theta\big(C, Q_i, W^q_{i-1}Q_{i-1}, W^a_{i-1}\overline{A}_{i-1}, \\ \ldots, W^q_1 Q_1, W^a_1 \overline{A}_1\big) \end{aligned} \quad (3)$$

Considering the real-world inference scenario, we assign attention weights to the questions and answers at each turn of the conversation history. The attention weight W^q_j of the question Q_j represents the degree of relevance to the current question Q_i, that is, the more similar Q_j is to the Q_i, the more attention the model will give to this sequence. Notably, the attention weight is complemented by the cosine similarity between Q_j and Q_i, as shown in Equation 4. Since all questions in the conversation history are provided in the inference scenario, the whole attention weights can be directly calculated.

$$W^q_j = \text{Similarity}(Q_j, Q_i) \quad (4)$$

where W^q_j represents the attention weight of the question in the j-th turn of the conversation history, Similarity is used to compute the cosine similarity.

3.3 Confidence-based Attention Calculation

In this section, we aim to enhance the model's focus on the most relevant content of the predicted answers. However, it is impractical to calculate weights for all predicted answers, as some of them may be incorrect. To address this challenge, an A-Scorer is devised to automatically evaluate the confidence of predicted answers and is trained in conjunction with the Att-BERT model. In specific, after the Att-BERT model generates a predicted answer each turn, we input the question Q_j, the predicted answer $\overline{A}_j$, and the corresponding context C into the A-Scorer model to evaluate the confidence of the $\overline{A}_j$, and use the cosine similarity between the predicted answer and the actual answer as the ground truth for the confidence.

$$W^a_j = W^q_j \times \text{A-Scorer}(Q_j, \overline{A}_j) \quad (5)$$

where W^q_j and W^a_j represent the attention weights of the question and the predicted answer in the j-th turn of the conversation history, A-Scorer is the model we proposed for automatic confidence evaluation.

Following the joint training of the Att-BERT model and the A-Scorer model, these two models become capable of operating in concert with real-world inference scenarios. The answer predicted by the Att-BERT model is evaluated for confidence by the A-Scorer model. Furthermore, during the prediction of an answer, the attention weight of each question or answer within the conversation history is ascertained contingent upon the predicted answer's confidence, as well as the degree of correspondence to the current question.

3.4 Overall Pipeline

In this subsection, we describe the training pipeline for the Att-BERT model and the A-Scorer model, which are trained together in a single step and are also applied together in the inference phase.

We divide the training data into batches, ensuring that (1) the same batch does not contain examples from the same conversation, and (2) for any two examples from the same conversation, the batch of the example that appears later in the conversation is also later. We do this to ensure that when an example is input into the model, all predicted answers for the questions in its conversation history have already been obtained, thus ensuring that only predicted answers are used during the training phase, not the ground-truth answers. Then, we train the Att-BERT model and the A-Scorer model together following the training protocol in Equation 3. The Att-BERT model assigns different attention weights to the questions and predicted

answers in the conversation, while the A-Scorer model evaluates the confidence of the answers predicted by the Att-BERT model.

For evaluation, we still use Equation 3 as the actual evaluation protocol. Instead of using ground-truth answers or sampling predicted answers based on the confidence obtained during training, we directly apply the predictions of the A-Scorer model to the attention weights of the Att-BERT model. Doing so not only bridges the gap between training and real-world inference scenarios but also avoids the need for additional training steps and reduces the demand for excessive computational resources.

4 Experiments

4.1 Datasets and metrics

QuAC (Choi et al., 2018) is a benchmark ConvQA dataset, which comprises 14K conversations and 100K question-context pairs and is designed to simulate realistic information-seeking conversations. In QuAC, questioners did not have access to the contexts during data collection. Since the test set is not publicly available, we use the development set for evaluation.

CoQA (Reddy et al., 2019) is another ConvQA dataset, containing 127K question-context pairs. Similar to QuAC, we use the development set for CoQA as the test set is not publicly accessible.

F1-score: To assess the performance of our models, we use the F1-score as the evaluation metric. This follows the standard evaluation protocol established by (Kim et al., 2021). The F1-score is a widely recognized metric that balances precision and recall, making it particularly suitable for evaluating the quality of predictions in natural language tasks.

Baselines: We compare Auto-ACE with several relevant baselines. Except for the gold and No Pred models, all other models used predicted answers as the conversation history in the inference phase.

- **Gold**: which uses an unrealistic setting in both training and inference phases, using ground-truth answers as conversation history.
- **No Pred**: which does not use predicted answers during training and inference.
- **All Pred**: which retains all predicted answers as conversation history during both training and inference.

Method	QuAC		CoQA	
	BERT	RoBERTa	BERT	RoBERTa
Gold†	59.86	65.08	72.79	77.62
No Pred†	55.44	61.24	70.83	75.56
All Pred†	55.76	61.53	71.28	75.42
CoQAM†	55.83	61.55	71.27	74.29
AS-ConvQA†	57.06	62.18	71.99	76.76
Auto-ACE (ours)	**58.38**	**63.04**	**72.56**	**77.29**

Table 1: Performance(%) on QuAC and CoQA. **Bold** indicates the model with the best performance. Results with † comes from Jeong et al. (2023).

- **CoQAM**: which dynamically adjusts the sampling rate to alternately select ground truth answers or predicted answers during training, and uses predicted answers during inference phase.
- **AS-ConvQA**: This method decides whether to include the predicted answer in the conversation history during the training and inference phases based on the confidence and uncertainty of the predicted answer.

4.2 Main Results

As shown in Table 1, the Auto-ACE framework, which includes an Att-BERT and an A-Scorer model, demonstrates significant performance improvements across all baselines. The evaluation results show that, our method outperforms the strongest baseline that does not use ground-truth answers by 1.32%. In addition, our model can be trained in one step, unlike AS-ConvQA, which requires additional training and prediction steps. It should be noted that since the Gold model uses an unrealistic evaluation setting where ground-truth answers are used as conversation history, it is not fair when compared with other methods.

It is worth mentioning that the No Pred methods outperform those using predicted answers or heuristic sampling of conversation history, which demonstrates incorrect predicted answers can introduce noise to the QA model. Moreover, our method shows a significant advantage, it might attribute to the A-Scorer can automatically evaluate the confidence of predicted answers, allowing the QA model to truly focus on relevant and correct answers, and minimize the impact of noise at the same time.

4.3 Ablation Study

We also conducted an ablation study on the use of the attention mechanism. Specifically, we di-

Method	QuAC		CoQA	
	BERT	RoBERTa	BERT	RoBERTa
Auto-ACE	**58.38**	**63.04**	**72.56**	**77.29**
w/o attention(Q&Q)	57.32	62.08	70.95	75.29
w/o attention(Q&A)	57.88	62.25	71.19	76.68
w/o attention(both)	56.46	60.98	69.23	74.57

Table 2: Performance (%) of ablation study on QuAC and CoQA datasets. **Bold** indicates the model with the best performance.

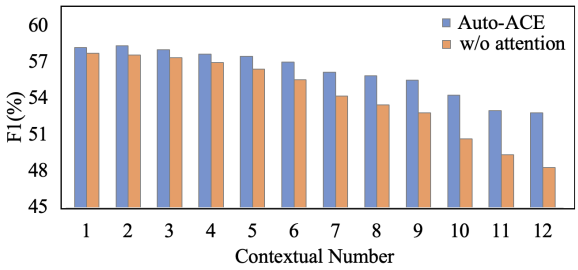

Figure 3: Results(%) of the effect of the different contextual number.

vided it into three scenarios: not considering the attention between the current question and the historical question-answer pairs, not considering the attention to the predicted answers in the history, and not using the attention mechanism at all. The evaluation results are shown in Table 2. The performances of all ablation models are worse than the complete model, which demonstrates the necessity of the attention mechanism.

The most significant performance drop occurs when neither of the two attention mechanisms is considered, so using either one of them alone can improve the performance, indicating that both of them are effective. An interesting finding is that disregarding the attention between the current and previous questions (Q&Q) often results in worse performance than disregarding the attention to the predicted answer (Q&A), indicating that the similarity between the current question and the historical question-answer pairs seems to have a greater impact on the model's performance.

4.4 Difference of Evaluator

To demonstrate that our proposed Auto-ACE framework can deliver robust performance across different QA models, we also evaluated it by replacing the evaluator with Roberta, with the results shown in Table 1. It can be seen that our model has performed well in both configurations using BERT and Roberta as evaluators. In the configuration using RoBERTa as an evaluator, Auto-ACE improved the F1 scores on the quac and coqa datasets by % and % compared to the best-performing baseline, respectively.

4.5 Effect of the Contextual Number

To examine and analyze the impact of the max length of utterances over the performance of our proposed Auto-ACE framework, we conduct experiments by varying the max length from 1 to 12. Since the maximum number of conversation turns in all sessions of the QuAC dataset is 12, setting the contextual number to 12 means that all samples retain the complete conversation history. In other cases, samples retain the most recent contextual number of conversation turns' history.

We plotted the experimental results in a bar chart and demonstrate them in Figure 3. It should be noted that, to consider the relationship between the attention in our proposed Att-BERT model and the contextual number, we also evaluated the model without applying the attention under different contextual number settings. From Figure 1, it can be seen that as the contextual number increases, the model's performance gradually decreases, with the best results for the model being achieved when the contextual number is set to 1-3. This also conforms to our intuition: the more recent Q&A turns in the conversation history tend to be more relevant to the current question.

Another point worth noting is: with the increase of contextual number, although the model's performance declines, the model without attention to the conversation history declines more significantly than our model. This is because, even though there are many irrelevant Q&A pairs in the long conversation history, the proposed Auto-ACE model can allocate attention to the conversation history based on relevance and predicted answer's confidence, thus allowing the model to focus on the information that is relevant and reliable to the current question in a long sequence.

4.6 Case Study

Two representative conversation scenarios are provided in Figure 4. These two examples demonstrate that our method of weighting the conversation history in the form of attention is highly practical and meaningful. From example (a), we can observe that the current question Q_3 has a strong correlation with the historical question Q_1, because "first film" in the current question refers to the answer

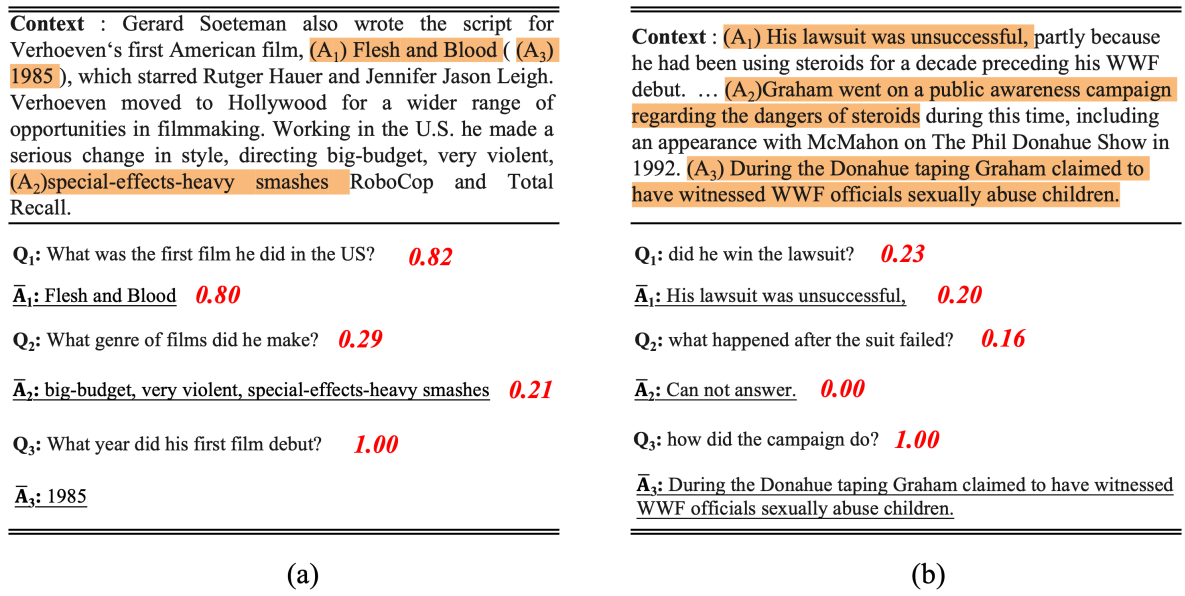

Figure 4: Examples of applying the attention weight to the history.

of question Q_1:"Flesh and Blood". In our method, due to the high similarity to Q_3, Q_1 is assigned a high attention weight. $\overline{A}_1$, as the predicted answer for Q_1, also receives a high final attention weight because the A-Scorer deems it to have a high degree of confidence. When these weights are applied in the form of attention to Att-BERT, the model can focus more on the useful information in the history: Q_1 and $\overline{A}_1$, and thus it is easier to predict the correct answer.

The opposite scenario is depicted in Figure 4(b). Although the predicted answer $\overline{A}_2$ is crucial for the current question as it includes the keyword "campaign" from Q_3, the model's prediction for question Q_2 is "Can not answer" at this point, which could introduce noise into the model when being used as part of the conversation history. In our method, "Can not answer" is assigned an attention weight of 0 by the A-Scorer, hence its final attention weight is 0. The Att-BERT does not pay attention to this incorrect answer, thus not affecting the prediction of the answer for the current turn.

5 Conclusion

In this paper, we introduce an automatic answer correctness evaluation method named Auto-ACE for ConvQA task, which can balance the inconsistency between training and inference. The proposed Auto-ACE method consists of two primary components including Att-BERT and A-Scorer. The Att-BERT effectively bridges the current question with historical Q&A pairs using attention mechanisms, enabling the model to focus on more relevant content. Furthermore, the A-Scorer is designed to evaluate the confidence of predicted answers and is applied to the Att-BERT as the confidence-based attention. Experiments conducted on QuAC and CoQA datasets demonstrate that our proposed Auto-ACE method significantly improves the performance and reliability of other baseline models.

Limitations

Although the Auto-ACE framework demonstrates promising results in the Conversational Question Answering task, there are still some limitations that require further attention: 1) The model's capability to process lengthy conversational histories needs enhancement to ensure consistent performance. In the future, we will consider the richness of real-world conversations to improve the model's performance. 2) The A-Scorer may still introduce noise due to inappropriate evaluation of predicted answers, future work could consider employing large language models to further enhance the accuracy of answer evaluation.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (62176076), the Natural Science Foundation of Guangdong (2023A1515012922), the Shenzhen Foundational Research Funding (JCYJ20220818102415032),

and the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

References

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 182–192.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2022. Conversational question answering on heterogeneous sources. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 144–154.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Soyeong Jeong, Jinheon Baek, Sung Ju Hwang, and Jong C Park. 2023. Realistic conversational question answering with answer selection based on calibrated confidence and uncertainty measurement. *arXiv preprint arXiv:2302.05137*.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. 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)*, pages 6130–6141.

Yueyang Liu, Hunmin Lee, and Zhipeng Cai. 2021. An attention score based attacker for black-box nlp classifier. *arXiv preprint arXiv:2112.11660*.

Angrosh Mandya, James O'Neill, Danushka Bollegala, and Frans Coenen. 2020. Do not let the history haunt you: Mitigating compounding errors in conversational question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2017–2025.

Yasuhito Ohsugi Itsumi Saito Kyosuke Nishida and Hisako Asano Junji Tomita. 2019. A simple but effective method to incorporate multi-turn context with bert for conversational machine comprehension. *ACL 2019*, page 11.

Mohamed Osama, Aya Zaki-Ismail, Mohamed Abdelrazek, John Grundy, and Amani Ibrahim. 2020. Score-based automatic detection and resolution of syntactic ambiguity in natural language requirements. In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 651–661. IEEE.

Kate Pearce, Sharifa Alghowinem, and Cynthia Breazeal. 2023. Build-a-bot: teaching conversational ai using a transformer-based intent recognition and question answering architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16025–16032.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. 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*, pages 1133–1136.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1391–1400.

Gonçalo Raposo, Rui Ribeiro, Bruno Martins, and Luísa Coheur. 2022. Question rewriting? assessing its importance for conversational question answering. In *European Conference on Information Retrieval*, pages 199–206. Springer.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 355–363.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014.