Combining Impression Feature Representation for Multi-turn Conversational Question Answering

Shaoling Jing[1,2,3,*] and **Shibo Hong**[2] and **Dongyan Zhao**[1] and **Haihua Xie**[2] and **Zhi Tang**[1]
1. Wangxuan Institute of Computer Technology, Peking University, Beijing, China
2. State Key Laboratory of Digital Publishing Technology,
Peking University Founder Group Co. LTD., Beijing, China
3. Postdoctoral Workstation of the Zhongguancun Haidian Science Park, Beijing, China
{jingshaoling, zhaody, tangzhi}@pku.edu.cn
{hongshibo, xiehh}@founder.com

Abstract

Multi-turn conversational Question Answering (ConvQA) is a practical task that requires the understanding of conversation history, such as previous QA pairs, the passage context, and current question. It can be applied to a variety of scenarios with human-machine dialogue. The major challenge of this task is to require the model to consider the relevant conversation history while understanding the passage. Existing methods usually simply prepend the history to the current question, or use the complicated mechanism to model the history. This article proposes an impression feature, which use the word-level inter attention mechanism to learn multi-oriented information from conversation history to the input sequence, including attention from history tokens to each token of the input sequence, and history turn inter attention from different history turns to each token of the input sequence, and self-attention within input sequence, where the input sequence contains a current question and a passage. Then a feature selection method is designed to enhance the useful history turns of conversation and weaken the unnecessary information. Finally, we demonstrate the effectiveness of the proposed method on the QuAC dataset, analyze the impact of different feature selection methods, and verify the validity and reliability of the proposed features through visualization and human evaluation.

1 Introduction

Conversational Question Answering (ConvQA) is a new question answering task that requires a comprehension of the context, which has recently received more and more attention (Zhu et al., 2018; Qu et al., 2019a; Qu et al., 2019b; Meng et al., 2019; Pruthi et al., 2020). Since conversation is one of the most natural ways for humans to seek information, it carries over context through the dialogue flow. Specifically, we ask other people a question, dependending on their answer, we follow up with a new question, and second answer with additional information will be given based on what has been discussed (Reddy et al., 2019). Therefore, multi-turn conversational question answering is formed in this way. It can be used in many fields as a personal assistant systems, such as, customer service, medical, finance, education, etc. Moreover, with the rapid development of artificial intelligence technology in theory and practical applications, many personal assistant products have been launched in the market, such as Alibaba AliMe, Apple Siri, Amazon Alexa, etc. Although these assistants are capable to cover some simple tasks, they cannot handle complicated information-seeking conversations that require multiple turns of interaction (Qu et al., 2019b).

In the tasks of two recent multi-turn ConvQA datasets, CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018), given a passage, a question, and the conversation context preceding the question, the task is to predict a span of passage as the answer or give an abstrctive answer based on the passage. So the machine has to understand a text passage and converstion history to answer a series of questions. Each conversation in the QuAC dataset is obtained by two annotators playing the roles of teacher (information-provider) and student (information-seeker) respectively. During the conversation, the student only has access to the heading of passage and tries to learn about a hidden Wikipedia passage by asking a sequence

Proceedings of the 19th China National Conference on Computational Linguistics, pages 863-873,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

of freeform questions. The teacher answers the question by providing a span of text in the passage, as in existing reading comprehension tasks SQuAD (Rajpurkar et al., 2016), and gives the dialog acts which indicate the student whether the conversation should follow up. The CoQA has abstractive answers involving adding a pronoun (Coref) or inserting prepositions and changing word forms (Fluency) to existing extractive answers (Yatskar, 2018). Both datasets contain yes/no questions and extractive answers. Compared with the CoQA[1], the QuAC[2] setting is similar to a user query on search engines. The latter is designed to model and understand information-seeking conversation, which is closer to the people's daily question-answering style conversation than other datasets. On the other hand, QAs in QuAC are mostly non-factoid QAs and 86% of the 100 questions are contextual questions which requires reading the history to resolve coreference to dialog and passage. Moreover, the main answer type of QuAC dataset is extractive, resulting experiments are not easily disturbed by other types of answer factors and are suitable for verifying the feasibility of the proposed method. Therefore, this article intends to use the QuAC dataset for ConvQA experiments.

Existing multiple turns of question answering methods (Qu et al., 2019b; Zhu et al., 2018; Yatskar, 2018; Huang et al., 2018) emphasize the influence of historical context on current questions. Some of methods (Zhu et al., 2018; Reddy et al., 2019) prepend history turns to the current question or use a recurrent structure to model the representations of history turns (Huang et al., 2018), which obtain a good performance but a lower training efficiency. Some methods (Choi et al., 2018) adopt a simple heuristic method to select immediate previous turns, but they do not work for complicated conversational behaviors. Some researches attend history embedding (Qu et al., 2019a) or attend history position to the current question (Qu et al., 2019b), but not applicable to several no span-based answers. In addition, according to the literature available, there is a great lack of public studies on selecting or re-weighting of the conversation history turns, and re-representing the current questions and passages. Therefore, this paper proposes an impression feature combined with conversational history. By simulating the process of human question answering, we calculate the correlation from the deep historical context to the current question and the complete semantic unit of the passage to form impression features, and use this feature to replace the position information. This solves the problem that the abstrctive answer is difficult to learn position information, and enhance the knowledge representation ability of the model.

In this paper, we propose a multi-turn conversational question answering model combining with impression features. In order to learn the useful information from the conversation history, we separately calculate the word-level inter attention and turn inter attention from the conversation history to the current question and the passage. Then the learned representation is used as impression feature and fed to BERT (Devlin et al., 2018) with other inputs. The final representation is used to predict the answers.

Therefore, the contributions are as follows:

(1) Design an impression feature representation. This feature helps the model to learn more accurate information from the context of the historical conversation turns and assists the model in understanding passage and conversation, which provides new insights to the ConvQA task.

(2) Adapt different feature selection methods to verify the impact of the proposed impression feature representation on the model.

(3) A multiple turn conversational question answering model combining impression features is proposed.

2 Related Work

ConvQA is closely related to Machine Reading Comprehension (MRC) and conversational system.

The ConvQA task is similar to the machine reading comprehension task (Rajpurkar et al., 2016), but the major difference from MRC is that the questions in ConvQA are organized in conversations (Qu et al., 2019b), such as CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018). Some questions rely on the historical questions or answers through pronouns. For instance, there are two questions Q_1, Q_2 and an answer A_1 to Q_1. Q_1:"who is going to have a birthday?". A_1:"Grandma Li.". Q_2:"where she was

[1]https://stanfordnlp.github.io/coqa/

[2]http://quac.ai/

Proceedings of the 19th China National Conference on Computational Linguistics, pages 863-873,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

born?". Here, the pronoun "she" of Q_2 associates Q_2 with Grandma Li of A_1, which indicates that the A_2 depend on A_1. If the QA model does not use Q_1 and A_1, then it does not know who she refers to in Q_2, making it difficult for the model to accurately answer Q_2. However, the questions of traditional MRC datasets (such as SQuAD (Rajpurkar et al., 2016) and SQuAD2.0 (Rajpurkar et al., 2018)) are independent of each other and have no relevance. Compared with the traditional MRC task, multi-turn ConvQA based on MRC adds multiple turns of conversation history to the original MRC task, making the ConvQA task more suitable for human daily conversation habits.

The exsiting methods for ConvQA in (Qu et al., 2019a) and (Qu et al., 2019b) determine whether the token in the question and the passage appear in each round of the historical conversation, and take the distance from the history turn of answers to the current question as the relative position, finally use the embedding of the relative position as an input of BERT encoder (Devlin et al., 2018). These methods are simple and effective, but they are not applicable to some no span-based answers. Because the token in the abstractive answer may be synonymous with a word in the historical answer, not the same word. In this case, the relative position is invalid. Moreover, a large amount of redundant information may also be introduced, and there may be a possibility of over-learning. For example, for a long passage, the author divides the passage into several sub-passages, and learns the relationship between each sub-passage and the answers of the historical rounds. If a question is only related to one of the sub-passages, suppose p_0, and has nothing to do with another sub-passage p_1. The information learned by p_1 and the largely redundant information of history conversation turns might play a negative interference role for the model to find the answer, while answering the current question q_k. Therefore, this paper focuses on how to select historical context and integrate its information into current question and passage.

ConvQA is very similar to the Background Based Conversations (BBCs) which recently proposed in the field of conversational systems. The latter is proposed to generate a more informative response based on unstructured background knowledge. But most of the research is aimed at topic-specific field (Meng et al., 2019), such as the conversation for movies (Moghe et al., 2018; Zhou et al., 2018) and diverse set of topics of Wikipedia (Dinan et al., 2018). Therefore, question answering based on reading comprehension and BBCs, these two tasks have in common that when responsing to each current sentence, not only the passage or background, but also the historical conversational context must be considered. The difference is that the former pays more attention to the ability of the model to understand the passage. When answering questions, the passage is mainly learned, and the historical conversation is supplemented to make the answer more accurate. The latter pays more attention to the ability of the model to understand the conversational context. When making a response, the model mainly learns conversational context, and assists with reference to background knowledge, the purpose is to enable the conversation to continue while making the response more informative.

In terms of model structure, RNN-based structure and BERT-based model (Devlin et al., 2018) have certain effectiveness on ConvQA, MRC and BBCs tasks. The RNN-based model (Zhu et al., 2018) can learn the impact of historical questions and answers on the current question and passage, but it cannot learn the deep bidirectional context representation. The BERT-based model is proved to greatly improve the performance of ConvQA (Qu et al., 2019a; Qu et al., 2019b), but it lacks reasonable integration into the history turns of conversation. Therefore, this paper proposes a method to model the history turns of questions and answers, generate impression features, and integrate them into the current questio and passage to improve model performance.

3 Our Approach

3.1 Task and Notations Definition

The ConvQA task is defined as (Reddy et al., 2019) and (Choi et al., 2018), given a passage x, the k-th question q_k in the conversation and the history conversation H_k preceding q_k, the task is to predict the answer a_k to the question q_k. There are only extractive answers in dataset QuAC (Choi et al., 2018). So the task is to predict the text span a_k within passage x. For the question q_k, there is $k-1$ turns of history conversation, and i-th turn of history conversation H_k^i includes a question q_i and its groudtruth answer a_i, which is $H_k^i = \{q_k^i, a_k^i\}_{i=1}^{k-1}$.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 863-873,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

In order to ensure that the latter part of the long passage can be learened by the model, we divide the given passage x into N parts with sliding window following the previous work (Devlin et al., 2018), it is denoted as $x = \{x_n\}_{n=1}^{N}$ and $x_n = \{x_n(t)\}_{t=1}^{T}$, where $x_n(t) \in \mathbb{R}^h$ refers to the representation of the t-th token in x_n, T is the sequence length and h is the hidden size of the token representation. The k-th question is denoted as $q_k = \{q_k(j)\}_{j=1}^{J}$, $q_k \in \mathbb{R}^{J \times h}$, where $q_k(j) \in \mathbb{R}^h$ refers to j-th token in q_k and J is the maximum question length. All $k-1$ turns of history question and answer sequences are represented as $H_k = \{H_k^i\}_{i=1}^{I}$, $H_k \in \mathbb{R}^{I \times M \times h}$, where I is the maximum number of history turns for all conversations. The i-th turn history conversation of the k-th question is denoted as $H_k^i = \{H_k^i(m)\}_{m=1}^{M}$, $H_k^i \in \mathbb{R}^{M \times h}$, where $h_k^i(m) \in \mathbb{R}^h$ is m-th token in H_k^i and M is the maximum length of history questions and answers.

3.2 Impression Feature Representation

Multiple NLP tasks obtained state-of-the-art results by using pre-trained language model BERT, which learned the deep bidirectional representations through transformer (Vaswani et al., 2017). Adaptive to this paper, the encoder of BERT model encodes the question q_k, the passage x and the proposed Impression Feature (ImpFeat) that attend the conversaitonal histories H_k into contextualized representation, which is shown in Figure 1. The input sequences composed of token-level questions q_k and passages x_n are fed into the BERT model. Then the BERT encoder generates the token-level contextualized representation based on the token embedding, segment embedding, position embedding and the proposed impression feature (the different color row in the orange dotted lines of Figure 1). Finally, based on the output representation, the answer span predictor calculate the probability of each token as the beginning and end of the answer. Among them, the proposed impression feature (red-cyan row in the orange dotted frame) generation is detailed in Figure 2.

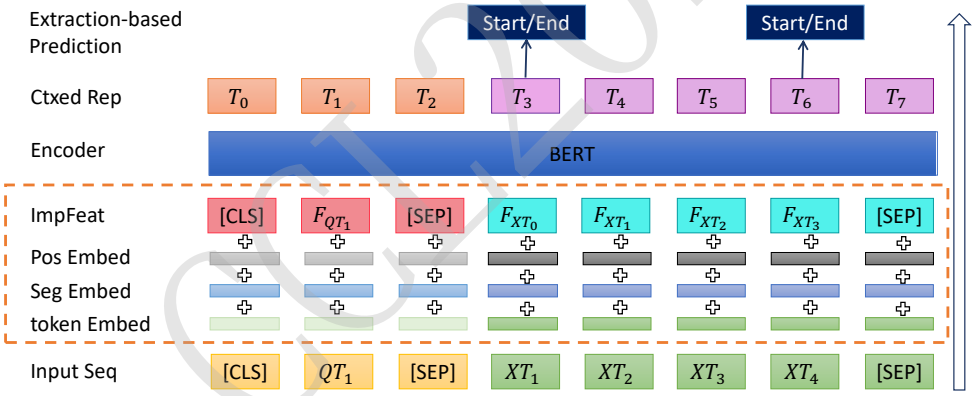

Figure 1: Our model with ImpFeat. It mainly reveals the process from the input of questions and passages (the light yellow-green row) to the contextualized representation (the pink-purple row), and then to the generation of answers (navy blue). This process includes the steps of inputting sequences, making features (marked by orange-dotted lines), BERT encoding, and predicting answers. The method of generating ImpFeat (red-cyan row in the right of Figure 2) from input sequence (the light yellow-green row in the left of Figure 2) is detailed in Figure 2.

As shown in Figure 2, the generation of impression features mainly includes two stages, word-level inter attention and turn inter attention. An input sequence contains a question q_k and a sub-passage x_n. For convenience, q_k is used as the representative of the input sequence in the following formula. The calculation method of the sub-passage x_n is the same as it. So the generation process is as follows.

Step 1: we follow word-level inter attention in the previous work (Zhu et al., 2018) to compute the attended vector from history turns of questions and answers to the input sequence. The relevance score matrix between j-th token of the current question and m-th history questions or answers is defined as Eq. 1:

$$r_j^i(m) = \tanh(U q_k(j)) D \tanh(U H_k^i(m)) \quad (1)$$

Proceedings of the 19th China National Conference on Computational Linguistics, pages 863-873,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

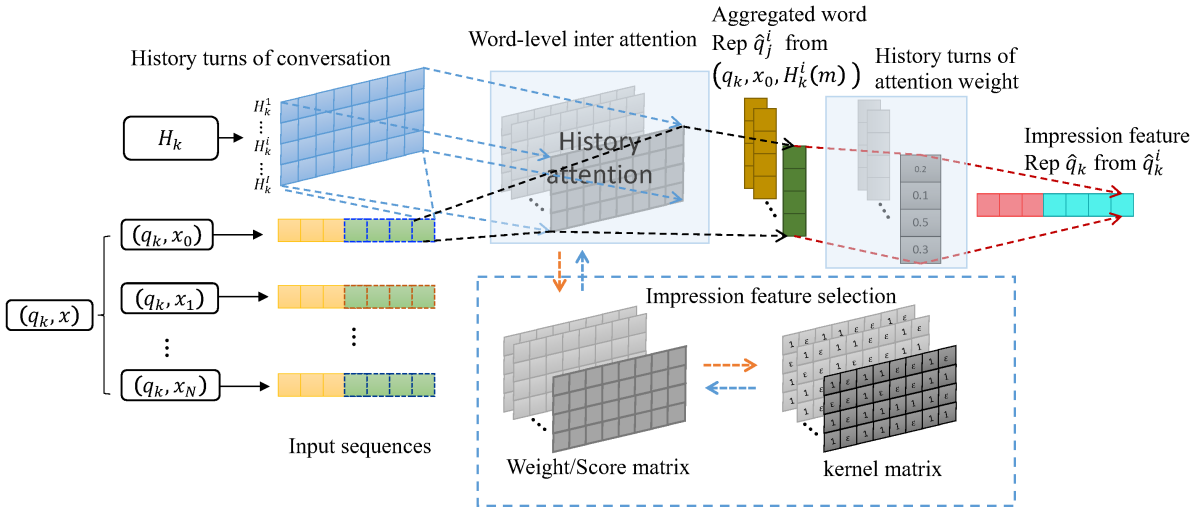

Figure 2: The proposed impression feature generation and selection using history attention. A sliding window approach is used to split a passage into sub-passages $(x_0, x_1, \cdots, x_N)$, which are then packed with the question q_k to form the input sequences $(q_k, x_0), (q_k, x_1), \cdots, (q_k, x_N)$. These input sequences share the same question. Then we generate the conversation history H_k of each input sequence. Take (q_k, x_0) for illustration, we did word-level inter attention and turn inter attention respectively. Word-level inter attention is applied to calculate attention $\hat{q}_k^i$ from each token of the conversational history to each token of the input sequence. Then turn inter attention is calculated from different history turns of conversation to the input sequence. In addition, we also make feature selection (in the blue dotted lines) for the obtained historical memory in word-level inter attention stage to make the memory is selective.

where, $r \in \mathbb{R}^{J \times I \times M}$, $D \in \mathbb{R}^{d \times d}$ is a diagonal matrix, and $U \in \mathbb{R}^{d \times h}$, d is the attention hidden size. The word-level attentive weight of m-th token in i-th history conversation to the j-th token of the current question q_k is represented as $\hat{\alpha}_j^i(m)$:

$$\hat{\alpha}_j^i(m) = \frac{e^{r_j^i(m)}}{\sum_{i'=1}^{I} \sum_{m'=1}^{M} e^{r_j^i(m)}} \tag{2}$$

Therefore, the aggregated word-level representation of all tokens in i-th history turn of conversation to the j-th token of the current question is represented as $\hat{q}_j^i$:

$$\hat{q}_j^i = \sum_{m=1}^{M} \hat{\alpha}_j^i(m) H_k^i(m) \tag{3}$$

Step 2: To learn the attention from different history turns of conversation to the input sequence, i.e. history turn inter attention, we learn an attention vector $D \in \mathbb{R}_I$ to compute attention weight from aggregated representation of i-th history turn of conversation to the current question. Initialize the weight matrix D with random values, then we get:

$$\hat{w}_i = \frac{e^{\hat{q}_j^i \cdot D}}{\sum_{i'=1}^{I} e^{\hat{q}_j^{i'} \cdot D}} \tag{4}$$

Further, the ImpFeat representation of all tokens of all history turns of conversation to the input question is denoted as $\hat{q}_k(j)$:

$$\hat{q}_k(j) = \sum_{i=1}^{I} \hat{w}_i \hat{q}_j^i \tag{5}$$

Proceedings of the 19th China National Conference on Computational Linguistics, pages 863-873,
Hainan, China, October 30 - November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Step 3: To learn the attention within the tokens of the input question and passage, self-attention in Transformer structure (Vaswani et al., 2017) is applied here. So $\hat{q}_k(j)$ is refered as impression feature representation, and is merged with the token embedding, segment embedding and position embedding as the input of BERT.

The proposed two attention methods, and the self-attention in Transformer (Vaswani et al., 2017) respectively learn the attention from the tokens of history conversation to the input sequence, the attention from history turns to the input sequences, and the attention within the input sequence. So the model learns the historical information from different dimensions. Just like human reading, the model has a deep impression on historical information, which is why we express the learned representation as the impression feature. In addition, we also make feature selection for the obtained historical memory in word-level inter attention stage to make the memory is selective.

3.3 Impression Feature Selection

In order to verify whether the attention learned above is effective, and remove some redundant information. In step1, we use a kernel matrix to disturb the weights learned by the input sequence and history turns of conversation. Make

$$r_j^i = \sum_{m=1}^{M} r_j^i(m) \tag{6}$$

Then we sort r_j^i for each token of input sequence, select the historical turn number corresponding to the top s of r_j^i as the selected useful turn, which is represented as $r_j^{s'}$, $0 \leq s' \leq I$, and generate the corresponding kernel matrix :

$$a = \{a_j^i(m)\}_{1\leq i\leq I, 1\leq m\leq M}, a_j^i(m) = \begin{cases} 1, & \text{if } i = s' \\ \epsilon, & \text{otherwise} \end{cases} \tag{7}$$

where, ϵ is equals to a very small value, it is 0.001 in this paper. s is from 3 to 5 in this paper. $a_j^{s'}(m) = 1$ for all m in the s'-th turn. The new weight matrix after selection is represented as:

$$\alpha_j^i(m) = \hat{\alpha}_j^i(m) \cdot a_j^i(m) \tag{8}$$

where, $\alpha_j^i(m)$ represents that which history turns of conversation are more useful to the input sequence. Then we use the new weight matrix $\alpha_j^i(m)$ to replace $\hat{\alpha}_j^i(m)$ in Eq.(3), the q_k after adding impression feature selection is represented as:

$$q_j^i = \sum_{i=1}^{I} \alpha_j^i(m) H_k^i(m) \tag{9}$$

At last, use Eq.(9) and Eq.(5) to recalculate the ImpFeat representation.

4 Experiments

4.1 Data Description

The QuAC (Choi et al., 2018) dataset mentioned in the introduction is used for our experiment. It is a large-scale dataset contained more than 8,850 conversations and 98,400 questions. Statistics for this dataset is summarizied in Table 1, we can only access the training and validation data.

4.2 Experimental Setup

4.2.1 Competing Methods

The methods with published papers on QuAC leaderboard[3] are considered as baselines. To be specific, the competing methods are:

[3]http://quac.ai/

Proceedings of the 19th China National Conference on Computational Linguistics, pages 863-873,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 1: Statistics of QuAC dataset.

Items	Training data	Validation data
Number of passages	6,843	1,000
Number of dialogs	11,567	1,000
Number of questions	83,568	7,354
Average questions per dialogs	7.2	7.4
Average tokens per passage	396.8	440.0
Average tokens per question	6.5	6.5
Average tokens per answer	15.1	12.3
Min/Avg/Med/Max history turns per question	0/3.4/3/11	0/3/5/3/11
% unanswerable	20.2	20.2

BiDAF++ (Choi et al., 2018; Peters et al., 2018): BiDAF++ is a re-implementation of a top-performing SQuAD model (Peters et al., 2018), which augments bidirectional attention flow (BiDAF) (Seo et al., 2016) with self-attention and contextualized embeddings.

BiDAF++ w/2-ctx (Choi et al., 2018): Based on BiDAF++, BiDAF++ w/r-ctx consider the context(ctx) from the previous r QA pairs. When $r = 2$, the model reached the best performance.

FlowQA (Huang et al., 2018): This model incorporate intermediate representations generated during the process of answering previous questions, thus it integrates the latent semantics of the conversation history more deeply than apporaches that just concatenate previous questions/answers as input.

BERT (Qu et al., 2019a): A ConvQA model with BERT is implemented and without any history modeling. We re-implement the model with batch size as 12 and marked with BERT_BZ12.

BERT + PHQA (Qu et al., 2019a): Based on BERT, this model adds conversation history by prepending history turn(s) to the current question. Here, PHQA prepends both history questions and answers. **BERT + PHA** prepends answers only.

BERT + HAE (Qu et al., 2019a): This approach model the conversation history by adding history answer embedding that denote whether a token is part of history answers or not.

BERT + PosHAE (Qu et al., 2019b): Based on BERT + HAE, This model learn position information of history turns by setting the distance from the historical turn to the current turn.

BERT + Att_PHQA : We implement a BERT-based ConvQA model that encode attention of history questions and answers (Att_PHQA), where, attention is computed from the prepended previous r QA pairs $(q_k, q_{k-1}, a_{k-1}, \cdots, q_1, a_1)$ to the input sequence (q_k, x_n). Here $r = 2$, i.e. $(q_k, q_{k-1}, a_{k-1}, q_{k-2}, a_{k-2})$.

BERT + Att_PHA: A BERT-based ConvQA model that encode attention of history answers only, where the prepended previous history is formed by $(q_k, a_{k-1}, a_{k-2}, \cdots, a_1)$. we set max answer length as 35 since it gives the best performance under this setting.

BERT + ImpFeat w/r-ctx: This is the solution we proposed in Section 3. The history turns of conversation H_k from the previous r QA pairs.

4.2.2 Hyper-parameter Settings and Implementation Details

In order to compare with methods similar to this article, such as BERT + HAE (Qu et al., 2019a), BERT + posHAE (Qu et al., 2019b), most of our experimental setting are the same as paper (Qu et al., 2019b), such as Tensorflow[4], v0.2 QuAC data, and BERT-Base Uncased model with the max sequence length of 384. The difference is that the batch size is set to 12, and the max answer length is set to 35 in BERT+ Att_PHA. The total training steps is set to 58000. Experiments are conducted on a single NVIDIA TESLA V100 GPU.

[4]https://www.tensorflow.org/

Proceedings of the 19th China National Conference on Computational Linguistics, pages 863-873,
Hainan, China, October 30 - November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4.2.3 Evaluation Metrics

The QuAC challenge provides two evaluation metrics, word-level F1 and human equivalence score (HEQ) (Choi et al., 2018). Word-level F1 evaluates the overlap between prediction and references. HEQ is used to check if the system's F1 matches or exceeds human F1. It has two variants: (1) the percentage of questions for which this is true (HEQ-Q), and (2) the percentage of dialogs for which this is true for every question in the dialog (HEQ-D).

4.3 Experimental Results and Analysis

4.3.1 Main Evaluation Results

Table 2: Evaluation results on QuAC. Validation result of BiDAF++, FlowQA are from (Choi et al., 2018) and (Huang et al., 2018) ."-" means a result is not available.

Models	F1	HEQ-Q	HEQ-D
BiDAF++	51.8	45.3	2.0
BiDAF++ w/2-ctx	60.6	55.7	5.3
FlowQA	**64.6**	-	-
BERT	54.4	48.9	2.9
BERT + PHQA	62.0	57.5	5.4
BERT + PHA	61.8	57.5	4.7
BERT + HAE	63.1	58.6	6.0
BERT + PosHAE	**64.7**	**60.7**	**6.0**
BERT_Batchsize12	53.26	46.15	2.6
BERT + Att_PHQA	54.3	47.45	2.2
BERT + Att_PHA	62.48	57.74	5.3
BERT + ImpFeat w/11-ctx	63.02	58.54	**6.2**
BERT + ImpFeat w/4-ctx	**63.67**	**59.17**	5.9

The results on the validation sets are reported in Table 2. To implement the method of this article, we re-implement the BERT-based question answering model on the QuAC dataset, and set the batch size as 12. The result is slightly smaller 1% than the result in paper (Qu et al., 2019a), which is caused by the different hyperparameters setting. Moreover, we summarize our observations of the results as follows: (1) BERT + Att_PHA brings a significant improvement compared with BERT + PHA. This shows the advantage of using attention and suggests that making attention from history answer to the current question and passage plays an important role in coversation history modeling. (2) Computing attention with PHQA and PHA are both effective. BERT + Att_PHA achieves a higher performance compared to BERT + Att_PHQA, which indicates that all history answers contribute more information to the model than just the previous two turns of conversation history. (3) Our model (BERT + ImpFeat) obtains a substantially significant improvements over the BERT + Att_PHA model, but suffer the poor performance than FlowQA and BERT + PosHAE. One possible reason is that the impression feature has learned the token relevance from the context history to the current and passage, but it seems that there is still lack of topic flow and positional information of the conversation history, so that there is not enough improvement. (4) BERT + ImpFeat w/4-ctx outperform BERT + ImpFeat w/11-ctx, which indicates that the number of history pairs still affect the performance of the model, but four turns of context history may not be optimal result since we have not yet do experiments for all different history turns.

4.3.2 Ablation Analysis

In order to verify whether the proposed impression feature selection method is effective, we set different selection methods for comparison. Specifically, we randomly set the element of a in Eq.(7) to 1 or ϵ, then predict the answer. The results in Table 3 shows that after removing or replacing our feature selection method, the model performance drops significantly, indicating the importance of our proposed selection method.

Table 3: Results for ablation analysis. “w/o” means to remove or replace the corresponding component.

Models	F1	HEQ-Q	HEQ-D
BERT + ImpFeat w/4-ctx	**63.67**	**59.17**	**5.9**
w/o ImpFeat Selection	62.06	57.49	5.5
w/o Random Selection	23.75	23.02	0.6

Table 4: Results for human evaluation. Correctness, Completeness, Fluency are abbreviated as Cor, Com and Flu.

Evaluator	Cor	Com	Flu
A	4.07	4.74	4.71
B	4.06	4.79	4.74
C	4.0	4.68	4.54
Average	**4.04**	**4.73**	**4.66**

4.3.3 Impression Feature Analysis

To further analyze the impression feature, we randomly select an example and visualize the relationship between current question, passage, and conversation history, as shown in Figures 3 and 4, respectively. In Figures 3, the passage is from “..., faced ratio for 1963, and subsequent years. On May 11, Koufax no-hit the San Francisco Giants 8-0, besting future Hall of Fame pitcher Juan Marichal–himself a no-hit pitcher a month later, ... ”. The current question is from “Are there any other interesting aspects about this article? ”, and the sixth turn of history answer is parts of the passage. We can see that the tokens that are more relevant to the passage have a higher score and the stronger correlation, their corresponding color are redder, even white. On the contrary, the tokens that are less relevant to the passage have a lower score and the worse correlation, their corresponding color are darker. Furthermore, we can clearly see that there is a diagonal score that is generally large, because its answer exactly corresponds to the original answer. Besides, from Figure 4, we can see that the tokens such as ”powerful”, ”graints” in history answers are more relevant to the tokens ”change”, ”walks”, ”affect” and ”basketball” in the current question, indicating that the impression feature has learned relevant information from conversation history, and it is helpful to predict answers.

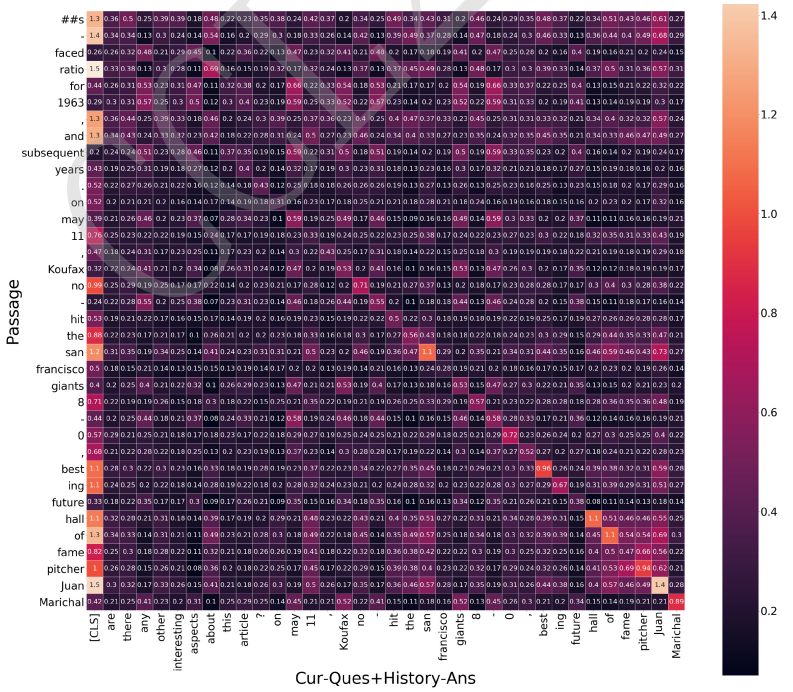

Figure 3: The heatmap of attention score from the current question and conversation history (Cur-Ques + History-Ans) to the passage. The first cloumn is the aggregated scores, the second to ninth tokens on the horizontal axis indicate the ninth current question, and the remaining tokens represent a part of the answer of the sixth turn conversation history. The vertical axis represents parts of passage tokens.

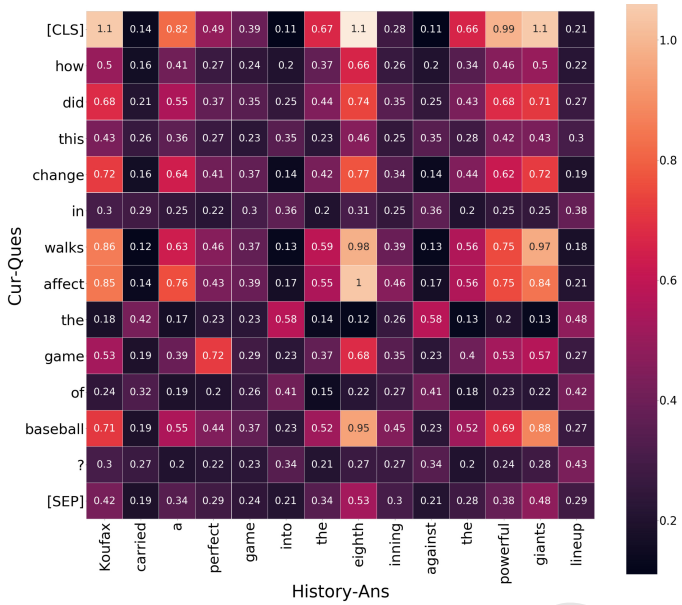

Figure 4: The heatmap of attention score from the conversational history answer (History-Ans) to the current question (Cur-Ques). The first row is the aggregated scores.

4.4 Human Evaluation

In addition, human evaluation is also conducted to verify the reliability of the proposed method. Three graduate students evaluate 100 randomly selected samples from the validation set results. Each sample contains one article and multiple QA pairs.

With reference to the subjective evaluation metrics commonly used in question generation research, we design correctness, completeness, and fluency to evaluate the predicted results. Correctness refers to the correctness of a predicted answer, evaluating whether a predicted answer is the same or related to the original answer, and whether it can be used to answer the question, etc. Completeness refers to the completeness of semantics, evaluating whether a predicted answer has the main components of the sentence, whether it is a complete sentence that is understandable to humans, and whether there are redundant words or missing words, etc. Fluency refers to the fluency of expression, evaluating whether a predicted answer is smooth, and whether the word order is correct, etc.

We divide the score into 1-5 based on three metrics. From 1 to 5, the predicted answer becomes more accurate, complete and fluent. Specifically, 1 means the predicted answer is completely incorrect, incomplete, or not fluent. And 5 means the answer is correct, complete, and fluent. Finally, the average score is calculated and shown in Table 4. The correctness, completeness, and fluency all exceed 4 points, indicating that most predicted answers are reasonable.

5 Conclusion and Future Work

Based on the general framework for ConvQA, we propose a new feature named impression feature, and combine the proposed feature with token embedding, position embedding and segment embedding as the input of BERT encoder. Then we introduce an impression feature selection method to select the important history information. Extensive experiments show the effectiveness of our method. Finally, we perform an in-depth analysis to show the different attention methods under different setting. Future work will consider to integrate multi-oriented information and a free-form answer type for ConvQA.

Acknowledgments

We thank all people who did human evaluation. This work are funded by China Postdoctoral Science Foundation (No.2019M660578), National Key Research and Development Program (No.2019YFB1406302), and Beijing Postdoctoral Research Foundation (No.ZZ2019-93).

References

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

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*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. *CoRR*, abs/1810.06683.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Refnet: A reference-aware network for background based conversation. *arXiv preprint arXiv:1908.06449*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Learning to deceive with attention-based explanations. *The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, July.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. Bert with history answer embedding for conversational question answering. *SIGIR'19: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information*, page 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. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1391–1400, Nov.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

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, Mar.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Mark Yatskar. 2018. A qualitative comparison of coqa, squad 2.0 and quac. *arXiv preprint arXiv:1809.10735*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *CoRR*, abs/1812.03593.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 863-873,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China