Graph-combined Coreference Resolution Methods on Conversational Machine Reading Comprehension with Pre-trained Language Model

Zhaodong Wang **Kazunori Komatani**
SANKEN, Osaka University

Abstract

Coreference resolution such as for anaphora has been an essential challenge that is commonly found in conversational machine reading comprehension (CMRC). This task aims to determine the referential entity to which a pronoun refers on the basis of contextual information. Existing approaches based on pre-trained language models (PLMs) mainly rely on an end-to-end method, which still has limitations in clarifying referential dependency. In this study, a novel graph-based approach is proposed to integrate the coreference of given text into graph structures (called coreference graphs), which can pinpoint a pronoun's referential entity. We propose two graph-combined methods, evidence-enhanced and the fusion model, for CMRC to integrate coreference graphs from different levels of the PLM architecture. Evidence-enhanced refers to textual level methods that include an evidence generator (for generating new text to elaborate a pronoun) and enhanced question (for rewriting a pronoun in a question) as PLM input. The fusion model is a structural level method that combines the PLM with a graph neural network. We evaluated these approaches on a CoQA pronoun-containing dataset and the whole CoQA dataset. The result showed that our methods can outperform baseline PLM methods with BERT and RoBERTa.

1 Introduction

In recent years, using a large-scale pre-trained language model (PLM) as a backbone for various challenging machine comprehension tasks (Devlin et al., 2019) has become fundamental, especially in conversational machine reading comprehension (CMRC) (Liu et al., 2019a). CMRC tasks not only require a model to fully understand the given articles but also propose to mimic the way humans seek information in conversations through question-answering. Most PLM utilize attention mechanism and achieve positive results on a broad range of CMRC datasets (Choi et al., 2018; Reddy et al., 2019). PLMs generally use an end-to-end approach trained from questions to answers. However, the explainability of the answers generated through the intrinsic multi-head self-attention mechanism remains insufficient. Although these PLMs have demonstrated great advantages in terms of solving questions that simply need semantic matching, limitations in logical comprehension (Ding et al., 2019) such as in coreference resolution still exist.

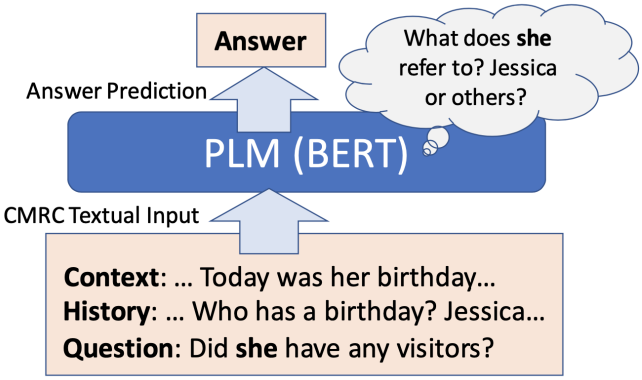

Figure 1: Coreference resolution is required for end-to-end PLM in CMRC task.

Coreference resolution such as for anaphora (von Heusinger and Egli, 2012) is commonly found in CMRC tasks. Anaphora can be described as a pronoun word (anaphor) contained in a current question, in which its referential entity (antecedent) has already been introduced earlier in the conversation history or article context. As shown in Figure 1, to answer the current question "*Did she have any visitors?*", the model requires that the pronoun "*she*" be resolved as an anaphor referring to the entity "*Jessica*" as its antecedent, on the basis of the given context and conversation history. Therefore, CMRC models require mechanisms that can resolve referential dependencies to properly understand the intent of current questions.

Considering the shortcomings of the PLM approach in logical comprehension such as in coref-

Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 72 - 82

erence resolution, research on how to better adapt models to learn reasoning is gradually gaining attention (Yeh and Chen, 2019; Qu et al., 2019; Song et al., 2018). FlowQA (Huang et al., 2019) was proposed to add a reasoning layer between questions and answers to incorporate intermediate representations of a conversation history. The question rewriting (QR) model (Vakulenko et al., 2021; Lin et al., 2020) was proposed to rewrite current questions on the basis of a conversation history. Specifically, the QR model simplifies complex multi-turn question-answering (QA) tasks into single-turn QA tasks, which can solve a current question without a conversation history.

However, because these models are built through the embeddings of a conversation history (Qu et al., 2019), they generally suffer from two drawbacks in coreference reasoning for CMRC tasks. (1) Since the input length of a conversation history is limited by the PLM's structure, the current question sometimes contains pronouns whose referential entity does not appear in the conversation history, so the model cannot accordingly resolve referential dependencies. (2) To achieve coreference reasoning, a CMRC model also needs to seek information from the context of articles. Due to the sequence nature of the PLM and the multiple referential dependencies in the context of an article, these models cannot handle each referential dependency precisely, as shown in Figure 2's context part in different colors.

In this paper, we propose solving the coreference of a target pronoun through additional mined information to enhance PLMs' coreference reasoning ability for CMRC. A novel graph approach is proposed that integrates the coreferences of given text into graph structures, which we call the coreference graph. The coreference graph is constructed separately by using the conversation history and article context as text information. Each entity in the graph holds a unique place label in accordance with the text information, which can be used to pinpoint every pronoun's referential dependency precisely. To better implement the coreference graph as an enhanced component into PLMs, we propose two graph-combined methods: the evidence-enhanced method and the fusion model method. These two methods integrate graph information from the textual and structural levels of the PLM architecture, respectively.

The **evidence-enhanced** method involves two textual level methods that enrich the PLM's input information for coreference reasoning: an **evidence** generator (EG) generates new text to elaborate pronouns, and an **enhanced** question (EQ) rewrites a pronoun into a referential entity in a question.

The **fusion model** is a structural level method that combines the PLM with a graph neural network. This model treats the PLM as an encoder to extract sequence features of pronouns and referential words from input. After that, the graph features of the corresponding words are computed by graph neural networks on the basis of the connectivity of the coreference graph. These two features are integrated using learnable weights to enhance the PLM's coreference reasoning ability.

For the experiments, we used questions from CoQA (Reddy et al., 2019) that contained pronouns to compose a new dataset (pronoun-containing dataset) specialized for the coreference reasoning ability of the CMRC model. We evaluated various combinations of our proposed methods on different PLMs, and we also compared them with the existing QR approach. The results showed that our methods can greatly outperform in terms of F1 score on the CoQA pronoun-containing dataset, 2.6 on BERT (Devlin et al., 2019) and 0.7 on RoBERTa (Liu et al., 2019b). We also used the whole CoQA dataset to evaluate the fusion model, which achieved the best performance in our methods, to compare its overall performance with RoBERTa. The contributions of this paper are as follows.

- We propose a novel graph approach for coreference resolution. This approach can establish referential dependency that appears not only in a conversation history but also in an article context.

- We show that both our evidence-enhanced and fusion model methods boost the performance of different PLMs in CMRC coreference resolution. Therefore, we prove that the introduction of additional information can further leverage the performance of PLMs in complex reasoning such as in coreference resolution.

- Our approaches provide a precise reasoning route for CMRC's coreference resolution and overcome the PLM model's weakness of interpretability.

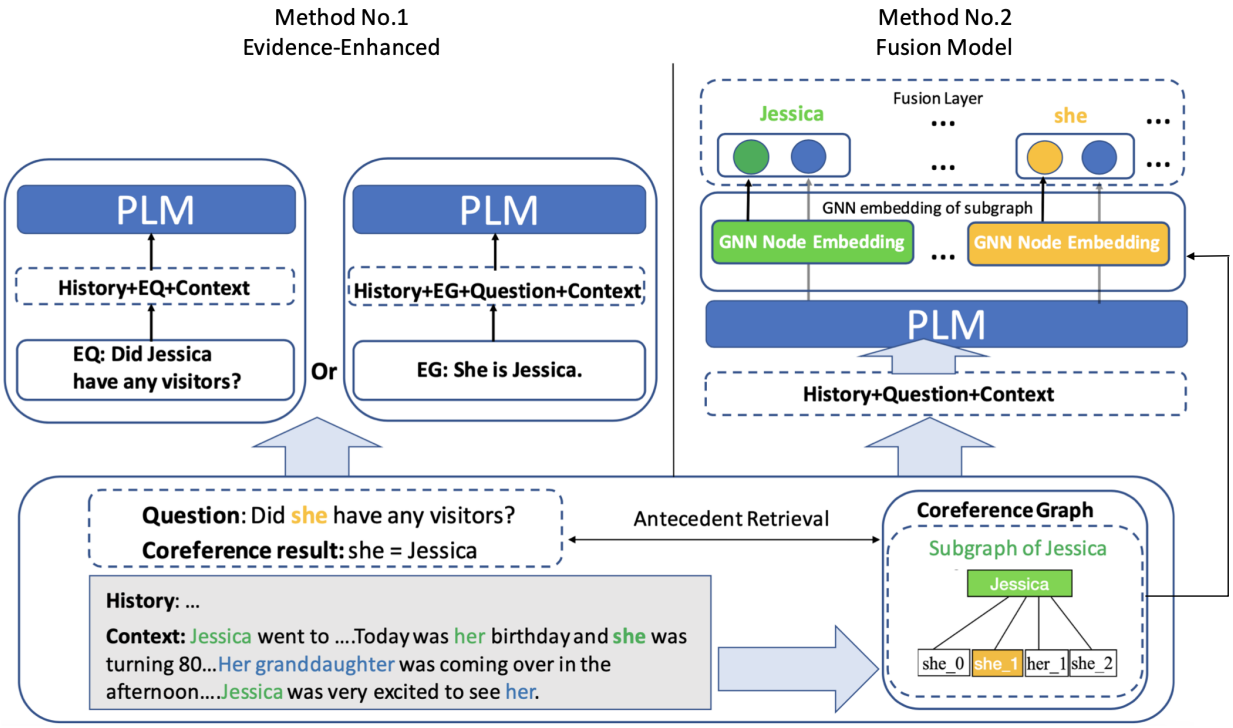

Figure 2: Overview of evidence-enhanced and fusion model. To answer current question, model should determine pronoun's referential entity through context or conversation history; graph-based coreference resolution can precisely determine dependency and add additional information to current question. Left part denotes textual level method of evidence-enhanced method. Right part denotes fusion model and fusion of PLM and graph embedding.

2 Background

2.1 Pre-trained Language Model

In recent years, the emergence of pre-trained language models (PLMs), including BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), XL-NET (Yang et al., 2019), and RoBERTa (Liu et al., 2019b), has refreshed the performance of various NLP tasks with advanced comprehension abilities. BERT is a representative model that is based on a multi-layer transformer (Vaswani et al., 2017). It is trained by using a massive amount of text data through a masked language model and next sentence prediction. There have been several improvements to the BERT model (Qiu et al., 2020), such as ConvBERT (Jiang et al., 2020), which specifically improves its performance in MRC. These PLM-based models mostly increase the scale of model parameters or improve the attention mechanism through their structure, but they still lack reasoning-level analysis and evidence support due to them using end-to-end learning methods (Chen and Yih, 2020).

2.2 Coreference Resolution

Coreference resolution is the task of retrieving all references in text that refer to the same entity. With the development of deep learning, the neural network has been gradually used to solve coreferencing, such as CoNLL-2012 (Pradhan et al., 2012), in recent years (Xu and Choi, 2020; Kirstain et al., 2021). Lee et al. (Lee et al., 2017) first applied the LSTM (Sak et al., 2014) network to coreference resolution; it can extract referential dependencies directly from text. Joshi (Joshi et al., 2019) provided a PLM baseline for coreference resolution through BERT. Joshi also provided SpanBERT Joshi et al. (2020), which enhanced the PLM's performance, especially in coreference extraction.

In this paper, we use AllenNLP Gardner et al. (2018)'s framework as an implementation of the approach by Lee et al.(Lee et al., 2017) with spanBERT for textual word embedding, and we achieve high-precision coreference extraction from a conversation history and article context.

2.3 Machine Reading Comprehension

Current machine reading comprehension (MRC) tasks can be classified into single-turn and multi-turn types, depending on whether the question-

answering relies on the conversation history. To tackle single-turn MRC such as SQuAD (Rajpurkar et al., 2018), many models based on semantic matching have been proposed, such as BiDAF (Seo et al., 2017), DrQA (Chen et al., 2017), (Lin et al., 2018), QANet (Yu et al., 2018), and BERT (Devlin et al., 2019), for MRC.

However, for multi-turn MRC like CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018), conversation-based questions and answers are introduced to enhance the connection between questions (known as CMRC (Liu et al., 2019a)). The ambiguity of a question increases (Min et al., 2020) due to the addition of a conversation history. Thus, to predict the answer $\hat{A}_i$ for the current question Q_i, the model should not only have to comprehend the article context C but also the conversation history H_i from the beginning (Q_1, A_1) to the previous turn (Q_{i-1}, A_{i-1}) for integration.

$$H_i = \{Q_1, A_1, ..., Q_{i-1}, A_{i-1}\} \quad (1)$$

$$\hat{A}_i = argmax(P(A_i|Q_i, C, H_i)) \quad (2)$$

For multi-turn MRC, several works (Huang et al., 2019; Yeh and Chen, 2019; Qu et al., 2019; Song et al., 2018) have incorporated reasoning representation to capture a conversation history's embedding. In comparison, approaches like question rewriting (QR) (Papakonstantinou and Vassalos, 1999) aim to break down multi-turn MRC into single-turn subtasks to minimize the complexity of multi-turn MRC (Vakulenko et al., 2021). CANARD (Elgohary et al., 2019) rewrites QuAC's questions and introduces this rewriting to the QR task. QR models (Vakulenko et al., 2021; Lin et al., 2020) rewrite current questions to incorporate a conversation history. However, due to the variable length of a conversation history, such models still have limitations in precisely resolving the coreference in questions.

3 Propsed Methods

In this section, we describe the architecture of our methods as an enhanced PLM component, as illustrated in Figure 2. The model contains two stages. (1) We construct a coreference graph from textural information towards solving the pronoun's referential entity in a question. (2) We use our two methods, evidence-enhanced and the fusion model, to integrate a referential entity's information into PLMs using textual and structural levels, respectively.

3.1 Coreference Graph

Inspired by the previous works (Song et al., 2018; Bastings et al., 2017; Dhingra et al., 2018), we introduce graph structures for the anaphora in questions. Specifically, our method uses the approach by Lee et al. (Lee et al., 2017) with SpanBERT word embedding to precisely extract all coreferences in text and organize them into graph structures. Additionally, we propose modeling the conversation history and article context separately in structures to fully use the graph information.

In the article context part, because there may be multi-identical pronouns referring to different entities in a context (e.g., "he" could refer to two males in the same article context), the current sentence number (order number) is kept after entities to ensure their uniqueness. As shown in Figure 3 with different numbers. To organize the entities into a graph, all of the anaphors (pronouns) are connected to the initially-occurring antecedent (referential entity). In this way, the entire context can be processed into a graph with multiple clusters, and each cluster holds a unique referential entity, as illustrated in Figure 3 in different colors.

In the conversation history part, to avoid multi-identical pronouns, the Q_i label for the i-th question and A_i label for the i-th answer are added behind an entity in a conversation history. In the construction part, considering the time-sequence nature of a conversation history, we use a conversation history's order sequences $(Q_1, A_1, Q_2, A_2, ...)$ to connect these entities into a queue structure.

3.1.1 Coreference Graph Construction

As illustrated in Figure 3, this procedure can extract the coreference information from text into a coreference graph. First, we extract reference words with relevant number labels as referential entities. In this way, each reference word can be classified into various clusters (shown in different colors in the top half of Figure 3). In the graph construction of the article context part, we use the first referential entity in one cluster and the initially-occurring antecedent as the head node. We connect all the remaining referential entities in the cluster to the head node. For the conversation history part, we connect the referential entities in the cluster in a queue in the order sequence $(Q_1, A_1, Q_2, A_2, ...)$. Accordingly, this step is repeated for every cluster

until each reference word has been processed into a graph structure as a unique entity.

3.1.2 Antecedent Retrieval

Antecedent retrieval is a process of querying the referential entity of a target pronoun through a coreference graph. For retrieval from an article context, the target pronoun and the sentence's order index are considered to form a query entity. When the node of the query entity is found, it is used as the starting node for a graph search until a non-pronoun entity is found as a referential entity for the result.

3.2 Method No.1: Evidence-Enhanced

We learned from previous studies (Zhou et al., 2019; Ding et al., 2019) that additional evidence is essential for a PLM's logical comprehension. Therefore, we present textual reformulation methods for resolving the referential dependency of current pronouns. As shown in Figure 2, after retrieving the referential entity ("she" refers to "Jessica"), the model needs to obtain this information for the current question Q_i. In PLMs like BERT (Devlin et al., 2019), the CMRC typically defines the model's input as the concatenation of three segments. Specifically, given a context C, the input for BERT is "**[CLS]**H_i**[SEP]**Qi**[SEP]**C." To ensure that new information is introduced with as little impact as possible for the PLM input, we propose two textual-level methods:

- **Evidence Generator (EG):** Generating inferential sentences to solve coreference on the basis of textual rules (like *"She" is "Jessica"*) and then adding the inferential sentence as evidence before the question. The input structure is "**[CLS]**H_i**[SEP]**EG_{Q_i}**[SEP]**Q_i**[SEP]**C."
- **Enhanced Question (EQ):** Reformulating a question by replacing the pronouns in the question with referential entities to create an enhanced question and replacing the enhanced question with the original one as input. The input structure is "**[CLS]**H_i**[SEP]**EQ_{Qi}**[SEP]**C."

3.3 Method No.2: Fusion Model

Inspired by Qiu et al. Qiu et al. (2019), we propose using the graph neural network to extract a coreference graph's features. We fuse these graph features with sequence features from the PLM to enhance the PLM's coreference reasoning ability.

3.3.1 Embeding Fusing

We want the model to learn both the graph and sequence features of an entity during computation. Additionally, we hope that the model can balance the two kinds of features by using learnable weights. Therefore, the final embedding $FinalEmb_k$ of all entities k in a coreference graph is calculated as follows ($[A:B]$ means to concatenate the two vectors A and B in a row, and $\odot$ means the Hadamard product).

$$w_k = ReLU(W \times [PLM_k : GNN_k]) \quad (3)$$

$$FinalEmb_k = w_k \odot PLM_k + (1 - w_k) \odot GNN_k \quad (4)$$

The computed final embedding is passed through the fully connection layer to compute the answer prediction for the current question.

4 Experiment Setup

4.1 Datasets Description

CoQA (Reddy et al., 2019) consists of 127K questions and answers from documents in 5 domains (Children, Literature, Middle& High School English Exams, CNN News, Wikipedia). The question-answering can be divided into extractive and non-extractive types (Niu et al., 2020). Similar to SQuAD, the extractive type selects a span from the context for the final answer to the question. The non-extractive type is defined as choices from Yes/No/Unknown for answering. We used two datasets to perform this experiment:

- **CoQA all:** The complete CoQA dataset.
- **CoQA pronoun-containing (38% of CoQA all):** Used to evaluate the model's performance in coreference resolution for anaphora. Samples in which questions contained pronouns from CoQA were extracted to form a partial dataset.

Compared with the evidence-enhanced method, the fusion model does not need the input of the model to be changed for learning. Therefore, we additionally used the CoQA-all dataset to evaluate the overall performance of the model.

All evaluations were conducted using the overall F1 score by using CoQA's official evaluation script[1].

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

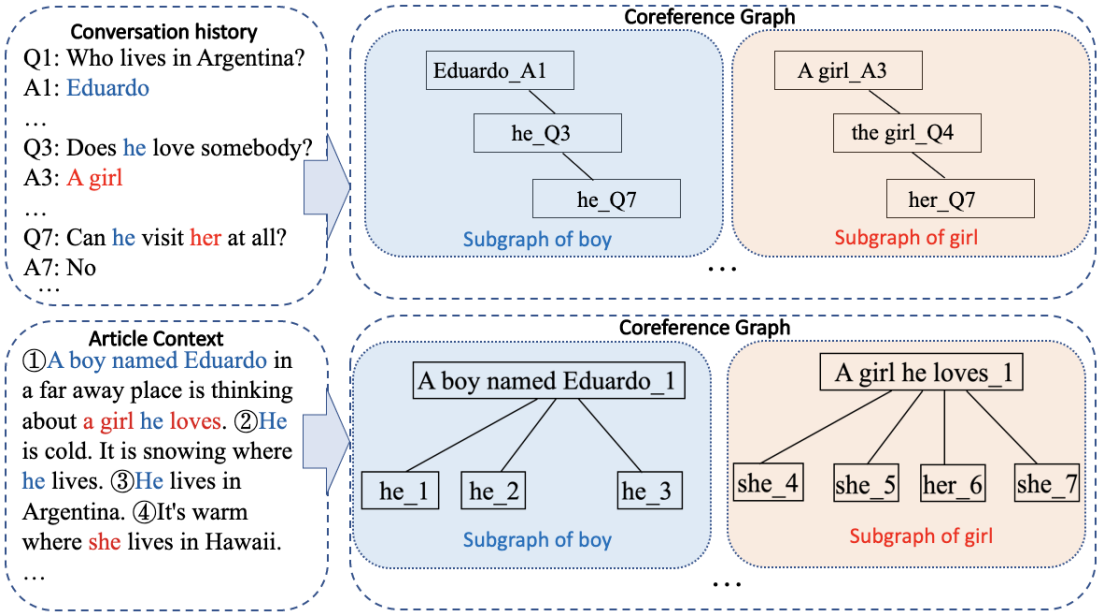

Figure 3: An example of converting conversation history and article context into the coreference graph. The same color represent entities has same referring entity, also in the same cluster as graph.

4.2 SpanBERT-based Coreference Extraction

We applied the coreference resolution model from AllenNLP[2]. This model adopts Lee et al. (Lee et al., 2017)'s approach to extracting the coreferences in clusters. Rather than using GloVe's word embedding in the initial model, SpanBERT (Joshi et al., 2020) for word embedding was used due to its superiority on the task of extraction.

4.3 Baseline of PLMs

4.3.1 BERT

Due to the multi-turn characteristic that CMRC retains compared with MRC tasks, the conversation history before Q_i should be considered as input into the model. In this experiment, a BERT-base-uncased (Devlin et al., 2019) fine-tuned by using all CoQA was used as our baseline model. It takes a concatenation of three segments as input (length of conversation history is 2). Specifically, given a context C, the input for BERT is [CLS]$(Q_{i-2}, A_{i-2}), (Q_{i-1}, A_{i-1})$[SEP]$Qi$[SEP]$C$, in which "[CLS]" is a classifier for "Yes/No/Unknown/Span" for CoQA's non-extractive questions.

4.3.2 RoBERTa

On the basis of BERT model's architecture, RoBERTa (Liu et al., 2019b) removes next sentence prediction and possesses better robustness through modifications and pre-training with larger data. RoBERTa can exceed almost all performances compared with the BERT model. In the experiment, we adopted a RoBERTa-base-uncased with the same training configuration as BERT. We found that RoBERTa achieves remarkable scores on the CoQA pronoun-containing dataset, which means that the capability RoBERTa holds towards coreference resolution is comparably higher than BERT accordingly.

4.4 GNN Embedding Algorithms

4.4.1 Graph Attention Networks

The graph attention network (GAT) (Velickovic et al., 2017) learns the structural features of graphs from the spatial domain through a multi-headed attention mechanism. In this experiment, we used PyTorch Geometric[3] as the implementation of GAT graph embedding, and the number of multi-heads was set to 8.

4.4.2 Graph Convolutional Network

The graph convolutional network (GCN) (Kipf and Welling, 2017) learns the structural features of graphs from convolution layers. It can be used to study the properties of a graph from the eigenvalues and eigenvectors of a Laplacian matrix. GCN has been successful in processing graph data by extracting structure-aware features. In this experiment, we

[2]https://demo.allennlp.org/coreference-resolution

[3]https://pytorch-geometric.readthedocs.io/en/latest/

CoQA pronoun-containing dataset							
Model	Approach	Child.	Liter.	M&H.	News	Wiki	Overall
BERT-base	baseline	76.0	70.0	72.9	73.8	77.7	73.9
	+QR	60.7	66.1	69.0	70.2	73.6	69.7
	+Hist.+EG	75.2	72.0	75.4	76.2	**81.1**	75.7
	+Hist.+EQ	76.1	74.0	**76.2**	**76.9**	80.0	**76.5**
	+Cont.+EG	**77.8**	72.6	**76.2**	76.1	80.3	76.4
	+Cont.+EQ	75.8	72.7	74.2	75.3	80.7	75.5
	+Hist.&Cont.+EG	77.1	**74.8**	75.0	76.0	81.0	**76.5**
	+Hist.&Cont.+EQ	76.8	72.4	75.2	75.3	79.2	75.6
RoBERTa-base	baseline	82.5	80.0	81.6	83.1	84.1	82.1
	+Hist.+EQ	80.9	77.9	81.2	80.5	84.7	80.8
	+Hist.&Cont.+EG	82.4	**80.4**	81.1	83.4	84.4	82.2
	+Hist.&Cont.+GCN	**83.5**	**80.4**	81.5	**83.7**	**85.9**	**82.8**
	+Hist.&Cont.+GAT	83.0	79.7	**82.6**	82.9	85.4	82.6
CoQA all							
RoBERTa-base	baseline	81.1	79.3	80.4	82.8	83.9	81.5
	+Hist.&Cont.+GCN	**82.3**	**80.0**	80.4	**84.2**	**84.6**	**82.3**
	+Hist.&Cont.+GAT	81.0	79.2	**80.7**	82.9	84.5	81.7

Table 1: Comparison of baseline method with QR model, evidence-enhanced method and fusion model for CoQA. "EG" and "EQ" denote evidence generator and enhanced question, respectively. For coreference graph in antecedent retrieval, "Hist." denotes using conversation history part, "Cont." denotes using article content part, "Hist.&Cont." denotes using both. "GCN" and "GAT" denote fusion model using graph embedding algorithms of GCN and GAT, respectively.

used PyTorch Geometric as the implementation of GCN graph embedding.

4.4.3 Initialization

For all nodes contained in the coreference graph, we initialize the node features using embeddings at the token level E_i generated through PLM. Here, we compute the average value for each node feature F_i for initialization. e.g.. the node "the girl" is composed of two tokens, "the" and "girl," and node feature $F_{the:girl}$ for initialization can be calculated as follows.

$$F_{the:girl} = \frac{1}{2}(E_{the} + E_{girl}) \tag{5}$$

4.5 Details

All experiments were implemented on PyTorch [4]. BERT and RoBERTa were implemented by using the Huggingface Transformers library [5]. The approach by Lee et al. (Lee et al., 2017) was implemented through the pre-trained model "coref-spanbert-large" from AllenNLP. We used three 11-GB GPUs (GTX 1080Ti), a batch size of 24 for BERT, and a batch size of 10 for RoBERTa in all experiments.

BERT and RoBERTa were utilized as our baseline, represent the basic and advanced PLMs respectively. To compare our approaches with others, we applied Question Rewriting (QR) model (Lin et al., 2020) using T5 (Raffel et al., 2020), trained on CANARD (Elgohary et al., 2019). To identify the effectiveness of coreference graph, we proposed to use information from different parts of coreference graph as comparisons.

5 Results & Analysis

The results are shown in Table 1, which presents a performance comparison of the baseline approaches, end-to-end QR, and our proposed methods integrated with different parts of the coreference graph. We can see that compared with the baselines, both the evidence-enhanced method and fusion model method improved the model's performance in different categories (Child., Liter., M&H., News, Wiki, and Overall).

Specifically, the combination of the EG with the coreference graph (Hist.& Cont.) improved the overall F1 score by 2.6 on the BERT baseline and

[4]https://pytorch.org/

[5]https://huggingface.co/docs/transformers/index

BERT-base				
Baseline	Our EG	F1>0	F1 ≥0.5	F1=1
False	False	342	503	997
True	False	145	175	180
False	True	219	263	244
True	True	2356	2121	1641
Total: 3062				

Table 2: Analysis of results for all answers for CoQA pronoun-containing test dataset (3062 samples in total). Comparison of baseline BERT with our best EG method "BERT+ EG + Hist.&Cont." "True" and "False" indicate whether each answer produced by QA model was correct or incorrect, respectively, in accordance with F1 thresholds provided in right-side columns.

by 0.1 on the RoBERTa baseline. Therefore, we concluded that while dealing with anaphora's coreference resolution, both the EG and EQ were effective as enhanced components of the PLM baseline model with BERT.

Comparing EG and EQ approaches comprehensively for BERT and RoBERTa, the EG one had generally higher scores. One possible reason is that generating additional evidence behind a question as input maintains the integrity of the original question. Although the EQ approach also achieved relevant performance, the textual substitution of pronouns may alter the intention of the question and mislead the model to make erroneous answer predictions.

To measure the effectiveness of the evidence-enhanced approach for each question, we compared the F1 scores of the answers produced by the baseline (BERT) and our evidence-enhanced model with the best performance ("EG+hist.&cont.," as shown in Table 2). "True" and "False" indicate whether the answer predicted by the model was correct or incorrect, in accordance with the F1 thresholds provided in the right-side columns. As shown in Table 2, the second row reflects the case where our model got an erroneous answer when the baseline's answer was correct, which can be interpreted as getting an erroneous referential entity of the target pronoun, thus leading to an erroneous prediction. The third row indicates that the answer of our model was correct and that of the baseline's was wrong. Compared with the second row, the third row shows the effectiveness of our model: introducing the correct referential entity and enhancing the model to output the correct answer. Additionally, in the third row, with the rise of the F1 threshold, the number increased from F1 $>$ 0 to F1 $\geq$ 0.5, which means that our model slightly corrected the baseline's answer from completely wrong into closer to correct. However, from the decline from F1 $\geq$ 0.5 to F1 $=$ 1, we can infer that our model still has limitations in making fully correct answer predictions.

From the results for the fusion model, we found that the fusion model achieved a further improvement (by up to 0.7 on RoBERTa) compared with the baseline and evidence-enhanced methods. This model also showed improvement on the CoQA-all dataset, which contains samples that are not needed for coreference resolution (without pronouns in questions), compared with the baseline. This indicates that the fusion model can effectively use coreference graph information. It can solve coreference resolution and maintain the ability to solve no-coreference questions. Therefore, compared with the evidence-enhanced approach, the fusion model has higher robustness.

Through comparing the two different graph embedding methods, GAT and GCN, we found that GCN generally outperformed GAT in terms of score in each category. We assume one reason is that the processed graphs always hold the same structure (a vertex containing multiple one-hop neighbor nodes), and such a simple structure is not adequately learned by GAT's multi-head attention, which is suitable for capturing features from the spatial domain. In contrast, GCN captures the graph features of each neighbor by using convolution layers, so it performed better in this experiment.

6 Case Study

We investigated how our approaches improve the coreference reasoning ability of the RoBERTa baseline approach. To compare the differences in answer prediction, we used RoBERTa-base as the baseline. RoBERTa-base + Hist.&Cont. + EG had the best performance in Table 1 as the evidence generator (EG), and RoBERTa-base + FusionMd.(+GCN) had the best performance as the fusion model. We selected several specific cases from CoQA for elaboration.

An example is shown in Figure 4. In this example, the coreference graph resolves that "he" refers to "Joseph Aloisius Ratzinger." Because of the absence of coreference resolution, the baseline incorrectly predicted the answer at the wrong place.

#Example
Article context C:
...Ratzinger established himself as a highly regarded university theologian by the late 1950s and was appointed a full professor in 1958...
Conversation History H_i:
...
Q_{i-1}: Did he have a lot of experience as a pastor?
A_{i-1}: No.
Current Question Q_i: What was his occupation immediately preceding his papacy?
Resolution in coreference graph:
his = Joseph Aloisius Ratzinger
Answer prediction:
Fusion model: Theologian.
Evidence-Enhanced: Academic and professor of theology.
Baseline: A major figure on the Vatican stage.
Gold answer: Theologian.

Figure 4: Answer predictions from different CMRC models.

EG resolved the referential dependencies, so the prediction's meaning was close to the correct answer. However, the fusion model could integrate the coreference information and predict the answer span accurately.

7 Conclusion

In this paper, we proposed the coreference graph, which can integrate coreferences from text into a graph structure. To use the information retrieved from a coreference graph, we introduced the evidence-enhanced method, which comprises two textual-level coreference resolution approaches to leverage BERT's performance on CMRC. However, the results showed that the improvement for RoBERTa is still limited. Therefore, we proposed the fusion model, using graph neural networks to incorporate the coreference graph into PLM structure. In comparison with the baseline and evidence-enhanced methods, the fusion model showed further improvement on RoBERTa, maintaining relatively higher robustness when learning coreference resolution. We confirmed that in conversational reading comprehension, a graph-structured representation of the article context and conversational history can both be an information supplement for answering a current question, especially with different PLMs. Rather than the end-to-end method in PLMs, our approaches can generate readable text as evidence when answering a question, which strengthens the interpretability of PLMs.

References

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1957–1967. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37, Online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2018. Neural models for reasoning over multiple mentions using coreference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 42–48. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2694–2703, Florence, Italy. Association for Computational Linguistics.

Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5917–5923. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2019. Flowqa: Grasping flow in history for conversational machine comprehension. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zihang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Convbert: Improving bert with span-based dynamic convolution. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5802–5807. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 14–19. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 188–197. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *CoRR*, abs/2004.01909.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019a. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5783–5797. Association for Computational Linguistics.

Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, and Minlie Huang. 2020. A self-training method for machine reading comprehension with soft evidence extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3916–3927. Association for Computational Linguistics.

Yannis Papakonstantinou and Vasilis Vassalos. 1999. Query rewriting for semistructured data. *ACM SIGMOD Record*, 28(2):455–466.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Delai Qiu, Yuanzhe Zhang, Xinwei Feng, Xiangwen Liao, Wenbin Jiang, Yajuan Lyu, Kang Liu, and Jun Zhao. 2019. Machine reading comprehension using structural knowledge graph-aware network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5896–5901, Hong Kong, China. Association for Computational Linguistics.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Hasim Sak, Andrew W. Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 338–342. ISCA.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *CoRR*, abs/1809.02040.

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.

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.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *stat*, 1050:20.

HK von Heusinger and Urs Egli. 2012. *Reference and anaphoric relations*, volume 72. Springer Science & Business Media.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8527–8533. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yi-Ting Yeh and Yun-Nung Chen. 2019. Flowdelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 86–90. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 892–901. Association for Computational Linguistics.