# Evaluating Co-reference Chains based Conversation History in Conversational Question Answering

Angrosh Mandya, Danushka Bollegala, Frans Coenen
*Department of Computer Science*
*University of Liverpool*
*Liverpool, United Kingdom*
$\{$*angrosh,danushka,coenen*$\}$*@liverpool.ac.uk*

***Abstract*—This paper examines the effect of using co-reference chains based conversational history against the use of entire conversation history for conversational question answering (CoQA) task. The QANet model is modified to include conversational history and NeuralCoref is used to obtain co-reference chains based conversation history. The results of the study indicates that in spite of the availability of a large proportion of co-reference links in CoQA, the abstract nature of questions in CoQA renders it difficult to obtain correct mapping of co-reference related conversation history, and thus results in lower performance compared to systems that use entire conversation history. The effect of co-reference resolution examined on various domains and different conversation length, shows that co-reference resolution across questions is helpful for certain domains and medium-length conversations.**

***Keywords*-Co-reference based Conversation History, Conversational Question Answering, QANet**

## I. Introduction

In recent times, the focus of Machine Comprehension (MC) has shifted from answering questions that most likely have an answer in the contextual passage [1], [2] to answering more difficult questions that are conversational in nature, with answers often absent in the contextual passage [3], [4], [5]. The **Co**nversational **Q**uestion **A**nswering (CoQA) dataset is developed for measuring the ability of systems to answer such conversation-style questions. An important aspect of this dataset is the presence of large amounts of co-reference links between questions. Almost half of the CoQA questions (49.7%) contain explicit co-reference markers (e.g. *he, she, it*) that refer back to previous questions [3]. For example, for the sample conversation in Figure 1, the pronoun *'she'* in $q_2$ an $q_3$ refers back to the name of the cat (*'cotton'*) in $q_1$.

Once upon a time, in a barn near a farm house, there lived a little white kitten named Cotton. Cotton lived high up in a nice warm place above the barn where all of the farmer's horses slept. But Cotton wasn't alone in her little home, but shared her hay bed with her mommy and 5 other sisters.

$q_1$ : What color was Cotton?
$a_1$ : white

$q_2$ : Where did she live?
$a_2$ : in a barn

$q_3$ : Did she live alone?
$a_3$ : No

$q_4$ : Who did she live with?
$a_4$ : with her mommy and 5 sisters

Figure 1: Example conversation from CoQA Dataset

A key characteristic of CoQA systems such as DrQA, PGNet, DrQA+PGNet [3], Bidaf++ [4], FlowQA [5] is to use previous conversational history to provide contextual information essential for answering the current question. For example, to answer $q_3$ in Figure 1, the CoQA model [3], [4], [5] uses previous set of questions and answers $\{q_2, a_2\}$ and $\{q_2, a_2, q_1, a_1\}$ to input one and two conversation histories, respectively as contextual information. A major drawback of this method, is that the CoQA model can easily miss out on key information vital for answering conversational questions. For instance, to answer $q_4$ in Figure 1, using $\{q_3, a_3, q_2, a_2\}$, does not provide key input *'cotton'* as contextual information useful for answering $q_4$. However, identifying the link between pronoun *"she"* in $q_4$ and *"cotton"* in $Q_1$ through resolving co-reference chains in $\{q_3, q_2, q_1\}$, can allow us to use $\{q_1, a_1\}$ as inputs to the CoQA system rather than $\{q_3, a_3, q_2, a_2\}$. Thus, resolving co-reference chains in conversation history and providing more relevant contextual information can be useful for improving the performance of CoQA systems.

Based on this motivation, we focus on examining the usefulness of resolving co-reference chains in conversation history for the CoQA task. The main contribution of this paper is not to propose a state-of-the-art (SOTA) model for CoQA but to provide an empirical analysis of the effect of using co-reference chains in CoQA. To this end, we conduct several experiments using co-reference based conversation history to examine its influence against using the entire conversation history. To identify co-reference chains, we use NeuralCoref[1], a neural network based co-reference resolution tool. For our experiments, we modify QANet [6], a SOTA model for MC to include the conversational history as an input to the model.

The empirical results presented in this paper shows that even though co-reference links are present in large number across conversational questions in CoQA, the abstract nature of questions in CoQA renders it difficult to map a given question to co-reference related conversation history, resulting in lower performance compared to systems that use entire conversation history.

## II. Related Work

The CoQA dataset was proposed by [3] for evaluating convesational question-answering systems. The dataset provides human style conversational questions and preserves the naturalness of the answers evident in typical conversations. Besides developing the CoQA dataset, [3] also evaluated several standard MC models such as sequence-to-sequence (seq2seq), pointer-generator network (PGNet), Document Reader Question Answering (DrQA) system and a combined DrQA+PGNet model for CoQA as baseline models. Following the availability of CoQA dataset, several models have been proposed for CoQA. The BiDAF++ model [4] based on the Bidirectional Attention Flow (BiDAF) model [2] augmented with self-attention [7] was proposed to compute similarities between the context and conversation history. A Flow mechanism was used to add intermediate representations obtained during the process of answering previous questions [5]. SDNet, a contextual attention-based deep neural network [8] was proposed to leverage inter-attention and self-attention for CoQA. *Google SQuad* $2.0 +$ *MMFT (ensemble)*[2], the latest model listed on CoQA Leaderboard currently outperforms human performance on CoQA.

QANet [6], the SOTA model for MC was proposed to combine CNNs and self-attention networks to model local interactions and global interactions, respectively. QANet is shown to outperform SOTA MC models such as BiDAF [2], R-Net [9], Reinforced Mnemonic Reader [10] on SQuAD 1.0 dataset [1], both in terms of speed and accuracy.

As stated previously, the focus of this paper is not to propose SOTA for CoQA but to investigate the influence of co-reference links in answering conversational questions. Since QANet provides an efficient and faster means for MC, we propose to modify the QANet model in the context of CoQA. The modification of QANet to use similarity between context and conversation history is similar to the method proposed in BiDAF++[4]. Although various models [3], [4], [5], [8] have been proposed for CoQA, none of the studies have specifically focused on examining the influence of co-reference links in CoQA. To the best knowledge of the authors, this is the first study that provides an extensive empirical analysis of co-reference chains in CoQA. To this end, we use the modified QANet model to examine the performance of using co-reference chains based conversation history against using the available previous conversation history.

## III. Problem Formulation

Given a context passage $c$, a question $q_i$ and the conversational history $(q_1, a_1, ...q_{i-1}, a_{i-1})$, the task is to predict the answer $\hat{a}_i$.

$$p(\hat{a}_i|q_i) = f(c_i, q_1, a_1, ..., q_{i-1}, a_{i-1}) \quad (1)$$

However, instead of using the available $(q_1, a_1, ...q_{i-1}, a_{i-1})$, we propose to use the set of co-reference chains based conversation history $(q_k, a_k, ...q_{k-1}, a_{k-1})$, defined as the set of previous question-answer pairs that have co-reference links to the current question $q_i$.

$$p(\hat{a}_i|q_i) = f(c_i, q_k, a_k, ..., q_{k-1}, a_{k-1}) \quad (2)$$

Given two questions $q_i$ and $q_j$, we say that there exists a co-reference link between $q_i$ and $q_j$, if a word $u \in q_i$ refer to the same *person* or *thing* $v \in q_j$. Thus, the question-answer pair $\{q_j, a_j\}$ forms the co-reference chains based conversation history for $q_i$. For example, in Figure 1, given $q_4$ and $q_1$, we consider a co-reference link between words 'she' $\in q_4 \rightarrow$ 'Cotton' $\in q_1$, thus providing $\{q_1, a_1\}$ as the co-reference chains based conversation history for $q_4$. To evaluate the use of such conversation history, the QANet model is modified for CoQA as explained in the following section.

## IV. QANet Model for CoQA

The architecture of the modified QANet model for CoQA is described in Figure 2. We briefly describe the main components of the model. For a detailed explanation of QANet model, please refer [6].

### A. Input Embedding Layer

The embedding for each word $w$ is obtained by concatenating its word embedding with the character embeddings. The hyper-parameters of QANet [6] are retained, with word embedding initialized using $p_1 = 300$ dimensional pre-trained GloVe embeddings [11] and character embedding as a trainable vector of dimensionality $p_2 = 200$.

### B. Embedding Encoding Layer

The embedded input comprising $c, q_i$, and $\{q_1, a_1, ...q_i, a_i\}$ is provided as input to the encoding layer that consists of a stack of convolution, self-attention and feed-forward layers. The default network settings of the residual block are retained in the encoding layer.

[1] https://github.com/huggingface/neuralcoref

[2] Listed on March 29, 2019

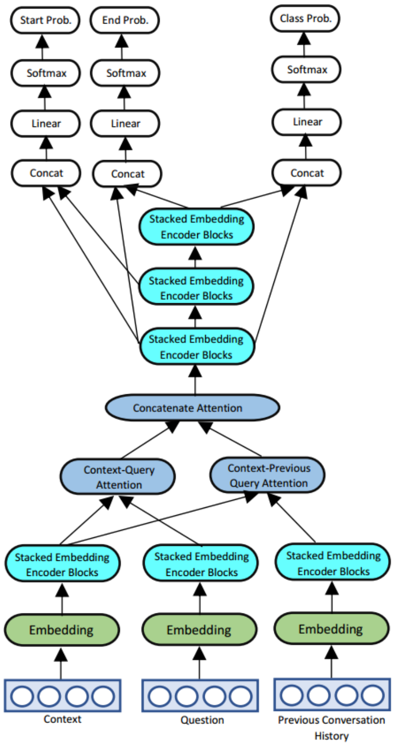
Start Prob.
End Prob.
Class Prob.
Softmax
Softmax
Softmax
Linear
Linear
Linear
Concat
Concat
Concat
Stacked Embedding Encoder Blocks
Stacked Embedding Encoder Blocks
Stacked Embedding Encoder Blocks
Concatenate Attention
Context-Query Attention
Context-Previous Query Attention
Stacked Embedding Encoder Blocks
Stacked Embedding Encoder Blocks
Stacked Embedding Encoder Blocks
Embedding
Embedding
Embedding
Context
Question
Previous Conversation History

Figure 2: Modified QANet model for CoQA. For residual network details, please refer [6]

The encoding layer receives as the input a vector of dimensionality $p_1 + p_2 = 500$ for each individual word and maps to one-dimensional convolution of dimensionality $d = 128$.

### C. Attention Layer

The main modification of QANet model for CoQA is in the attention layer, which besides computing similarity between $c$ and $q_i$ (context-query attention), also computes similarity between $c$ and conversation history $(q_1, a_1, ...q_{i-1}, a_{i-1})$ (context-conversation history attention).

Let C, Q and R be the encoded context, current question and conversation history, respectively. The similarities between each pair of words between $c$ and $q_i$, and $c$ and $(q_1, a_1, ..q_{i-1}, a_{i-1})$ is computed using similarity matrices $S_1 \in R^{n \times m_1}$ and $S_2 \in R^{n \times m_2}$, where $n$ is the length of $c$, and $m_1, m_2$ are the lengths of $q_i$ and $(q_1, a_1, ..q_{i-1}, a_{i-1})$, respectively. Each row of $S_1$ and $S_2$ is normalised using the softmax function to obtain matrices $\bar{S_1}$ and $\bar{S_2}$. The context-query and context-conversation history attention are computed as $A_1 = \bar{S_1} \cdot Q^T \in R^{n \times d}$ and $A_2 = \bar{S_2} \cdot R^T \in R^{n \times d}$. The tri-linear function [2] is used as the similarity function: $f(q_i, c) = \mathbf{W}_0[q_i, c, q_i \odot c]$, $f(q_j, c) = \mathbf{W}_1[q_j, c, q_j \odot c]$, where $\odot$ is the element-wise multiplication and $\mathbf{W}_0, \mathbf{W}_1$ are trainable vectors. To compute query-context attention $(B_1)$ and conversation history-context attention $(B_2)$, column normalized matrices $\bar{\bar{S_1}}$ and $\bar{\bar{S_2}}$ of $\bar{S_1}$ and $\bar{S_2}$ are computed using softmax function and $B_1$ and $B_2$ are obtained by $B_1 = \bar{S_1}\bar{\bar{S}}_1^T C^T$ and $B_2 = \bar{S_2}\bar{\bar{S}}_2^T C^T$.

### D. Model Encoding Layer

The input to the model encoding layer is $[c, a_1, c \odot a_1, c \odot b_1, a_2, c \odot a_2, c \odot b_2]$, where $a_1, a_2, b_1, b_2$ is a row of attention matrix $A_1, A_2, B_1, B_2$, respectively. The default settings of QANet are retained to share weights each of the 3 repetitions $(M_0, M_1, M_2)$ of the model encoder.

### E. Output Layer

The span selection method [2], [9] is used to predict the probability of each position in the context as being the start or end of an answer span. Specifically, the start and the end position probabilities are modelled as: $p^1 = \text{softmax}(\mathbf{W_1}[\mathbf{M_0}; \mathbf{M_1}])$ and $p^2 = \text{softmax}(\mathbf{W_2}[\mathbf{M_0}; \mathbf{M_2}])$. Simultaneously, we also output the probability $p^c$ of belonging to one of the four classes $\{$yes, no, unknown, span$\}$: $p^c = \text{softmax}(\mathbf{W_3}[\mathbf{M_0}; \mathbf{M_2}])$, where $\mathbf{W_1}, \mathbf{W_2}, \mathbf{W_3}$ are trainable variables and $\mathbf{M_0}, \mathbf{M_1}, \mathbf{M_2}$ are the output of the model encoder from bottom to top, respectively.

The loss function to learn the start and end probabilities is defined as the negative sum of the log probabilities of the predicted distributions of true start and end indices, averaged over all training examples:

$$L_0(\theta) = -\frac{1}{N} \sum_{i}^{N} \log(p^1_{y^1_i}) + \log(p^2_{y^2_i}) \tag{3}$$

The loss function to learn class probabilities is defined as the negative sum of the question belonging to a particular class, averaged over all training examples:

$$L_1(\theta) = -\frac{1}{N} \sum_{i}^{N} \log(p^c_{y^c_i}) \tag{4}$$

Here $y^1_i, y^2_i, y^c_i$ are respectively the groundtruth start and end positions, and the class of example $i$ and $\theta$ contains all trainable parameters. The total loss is:

$$L = L_0(\theta) + L_1(\theta) \tag{5}$$

### F. Inference

In the inference stage, for each question $q_i$, we first use $p^c$ to predict whether $q_i$ is answerable. If it is answerable, we predict the span $(s, e)$ with the maximum $p^1, p^2$, otherwise we predict the class as the answer for $q_i$.

## V. Experiments

### A. Evaluation metric

We conduct experiments on the CoQA dataset [3]. However, because the test set in CoQA is not publicly available and the main objective of this paper is to primarily investigate the effect of co-reference resolution in CoQA and not compete with systems listed on the CoQA leaderboard, we report our results only on the development set and not on the test set. To this end, we randomly choose 80% of CoQA training data as our train set and the remaining 20% as the development set, to develop the model. The learnt model is tested on the CoQA development set. Further, following [3], we report macro-average F1 score as the evaluation metric.

### B. Implementation

The original settings of the QANet model [6] is retained while the modifying the QANet model for CoQA. The co-reference chains were derived employing NeuralCoref[3], a pipeline extension for spaCy 2.0 that annotates and resolves co-reference clusters using a neural network.

### C. Results

The following explains the key results of this study

*1)* ***Using co-reference chains based history vs. Using available previous history****:* In order to examine the influence of co-reference chains in answering conversation questions, the following models were evaluated:

- **QANET-1-CCQ** and **QANET-2-CCQ**, model that uses previous one and two co-reference chain linked questions, respectively;
- **QANET-1-CCQA** and **QANET-2-CCQA**, model that uses previous one and two co-reference chain linked questions and answers, respectively;
- **QANET-1-PQA** and **QANET-2-PQA**, that uses previously available one and two questions and answers, respectively;

The overall performance of different models on the development set of the CoQA, in Table I shows that models using the entire previous conversation history (QANET-1-PQA and QANET-2-PQA) performs slightly better than models that use co-reference chains based conversation history (QANET-1-CCQ, QANET-2-CCQ, QANET-1-CCQA, QANET-2-CCQA). Interestingly for two domains "Children Stories" and "Literature", the co-reference chains based model (QANET-2-CCQA) achieves the best performance, indicating that the set of question-answer pairs identified based on co-reference resolution is helpful in answering conversational questions, particularly for these two domains. However, for other three domains the model using the available previous conversation history (QANET-2-PQA) achieves the highest performance. Though not conclusive, these results indicate that co-reference chains based conversation history can be helpful for CoQA in some cases.

| | Child. | Liter. | Mid-High. | News | Wiki. | Overall |
|---|---|---|---|---|---|---|
| QANET-1-CCQ | 62.4 | 56.7 | 63.1 | 66.9 | 67.4 | 63.4 |
| QANET-2-CCQ | 61.3 | 57.4 | 63.5 | 68.5 | 69.2 | 63.9 |
| QANET-1-CCQA | 65.7 | 59.3 | 64.6 | 70.2 | 68.2 | 65.3 |
| QANET-2-CCQA | **66.8** | **60.1** | 62.8 | 71.5 | 70.2 | 66.2 |
| QANET-1-PQA | 64.9 | 57.8 | 65.8 | 74.1 | 73.7 | 67.2 |
| QANET-2-PQA | 65.2 | 58.9 | **66.2** | **75.5** | **73.9** | **67.9** |

Table I: F1 scores of QANet based models for different domains in CoQA Development Set.

**Absence of contextual information**. The main reason for the poor performance of co-reference chain based models can be attributed to the absence of contextual information necessary for answering conversational questions, for co-reference based models. As seen in Table II, NeuralCoref facilitates identification of co-reference chain linked questions for about 80% of questions in the CoQA development set. This means that for the rest 20% of the questions, the contextual information in terms of previous questions and answers is not available for co-reference based models. Thus, these models have to entirely rely on the information available in the current question to answer it, resulting in a lower performance compared to QANET-1-PQA and QANET-2-PQA, which have conversation history for all questions, except the first. To address the problem of questions without co-reference chains based previous questions, we conducted experiments using the available previous conversation history for those questions where co-reference related previous conversation history was not available. However, the results (not reported here) showed that the inclusion did not help in improving the performance.

The co-reference chains based questions obtained for a sample paragraph in CoQA development set provided in Table V shows that there are no co-reference chains based previous questions for $q_2$ to $q_5$. The problem of not identifying co-reference linked previous questions for $q_2$ to $q_5$ is not because of the poor performance of NeuralCoref, but rather due to missing clues in $q_2$ to $q_5$ that does not help NeuralCoref in identifying co-reference links in previous questions. Further, as may be seen in Table V, questions $q_2$ to $q_5$ are quite abstract and change the topic of discussion, without providing any information about the change in the topic. This further makes it difficult to identify co-reference links in previous questions. These aspects further establish the complex nature of questions in CoQA dataset. The above results indicates that even though there are a high number of questions with co-reference links, connecting a given question to more relevant previous questions is quite challenging.

**Incorrect contextual information**. The poor performance of co-reference chains-based models can also be attributed to

[3] https://github.com/huggingface/neuralcoref

|  | Child. | Liter. | Mid-High. | News | Wiki. | Total |
|---|---|---|---|---|---|---|
| TQ | 1425 | 1630 | 1653 | 1649 | 1626 | 7983 |
| TQ_coref_links | 1181 | 1274 | 1385 | 1313 | 1223 | 6376 |
| (%) | 82.87 | 78.15 | 83.78 | 79.62 | 79.33 | 80.70 |

Table II: Number of co-reference chain linked questions for various domains in CoQA Development Set

|  | Questions in sequence | Co-reference chains based questions |
|---|---|---|
| 1. | What was the name of the fish? | - |
| 2. | What looked like a birds belly? | - |
| 3. | Who said that? | - |
| 4. | Was Sharkie a friend? | - |
| 5. | Did they get the bottle? | - |
| 6. | What was in it? | Did they get the bottle? |
| 7. | Did a little boy write the note? | Did they get the bottle? |
| 8. | Who could read the note? | Did they get the bottle? |
| 9. | What did they do with the note? | Did they get the bottle? |
| 10. | Did they write back? | Did a little boy write the note? Did they get the bottle? |
| 11. | Were they excited ? | Did a little boy write the note? Did they get the bottle? |

‘

Table III: Co-reference chains based questions obtained using NeuralCoref for a sample paragraph in domain “Children Stories” in CoQa development set.

combining incorrect contextual information with the current question. For example for questions $q_8$ to $q_{10}$ in Figure 1, the same question ($q_5$) is used as the co-reference chains-based previous question. However, information provided by $q_5$ is not very helpful in answering questions $q_8$ to $q_{10}$, and thus results in lower performance of the model. Although, experiments were conducted to include questions within a certain window in the question sequence, the performance (not reported here) almost remained the same.

**Paragraphs with higher proportion of co-reference based conversation history.** The performance of QANet model on conversations that have higher proportion of co-reference chains-based conversation history (80% and 60% questions have conversation history) (shown in Table IV), achieves a slightly better F1-score of 66.5 and 65.9, respectively, against a lower F1-score of 65.3 achieved with considering all conversations with co-reference linked questions. Although, there is a slight improvement the difference is not significant, indicating that even a lower percentage of questions that do not have any previous history can affect the model's performance. Further, it is also important to note that the errors induced by the co-reference resolution system can be compounding in nature and thus, can significantly lower the performance.

**Using answers with questions.** The results provided in Table I also indicates that co-reference chain based conversation history alone is not sufficient for answering conversational questions. As seen in Table I, the QANET-1-CCQ QANET-2-CCQ models which uses coreference-chain based questions alone perform poorly in comparison to the models QANET-1-CCQA, QANET-2-CCQA, QANET-1-PQA, QANET-2-PQA which employs both questions and corresponding answers together. Thus, it is useful to use previous answers along with questions, to augment the contextual information necessary to answer conversational questions.

|  | Overall |
|---|---|
| QANET-80%-CON-CCQA | 66.5 |
| QANET-60%-CON-CCQA | 65.9 |
| QANET-ALL-CON-WITH-CCQA | 65.3 |

Table IV: F1 scores of co-reference based QANet models for conversations with different percentage of co-reference chains in CoQA development set.

|  | Questions in sequence | Questions with replaced pronouns |
|---|---|---|
| 1. | What color was Cotton? | What color was Cotton? |
| 2. | Where did she live? | Where did Cotton live? |
| 3. | Did she live alone? | Did Cotton live alone? |
| 4 | Who did she live with? | Ho did Cotton live with ? |

‘

Table V: Replacing co-referenced pronouns in questions with referenced words from previous questions.

*2) Replacing co-referenced pronouns in questions:* . Experiments were also conducted to evaluate the performance co-reference based models by replacing co-referenced pronouns in the current question with referenced words in previous questions. For example, using NeuralCoref facilitates identification of co-reference link between the pronoun “she” $\in q_2, q_3, q_4$ and the noun “Cotton” $\in q_1$ (Figure 1). Using this co-reference link, the pronoun “she” is replaced with noun “Cotton” as shown in Table V.

The performance of the QANet model using current question alone (QANET_REG_QUEST) and using questions with replacing co-reference pronouns (QANET_COREF_REP_QUEST) is provided in Table VI. The results in Table VI shows that it is difficult to obtain a comparable score using current question alone and thus, contextual information in terms of conversation history plays an important role in achieving optimum performance for CoQA. However, interestingly a small improvement (F1-score of 58.88 vs. 57.30) is achieved when co-referenced pronouns in questions are replaced with either *person* or *thing* that it refers to in the previous questions. The replacement of co-referenced pronouns particularly seem to help in answering “No”, “Unknown”, and “Span prediction” type questions.

The minimum, maximum and the average number of questions for paragraphs in the domain of “Children Stories” in CoQA development set are 10, 25 and 14, respectively.

| | QANET_REG_QUEST | QANET_COREF_REP_QUEST |
|---|---|---|
| ‘ Yes | 80.62 | 54.84 |
| No | 34.57 | 66.60 |
| Unknown | 37.50 | 48.48 |
| Span | 56.69 | 58.64 |
| Overall | 57.30 | 58.88 |

Table VI: F1-scores of model using current question with replacing co-reference pronouns for the domain of “Children Stories” in CoQA dataset.

| | QANET_REG_QUEST | QANET_COREF_REP_QUEST |
|---|---|---|
| ‘ | *Conversation length* $\leq 14$ | |
| Yes | 82.81 | 54.68 |
| No | 35.71 | 66.32 |
| Unknown | 41.66 | 63.63 |
| Span | 54.42 | 58.75 |
| Overall | 55.76 | 58.95 |
| | *Conversation length* $> 14$ | |
| Yes | 79.16 | 54.94 |
| No | 33.86 | 66.77 |
| Unknown | 33.33 | 33.33 |
| Span | 58.68 | 58.54 |
| Overall | 58.59 | 58.83 |

Table VII: F1-scores of model using current question with replacing co-reference pronouns on different conversation length for domain “Children Stories” in CoQA dataset.

Therefore the performance of QANET_REG_QUEST and QANET_COREF_REP_QUEST was examined on two groups: (a) paragraphs with $\leq 14$ questions; and (b) paragraphs with $> 14$ questions as shown in Table VII. As Table VII shows, replacing co-referenced pronouns in paragraphs with $\leq 14$ questions significantly helps in answering question types such as “no” (66.32 vs. 35.71), “unknown” (63.3 vs. 41.66) and “span prediction” (58.75 vs. 54.42). These results indicates that more accurate co-reference links are obtained in conversations with lower to medium (around 14) number of questions. However, as the length of conversations increase, there seems to be little effect of using co-reference links, which is most likely due to poor co-reference links between questions.

*3) Comparison with CoQA baseline models.:* As mentioned previously, the objective of this paper is not to compete against SOTA for CoQA task. However, it needs to be noted that the QANet models using one and two conversation history (QANET-1-PQA and QANET-2-PQA) achieves an F1-Score of 67.2 and 67.9, respectively on the CoQA development set. These results are slightly better than the performance of baseline models: Seq2Seq (27.5); PGNet (45.4); DrQA (54.7); DrQA+PGNet (66.2), obtained on the CoQA development set [3]. The results of QANET-1-PQA and QANET-2-PQA are also comparable with scores of BiDAF++w/0-ctx (63.4);BiDAF++w/1-ctx (68.6); BiDAF++w/2-ctx (68.7) [4] on CoQA development set. The modified QANet model described in this paper follows a similar approach of BiDAF++ to combine context with conversation history, indicating the usefulness of QANet in the context of CoQA.

## VI. Conclusion

We presented in this paper an empirical analysis of using co-reference chains based conversation history for CoQA. The results presented in this paper shows that although there exists a large proportion of co-reference links across questions in CoQA, the abstract nature of questions renders it difficult to map together co-reference related questions for large number of questions, resulting in lower performance in comparison to models that use previously available conversation history. The results also show that using co-reference related questions can help in conversations which have fewer questions.

## References

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” *arXiv preprint arXiv:1606.05250*, 2016.

[2] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bidirectional attention flow for machine comprehension,” *arXiv preprint arXiv:1611.01603*, 2016.

[3] S. Reddy, D. Chen, and C. D. Manning, “Coqa: A conversational question answering challenge,” *arXiv preprint arXiv:1808.07042*, 2018.

[4] M. Yatskar, “A qualitative comparison of coqa, squad 2.0 and quac,” *arXiv preprint arXiv:1809.10735*, 2018.

[5] H.-Y. Huang, E. Choi, and W.-t. Yih, “Flowqa: Grasping flow in history for conversational machine comprehension,” *arXiv preprint arXiv:1810.06683*, 2018.

[6] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, “Qanet: Combining local convolution with global self-attention for reading comprehension,” *arXiv preprint arXiv:1804.09541*, 2018.

[7] C. Clark and M. Gardner, “Simple and effective multi-paragraph reading comprehension,” *arXiv preprint arXiv:1710.10723*, 2017.

[8] C. Zhu, M. Zeng, and X. Huang, “Sdnet: Contextualized attention-based deep network for conversational question answering,” *arXiv preprint arXiv:1812.03593*, 2018.

[9] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, “Gated self-matching networks for reading comprehension and question answering,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 189–198.

[10] M. Hu, Y. Peng, Z. Huang, X. Qiu, F. Wei, and M. Zhou, “Reinforced mnemonic reader for machine reading comprehension,” *arXiv preprint arXiv:1705.02798*, 2017.

[11] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.