COSE474-2024F: Final Project Proposal Dialogue-Based Question Answering: Evaluating NLP Models on Multi-Participant Conversations

Jeongmin Moon

1. Introduction

These days, it is important to provide concise summaries of long texts. Many people often seek summaries of news articles, academic papers and books, often using large language models (LLMs) like ChatGPT to generate those summaries. As a result, Natural Language Processing (NLP) models such as BERT (Devlin et al., 2019) have shown great performance on Question Answering (QA) tasks.

However, most of QA datasets such as Stanford Question Answering Dataset (SQuAD (Rajpurkar et al., 2016)) include document-style input texts. This means NLP models are good at understanding those texts and answering questions based on them. But it is unclear how well the models perform on QA datasets containing different styles of input texts. For example, if conversational texts from a group chat involving multiple participants are given, can the model answer to the question asking about the content of the conversation?

Therefore, it is important to address that uncertainty. By using pre-trained model that shows high performance on document-based QA datasets, I analyzed its capability on conversation-based QA datasets.

2. Related Works

Question Answering (QA) is a key task in NLP that has gained significant attention due to its broad applicability in areas such as information retrieval, customer service, and chat bot system. This section explores existing QA tasks, highlighting diverse approaches and datasets.

Traditional QA Systems Early QA systems usually relied on rule-based approaches, leveraging manually created knowledge bases and linguistic patterns. However, these approaches lacked scalability, as they required extensive domain-specific knowledge and engineering.

Extractive QA The introduction of large-scale datasets, such as SQuAD, marked a turning point for extractive QA. These tasks involve identifying a span of text from a given passage that directly answers the question. Transformer-based models, particularly BERT and its variants like RoBERTa (Liu et al., 2019), have shown outstanding performance in those tasks by leveraging self-attention mechanisms to understand contextual relationships.

Conversational QA Conversational QA focuses on multi-turn interactions including multiple participants where conversational context is maintained across sequential queries. Datasets like CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) evaluates conversational capabilities. Transformer-based approaches have shown great performance in understanding conversational context.

3. Method

3.1. Problem Definition

Given a conversational dataset D which includes several conversations between two speakers in chatting system, I created a question-answer pair D_{QA} for each conversations using LLM: $D_{QA} = LLM(D_C)$.

Then, given D and D_{QA}, I created a conversation-based QA dataset D_{conv} and a document-based QA dataset D_{doc} using a rule-based program P: D_{conv}, D_{doc} = $P(D, D_{QA})$.

Then, given D_{conv}, D_{doc} and a pre-trained NLP model M, the performance for each dataset was calculated and compared.

4. Problem definition & challenges

I will utilize a pre-trained model M, a dialogue-based QA dataset D, and a document-based QA dataset D', which is produced by preprocessing the dialogue-style input texts of D into a document-style.

Then, the performance of M on the QA task when using D and D' will be compared. Additionally, the performance of M after fine-tuning on D will be also compared.

The main challenges are as follows: it is not difficult to find conversational datasets, but most of them doesn't include any QA labels for the conversations in the datasets. Furthermore, converting D into the document-style D' is another difficulty.

5. Datasets

It is planned to use a conversational dataset from AI-Hub. Because it doesn't include QA labels, proper questions and answers about the conversations should be generated to use it as a dataset D. Alternatively, it can be another choice to find a dialogue-based QA dataset.

6. State-of-the-art methods and baselines

BERT outperformed not only previous state-of-the-art (SOTA) models but also human performance on QA tasks with SQuAD dataset. The performance of BERT in QA task with D and D' will be compared.

7. Schedule & Roles

October 26 $\sim$ November 8: Construct and preprocess dataset D.

November 9 $\sim$ November 22: Evaluate the performance of M on QA task using D and D'.

November 23 $\sim$ November 29: Evaluate the performance of M after fine-tuning on D.

November 30 $\sim$ December 6: Write the report.

References

Choi, E., He, H., Iyyer, M., Yatskar, M., tau Yih, W., Choi, Y., Liang, P., and Zettlemoyer, L. Quac : Question answering in context, 2018. URL https://arxiv.org/abs/1808.07036.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge, 2019. URL https://arxiv.org/abs/1808.07042.