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, NLP models like such as BERT(Devlin et al., 2019) have shown great performance on Question Answering (QA) tasks.

However, most of QA datasets such as SQuAD (Stanford Question Answering Dataset) use document-style input texts. This means 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 will analyze its capability on dialogue-based QA datasets.

2. 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.

3. Related Works

Question Answering is a fundamental task in NLP that involves answering questions based on a given input text. Many NLP models typically process document-style inputs and are designed to answer fact-based questions by extracting or generating relevant information from the text. These models often leverage transformer-based architectures, including BERT, which have shown significant improvements in understanding the text.

4. 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.

5. 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.

6. 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

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