Summarization on SParC

Shreya Dixit

Department of Computer Science, Yale University

Introduction

The overarching goal of the LILY Dialog to SQL project is to create a State-of-the-Art Dialog System that can help facilitate natural language interactions with databases. Within this goal, first, I in conjunction with 14 other college students under the direction of Tao Yu to create a cross-domain, large scale dataset for conversational interactions, now called SParC. This dataset also captures contextual dependencies between questions in the same example, and includes a diverse range of semantic content.

I also worked with Alex Fabbri, and Tao Yu on a summarization model on the SParC dataset that summarizes the sub-questions in each dialog to predict the target question. This problem intends to create a model that preserves the logical correctness, not just attains a high accuracy measure for a sample dataset.

My final project will establish the baseline that makes great use of “Get to the point: Summarization with Pointer-Generator Networks” by Abigail See, Peter J. Liu, and Christopher D. Manning. This original work was extended to create a State

A domain-adaptable summarization model that preserves logical correctness can have applications in every sector of society from education, business and will significantly aid in decision making. The difficulty of the task stems from the fact that effective summaries must be concise, comprehensive, informative, and relevant. Existing models tend to reproduce facts incorrectly and subject to the question. This suggests that the model has a lot of room for improvement, especially in its generation.

Why is this task interesting?

The Error Analysis thus far has consisted of a comparison of 100 predicted sequences their original test string and the categorization of the errors. The categories are inspired by their SQL counterparts. To explain, if both the predicted and actual strings were converted into SQL, the categories are the errors that would the strings incur. The dataset could incur multiple errors.

The model was able to produce only 2 logically correct output. Filter errors and logical errors consisted of the largest area of weakness. Specifically, the combination of filter errors and select table errors also manifested strongly. As a whole, the model is able to correctly predict the correct construction of the question 70 percent of the, but often connects the wrong subject to the question. This suggests that the model has a lot of room for improvement, especially in its generation.

Conclusion

In this project, we tackle the problem of dialog summarization. I adapt the pointer-generator model with coverage by See et. al to the SParC dataset and conducted error analysis. Thus far, we can reasonably conclude that the current model needs to improve on correct identification of the subject of outputs and on logical errors. Moving forward, we can continue this analysis by also creating a method of and evaluating the abstraction ability of this model. Also, it would be robust to conduct similar error analysis on an extractive model that only uses the input to predict words, in the interest of creating an accurate baseline.

Acknowledgement

I sincerely appreciate the instruction from professor Drago, Tao Yu and Alex Fabbri.