Summarization on SParC

Shreya Dixit

Department of Computer Science, Yale University

LILY Lab

Introduction

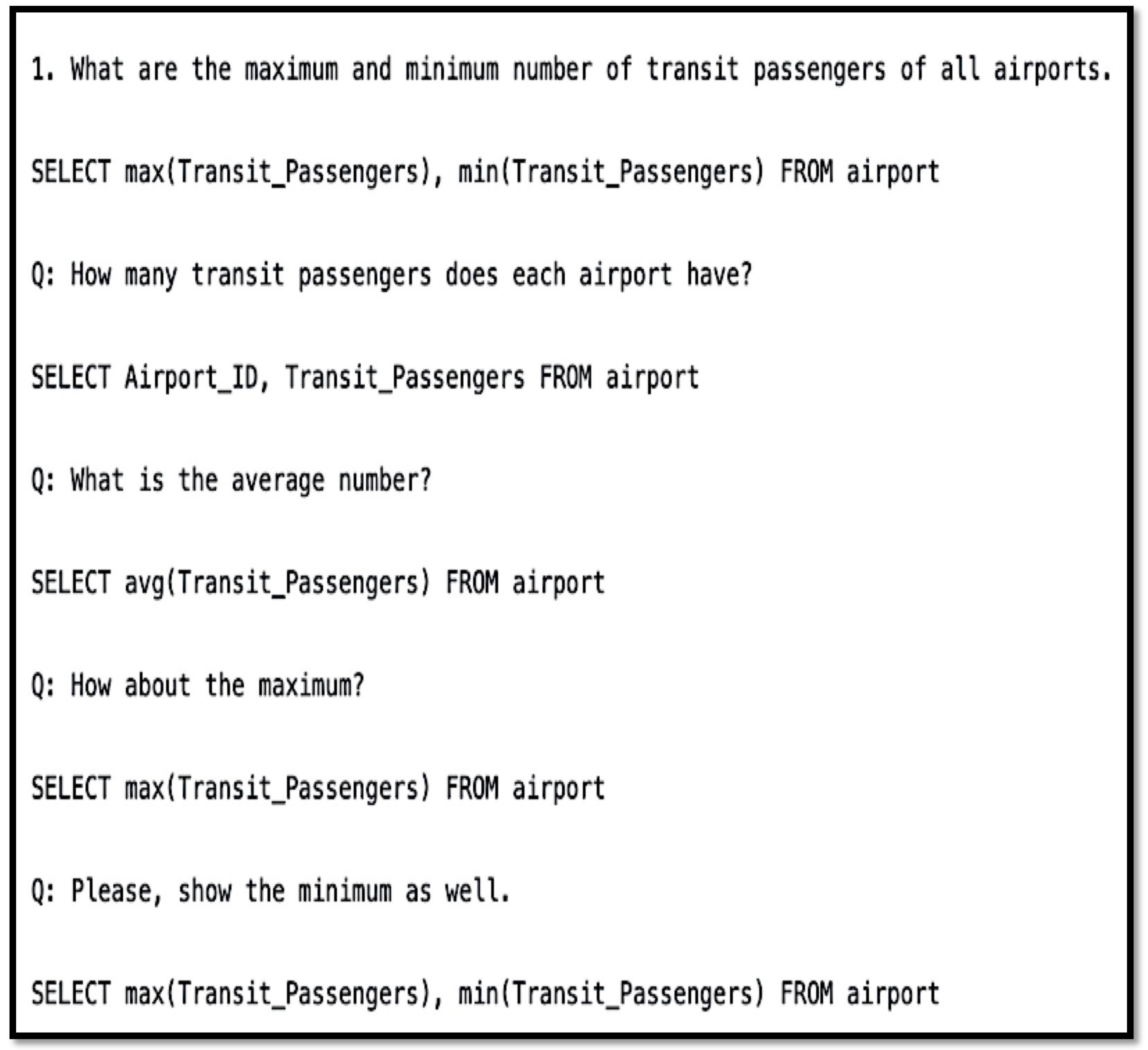
The overarching goal of the LILY Dialog to SQL project is to create a State-of-the-Art Dialog System that can help facilitate natural language interactions with databases. Within this goal, first, I in conjunction with 14 other college students under the direction of Tao Yu to create a cross-domain, 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 applied to the CNN Daily Mail dataset.]I extend this model to our dataset, a more difficult task than the news dataset due to more variation in the way information is presented (not a common lead sentence structure), and a smaller input size.

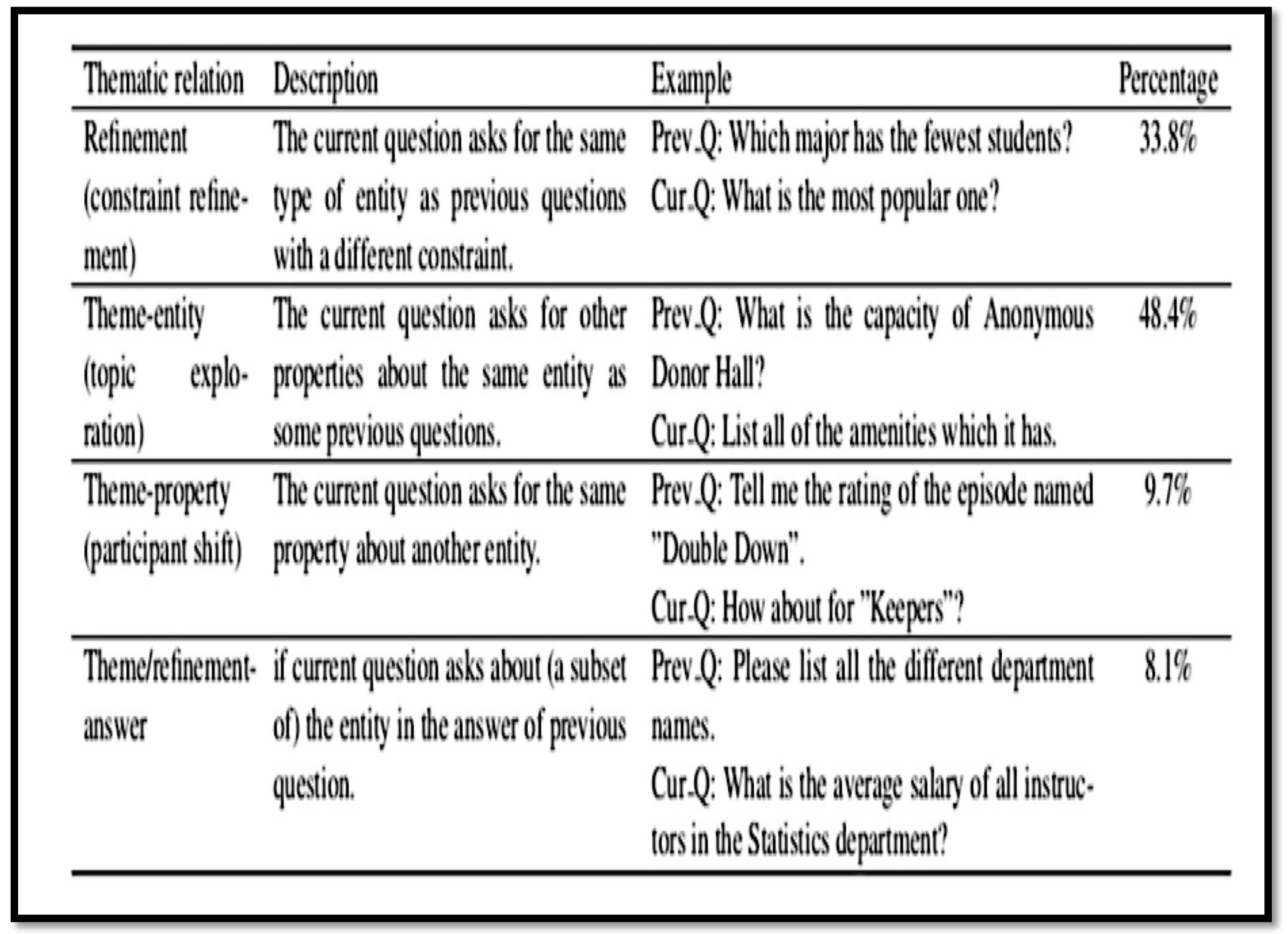
Why is this task interesting?

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 repeat themselves (See et.al, 2017). Creating a robust summary model must be able to be flexible among different domains, different lengths of inputs, as well as different types of documents. The SParC dataset contains many of these challenges for existing models due to the smaller input size of the dataset, the variability of question structures, and the thematic relations embedded into the input questions (shown in Figure 3).



```
1. What are the maximum and minimum number of transit passengers of all airports.

SELECT max(Transit_Passengers), min(Transit_Passengers) FROM airport

Q: How many transit passengers does each airport have?

SELECT Airport_ID, Transit_Passengers FROM airport

Q: What is the average number?

SELECT avg(Transit_Passengers) FROM airport

Q: How about the maximum?

SELECT max(Transit_Passengers) FROM airport

Q: Please, show the minimum as well.

SELECT max(Transit_Passengers), min(Transit_Passengers) FROM airport
```

Figure 1. Input example from SparC dataset. For the dialog summarization task, we are using the subset questions as our input, and the summary question as our output.



Thematic relation	Description	Example	Percentage
Refinement (constraint refinement)	The current question asks for the same type of entity as previous questions with a different constraint.	Prev.Q: Which major has the fewest students? Cur.Q: What is the most popular one?	33.8%
Theme-entity (topic exploration)	The current question asks for other properties about the same entity as some previous questions.	Prev.Q: What is the capacity of Anonymous Donor Hall? Cur.Q: List all of the amenities which it has.	48.4%
Theme-property (participant shift)	The current question asks for the same property about another entity.	Prev.Q: Tell me the rating of the episode named "Double Down". Cur.Q: How about for "Keepers"?	9.7%
Theme/refinement-answer	if current question asks about (a subset of) the entity in the answer of previous question.	Prev.Q: Please list all the different department names. Cur.Q: What is the average salary of all instructors in the Statistics department?	8.1%

Figure 3. Thematic Relations between Questions in the SParC dataset. Adapted from “SParC: Cross Domain Semantic Parsing in Context”

Model Used

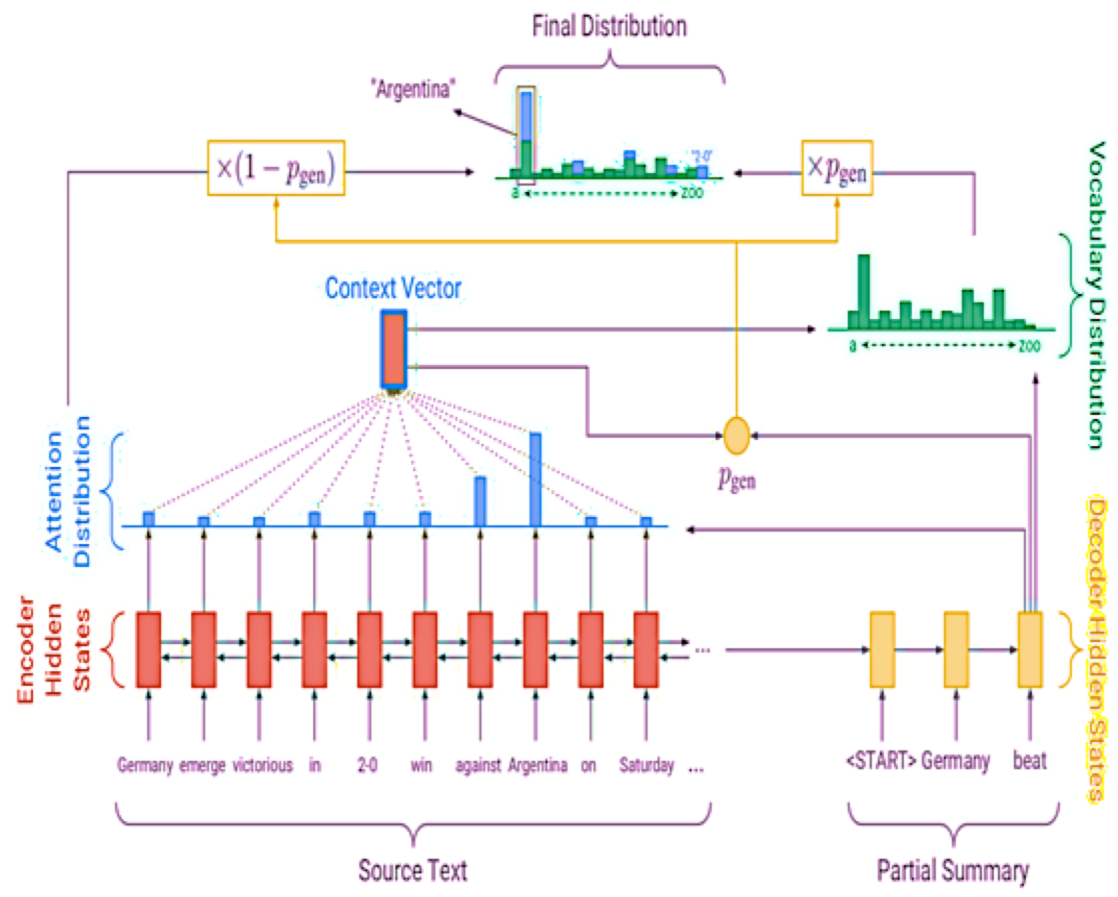


Figure 3. Baseline sequence to sequence model with attention. Adapted from “Get to the Point: summarization with Pointer-Generator Networks” by See et. al.

The pointer-generator with coverage model is the model we adopt in this dataset.

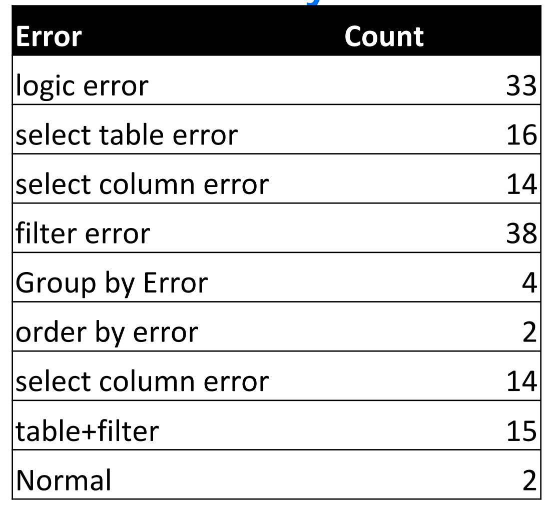
The model iterates on the previous encoder-decoder model with attention with its three features: pointing, which draws from the input words to create a more correct representation of information, generator, which allows the model to generate words from the vocabulary outside of the input, and coverage, which discourages the model repeatedly choosing the same words (See et. Al).

We found that the best model used 15,000 steps and additional separation delimiters between each sub-question in dataset.

Actual	Predicted
What is the average height and width of paintings that are oil medium in gallery 241?	What is the average height and resolution of paintings whose oil ACCT-211?

Figure 4. Example of one of 2 correct logical actual and prediction sequences returned from model

Error Analysis



Error	Count
logic error	33
select table error	16
select column error	14
filter error	38
Group by Error	4
order by error	2
select column error	14
table+filter	15
Normal	2

Figure 5. Results from Error Analysis

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 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.