TabCSE: Tablue Understanding with Contrastive Learning with aggregation results

Whanhee Cho
University of Utah
Comput Science
whanhee@cs.utah.edu

Abstract

erw

1 Introduction

Understanding relation between columns are imporatnt for data lake data analysis.

What could be the useful relationship for models? - use prompt or table to text. - current data use existing table -> instruction tuning?

How about adding more information to the model? When humans try to interpret table, we usually see the caption, table title, column values, and relationship between columns.

The caption is a sort of summarization of table. Even when the column header/type is missing we can summarize the table. : or the metadata could be enough.

train with wikitables, masking column headers, training with relationship between columns,

Table itself understanding very good. But need metadata while training. But how about learning from multiple tables?snetence embedding get their success from comparing their sentences like SBERT or SimCSE. However, in the table data, it is hard to find similar dataset such as NLI data SBERT or SimCSE used.

However, in the sense of understanding hypothesis can derive premise, in the tabular data, we can use groupby as the proof of understanding the table. Note that this is the first trial of using a pair of tables to understand a table.

2 Architecture

2.1 Serialization

There are lots of serialization methods for tabular data. Many works try to embed row-wise or column-wise for understanding the structure of tabular data.

Understanding the structure of the tabular data somewhat important. However, I could not find any related works the performance of learning the tabular data structure. TaBERT uses utternance.

In our work, we only uses the linearzation of table to see the effectiveness of tables.

3 Result

We used BERT-base-uncased from huggingface.

4 Conclusion

5 Related Works

Table-GPT is GPT for tabular data. They trained GPT with instruction tuning of number of data : . They added table for the instruction pair. They prompt the only one column and candidate column type that the model can choose from them. “Column type annotation using ChatGPT” only uses prompts to get column type annotation, they consist of prompts with a partial table. Doduo annotates columns based on each column and relations between columns using contextualized column embedding. TURL focuses on entity in tables utilzing other metadata such as table title, table caption in wikipeida table. Moreover, watchhog did contrastive learning inside the table using self-supervised learning method.