January, 2025

APM_5AI29_TP

Language Models and Structured Data

Mid-term Project Report

Acronym of the Team: Table Turners
Name: Dmitry Timkin, Ivanina Ivanova, Mark Daychman, Renata Mindiyarova

---

Enhancing Text-to-Table Information Extraction with Updated BART Models

Abstract

In this paper, we revisit the proposed 'text-to-table' information extraction task, which converts unstructured text content into structured tables using sequence-to-sequence models. We explain the core idea of the original paper [1] to use table constraints and relation embeddings to improve accuracy. In our work, we further evaluate whether more recent versions of BART-base and BART-large used in the original paper can further enhance the performance on the same datasets.

Problem Statement

Information extraction (IE) is one of the fundamental tasks in natural language processing. It aims to convert text into more dense and structured knowledge representations, making the information more accessible for further applications. One of the examples of such structured data type are tables. Before this paper was released, IE into table usually required predefined schemas, making it inflexible and domain specific. Accurate text-to-table generation, especially without a predefined schema, has many real-life applications across various domains, including automated structured reports from legal documents, financial statements, and medical records.

The paper formalizes this task as a sequence-to-sequence (seq2seq) problem and proposes a solution based on BART, which is known for its strong performance in text generation. The proposed method incorporates additional techniques such as table constraints and table relation embeddings to ensure that the generated tables are correctly structured. However, the original study only evaluates the performance of BART-base and BART-large from 2020, which were state-of-the-art at the time. We extend the work done in the paper by testing the provided implementation with more modern models (INSERT THE NAME OF THE CHOSE MODEL HERE) to see if the accuracy can be further improved. We think the results will significantly improve,

because newer models are specifically optimized for long-text generation. Because of all the extra tokens that are needed to properly encode a table using text, the output is usually fairly long, so long-text generation models are expected to perform better.

Table constraints are implemented during the coding process to enforce the that tables have a consistent structure. Specifically, the model determines the number of columns from the first generated header row and ensures that every generated row has the same number of columns. This prevents incomplete rows or irregular table formats. This method improves the syntactic accuracy of the model's output.

Table relation embeddings help with the table cell alignment by incorporating their relationship with their respective row and column headers. During the table generation, the model uses row and column relation embeddings to identify which row each non-header cell belongs to. These embeddings are added as relational vectors in the self-attention function. This helps improve the table's coherence and accuracy.

We aim to address the following key questions:

- Do more recent versions of BART-base and BART-large outperform the older versions used in the original text-to-table study?
- Do updated models reduce common errors in table generation, such as missing headers, incorrect row alignments, or incomplete tables?
- Are improvements consistent across different datasets and domains, or do they vary depending on the table complexity and text length?

Ultimately, we simply seek to address is understanding the extent to which advancements in pre-trained language models can enhance performance on complex IE tasks such as text-to-table generation.

Method/Overall Architecture

Description of your model

Experimentation

We faced several challenges running the original code from the paper due to its reliance on outdated Python versions and deprecated modules. The installation required extensive manual version matching, use of an older pip version, and the manual cloning of certain packages from GitHub repositories to resolve compatibility issues. Despite these hurdles, we managed to obtain preliminary experimental results on January 5th. We remain optimistic that the complete dataset will be processed and ready for inclusion in the final submission.

- Dataset & Dataset Statistics
- Experimental Results based on Evaluation Metrics
- Error Analysis

TO BE COMPLETED AFTER ALL THE RESULTS ARE OBTAINED

Discussion

TO BE COMPLETED AFTER THE EXPERIMENTATION IS DONE

Comparison with expectations, limitations, lessons learned, and perspectives.

References

[1] Xueqing Wu, Jiacheng Zhang, and Hang Li. Text-to-table: A new way of information extraction, 2022.