P11-13

Large Language Models as Generalizable Text-to-Table Systems

Steven Coyne[1]
[1]Tohoku University

Yuyang Dong[2]
[2]NEC Corporation

Contact: coyne.steven.charles.q2@dc.tohoku.ac.jp

Summary

- **Text-to-Table** is the task of summarizing text in table form with **no user query**.
- Previous studies focused on **fine-tuned models** that **do not generalize well** to new datasets.
- We explored the generalizability of prompted LLMs in this task.
- LLMs **generalized** but **did not outperform baseline systems** on in-distribution data.
- LLMs **wrote tables well** but **struggled with schema writing** and other high-level concerns.

Background

- Task introduced in Wu et al. (2022)
- Input: Sentences or paragraphs of text
- Output: Text-based tables
- Previously studied with fine-tuned models (e.g., BART)

Input Text:

```
The Wizards launched yet another comeback on Tuesday, this time feasting on the
relatively inexperienced Los Angeles Lakers. Washington not only overcame a 13 -
point fourth quarter deficit, but won by double digits as well. The team shot over
51 percent from the field on the night and outscored LA 37 - 13 in the fourth. At
the crux of the win was All-Star point guard John Wall, who scored 34 points...
```

Output Tables:

Team

Team	Percentage of field goals	Points in 4th quarter
Lakers	51	13
Wizards		37

Player

Player	Assists	Field goals attempted	Field goals made	Points	Total rebounds	Steals
John Wall	14	25	14	34		4
D'Angelo Russel	9	21	10	28	6	
Jordan Clarkson		19	10	22		

Example from the RotoWire dataset, adapted from Wu et al. (2022)

Challenges

Generalizability:

- Fine-tuned models struggle to generalize to other data

Training Dataset ↓	E2E Scores	WikiTableText Scores	WikiBio Scores	RotoWire Scores
E2E	98.56	12.66	5.22	0
WikiTableText	42.47	80.44	25.02	0
WikiBio	24.89	33.34	74.79	0
Rotowire	0	0	0	91.5

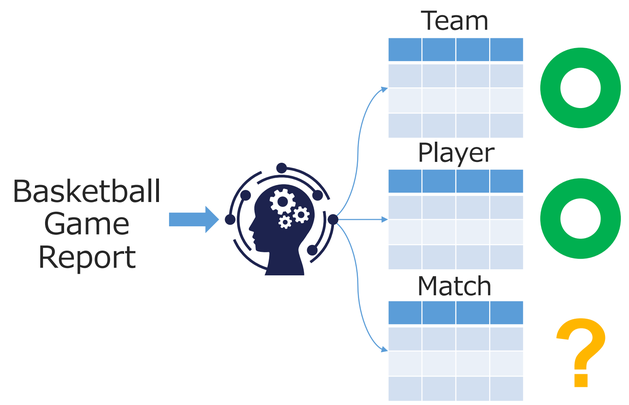
Average Non-header cell BERTScore F1 results of models trained following Wu et al. (2022)

Reasoning:

- Entity Linking
- Contextual Understanding
- Number of tables to make
- Design of columns & rows

Evaluation:

- Open-ended task
- Reference-based metrics
- What about alternatives?



Generation Approaches

Three Subtasks:

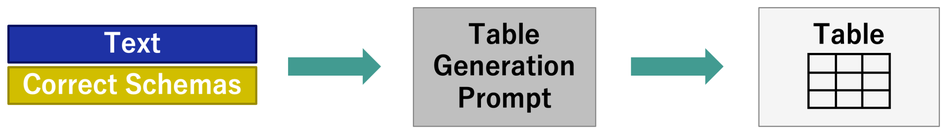
Group Generation:
- Entities are listed and sorted into groups
- Output: List of groups

Schema Generation:
- Columns are defined and schemas written
- Output: One JSON schema per planned table

Table Generation :
- Tables generated based on schema(s) and text
- Output: Tables written in tabular JSON

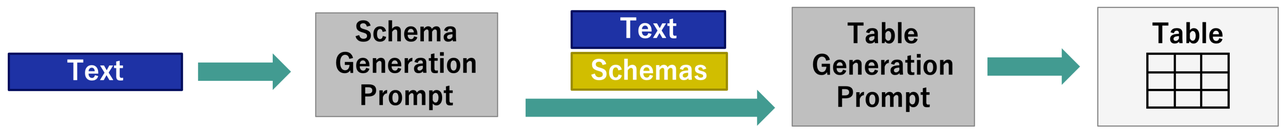
Prompt Settings:

◆ "Gold Schema" + Table Generation

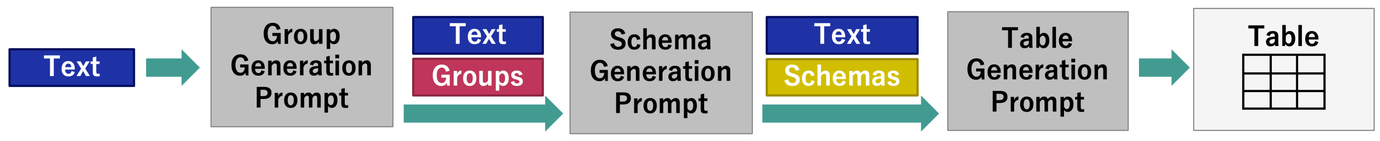


"Gold Schema" Setting
- Isolates table generation
- Establishes upper bound

◆ Schema Generation + Table Generation



◆ Group Generation + Schema Generation + Table Generation



Results

- The LLMs show moderate performance on all datasets
- "Gold Schema" scores are high; some exceed baseline
- The fine-tuned baselines outperform the LLMs in other settings
- The grouping step degrades performance except for few-shot RotoWire

Approach	Cell Type	E2E	WikiTableText	WikiBio	RotoWire
Baseline (Wu et al. 2022)	Header	99.88	91.98	93.13	93.14
	Non-header	98.56	74.63	78.18	92.97
Gold Schema + Text (Zero-shot)	Header	**100**	**99.63**	**99.93**	**99.86**
	Non-header	69.56	66.25	72.29	**93.76**
Gold Schema + Text (5-shot)	Header	**100**	**99.71**	**99.94**	**100**
	Non-header	88.45	**77.49**	**78.99**	**97.23**
Schema + Text (Zero-shot)	Header	77.46	58.38	69.66	55.10
	Non-header	41.87	28.57	42.51	44.19
Schema + Text (5-shot)	Header	98.89	82.42	90.28	83.65
	Non-header	86.27	50.07	65.11	84.94
Group + Schema + Text (Zero-shot)	Header	74.00	58.98	75.14	39.75
	Non-header	30.86	26.70	32.78	28.05
Group + Schema + Text (5-shot)	Header	98.52	82.72	90.02	84.03
	Non-header	85.43	48.16	64.20	85.38

Comparison of baseline and various prompt settings using `gpt-3.5-turbo-1106`. All results are BERTScore F1 scores.

Conclusions

- **The LLMs showed generalizable performance** in this task.
- For a specific domain, they typically **performed worse than in-domain fine-tuned baselines**.
- They **performed well on table writing** when given a schema.
- This suggests the challenge lies in **identifying schemas appropriate for the domain**.
- **Reference-free metrics** would help this task greatly.

ANLP 2024 Poster Session