AIREV at the NTCIR-18 U4 Task

Xin Fan, Kazuya Uesato, Yuma Hayashi, Tsuyoshi Morioka

AIREV, Inc.

{xfan, kuesato, yhayashi, tmorioka}@airev.co.jp

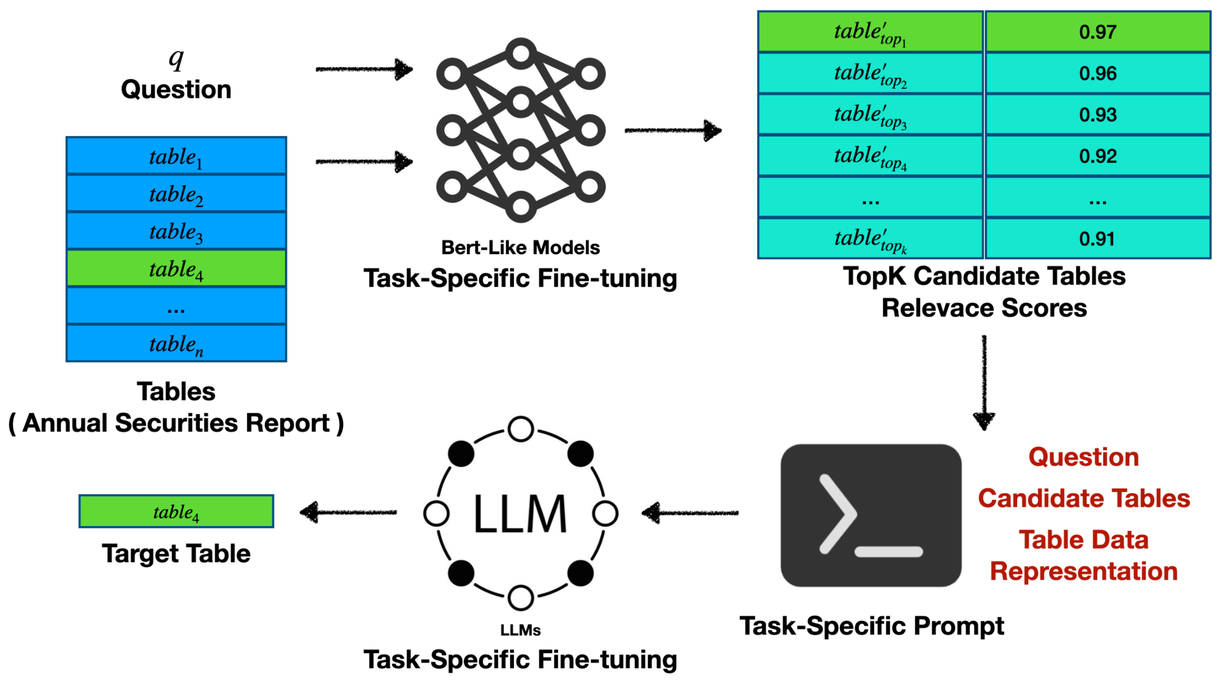
Results and Contributions

	Table Retrieval (TR)	Table Question Answering (TQA)	
		Cell Id	Cell Value
Public	0.9117 [2nd prize]	0.9785 [1st prize]	0.9493 [1st prize]
Private	0.9115 [2nd prize]	0.9473 [2nd prize]	0.9047 [2nd prize]

- Proposing two-stage approaches to TR task, which can efficiently reduce the input tokens
- Designing TQA-task specific prompts to the LLMs and several postprocessing for TQA Cell Value retrieval

Approaches to TR task

■ Two-stage retrieval approaches



- Utilizing LLMs (Llama-3 JP Elyza 8B) to predict the most relevant table with the query
- Reducing the candidate tables to input to the LLMs by fine-tuned BERTs
- The mean of # tables per a report is 221.9, which might lead to a significant reduction in the reasoning power of LLMs and a significant increase in computational consumption

Approaches to TQA task

■ Cell Id Retrieval by LLMs

- Utilizing LLMs to predict the most relevant cell in the table with the query (question)
- Unlike TR task, all cell values in the table can be contained in a prompt

■ Postprocess for Cell Value Retrieval

Detecting and multiplying by a unit

100 [百万円] → 100000000
300 [千株] → 300000

Reversing sign in facing with triangle marks

△ 150000000 → -150000000
▲ 300000000 → -300000000

Converting a format of date

2023年03月31日 → 2023-03-31

Removing redundant characters

(※1) 100,000 → 100000
(100,000) → 100000

■ Design of LLM prompt for TR task

Prompt:

"あなたのタスクは、候補の表の中から、質問に最も関連している表を選んでください。\n",

"以下は質問です： \n",

"***質問***: \n",

"**{Question}** \n",

"各表については、表のIDと内容があります。\n",

"以下は候補の表です： \n",

"***候補の表のリスト***: \n",

"[\n",

"**{List}** \n",

"] \n",

"答えは、候補の表の中から、質問に最も関連している表のIDだけ答えてください。\n",

"答えの表のIDは、候補の表にあるものでなければならない。\n",

"答えには表のID以外の余計なものを書かないでください。\n",

{List}:

$table'_{top_1}$: *table data representation*$_{top_1}$
$table'_{top_2}$: *table data representation*$_{top_2}$
$table'_{top_3}$: *table data representation*$_{top_3}$
...
$table'_{top_5}$: *table data representation*$_{top_5}$

■ Design of LLM prompt for TQA task

Prompt:

"あなたのタスクは、表のCell情報を基づいて、質問にもっとも関連しているCellを選んでください。\n",

"以下は質問です： \n",

"***質問***: \n",

"**{Question}** \n",

"各表については、表の各CellのIDとValueがあります。\n",

"それに、表の各CellのIDについては、このCellの行と列をしめしています。\n"

"以下は表のCell情報です： \n",

"***表のCell情報***: \n",

"[\n",

"**{List}** \n",

"] \n",

"答えは、表のCell情報中から、質問に最も関連しているCellのIDだけ答えてください。\n",

"答えのCellのIDは、表のCell情報にあるものでなければならない。\n",

"答えにはCellのID以外の余計なものを書かないでください。\n",

{List}:

$r1c1$: *cell value*$_{1.1}$
$r1c2$: *cell value*$_{1.2}$
$r1c3$: *cell value*$_{1.3}$
...
$r10c7$: *cell value*$_{10.7}$