Korean-Specific Dataset for Table Question Answering

Changwook Jun, Jooyoung Choi, Myoseop Sim, Hyun Kim, Hansol Jang, Kyungkoo Min

LG AI Research

Abstract

Existing question answering systems mainly focus on dealing with text data. However, much of the data produced daily is stored in the form of tables that can be found in documents and relational databases, or on the web. To solve the task of question answering over tables, there exist many datasets for table question answering written in English, but few Korean datasets. In this paper, we demonstrate how we construct Korean-specific datasets for table question answering: Korean tabular dataset is a collection of 1.4M tables with corresponding descriptions for unsupervised pre-training language models. Korean table question answering corpus consists of 70k pairs of questions and answers created by crowd-sourced workers. Subsequently, we then build a pre-trained language model based on Transformer and fine-tune the model for table question answering with these datasets. We then report the evaluation results of our model. We make our datasets publicly available via our GitHub repository and hope that those datasets will help further studies for question answering over tables, and for the transformation of table formats.

Introduction

In this paper, for the Korean-specific table question an- swering task, we present KO-TaBERT, a new approach to train BERT-based models that learn jointly textual and structured tabular data by converting table struc- tures. To address this, we first create two datasets writ- ten in the Korean language: the tabular dataset contains conversion formats of around 1.4M tables extracted from Korean Wikipedia documents for pre-training lan- guage models, and the table question answering dataset for fine-tuning the models. The table question answer- ing dataset consists of 70k pairs of questions and answers, and the questions are generated by crowd- sourced workers considering question difficulty. Addi- tionally, we introduce how structured tables are con- verted into sentence formats. The conversion formats play a crucial role for models to learn table structural information effectively without changing embeddings. Second, we follow BERT architecture (Devlin et al., 2018) to pre-train a language model with the converted strings from millions of tables, and fine-tune models on the table question answering dataset.

Methodology

Figure 1: Example of *Infobox* and *WikiTable* in a Wikipedia document

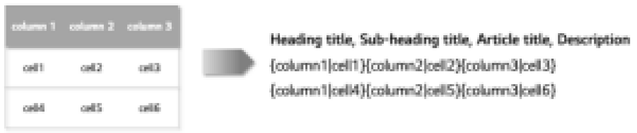

Figure 2: Converting *WikiTable* format into string sequences for pre-training input. The converted table strings are added with descriptions for the Wikipedia article.

describes question types generated regarding the question difficulties.

- Level1: Question [column−others] where [column−base] has [value]
- Level2: Question [column−others] where [column−base] has [condition]
- Level3: Question [column−base] where [column−others] has [value]
- Level4: Variation of the questions in other levels
- Level5: Question min or max in [column−base] where [column−others] has [value of numbers, dates, ranks, etc.]

Figure 3: Examples of a natural language question set related to a table for football match-results in UEFA Euro 2004 according to question levels

Experiments

Question difficulty	EM	F1
Level1	89.6	93.1
Level2	89.1	92.3
Level3	86.1	89.4
Level4	81.7	85.5
Level5	67.8	70.6
Overall	83.9	87.2

Table 2: Comparison of model performance according to each level of questions in the crowdsourced dataset.

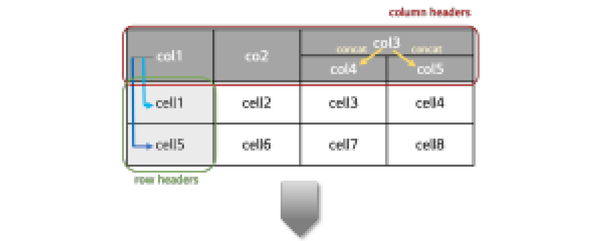

Figure 4: Example of the new conversion approach for complicated structured tables that consisting of merged and multi column headers into sentence strings.

Dataset source	Format	EM	F1
KorQuAD 2.0	v1	64.5	74.5
KorQuAD 2.0	v2	**69.1**	**78.4**
Crowd-sourced	v1	83.9	87.2
Crowd-sourced	v2	**87.2**	**91.2**

Table 3: Comparison of model performance with different table parsing approaches. Format v1 is the table conversion described in Figure 2.

Conclusion

we demonstrate how tabular data is converted into linearlised texts containing struc- tural information and properties. We construct a tabular dataset by extracting tables and converting them into sentence strings with tabular structural informa- tion for pre-training a language model. We also cre- ate a table question answering corpus with paid crowd- sourced workers. The corpus consists of 70k pairs of questions and answers related to tables on Wikipedia articles, and those questions are generated specifically considering levels of question difficulty. We conduct experiments on the table question answering task. Our model achieves the best performance when converted table sentence strings include richly structural features. In future work, we aim to extend the model for com- plex question answering over texts and tables with the generation of multimodal questions to jointly handle question answering from textual and tabular sources. We hope that our datasets will help further studies for question answering over tables, and for the transforma- tion of table formats.