Few-Shot Table Understanding: A Benchmark Dataset and Pre-Training Baseline

Ruixue Liu*, **Shaozu Yuan***, **Aijun Dai, Lei Shen**
Tiangang Zhu, Meng Chen, Xiaodong He
JD AI, Beijing, China
{liuruixue, yuanshaozu, daiaijun1, shenlei20}@jd.com
{zhutiangang3, chenmeng20, xiaodong.he}@jd.com

Abstract

Few-shot table understanding is a critical and challenging problem in real-world scenario as annotations over large amount of tables are usually costly. Pre-trained language models (PLMs), which have recently flourished on tabular data, have demonstrated their effectiveness for table understanding tasks. However, few-shot table understanding is rarely explored due to the deficiency of public table pre-training corpus and well-defined downstream benchmark tasks, especially in Chinese. In this paper, we establish a benchmark dataset, FewTUD, which consists of 5 different tasks with human annotations to systematically explore the few-shot table understanding in depth. Since there is no large number of public Chinese tables, we also collect a large-scale, multi-domain tabular corpus to facilitate future Chinese table pre-training, which includes one million tables and related natural language text with auxiliary supervised interaction signals. Finally, we present FewTPT, a novel table PLM with rich interactions over tabular data, and evaluate its performance comprehensively on the benchmark. Our dataset and model will be released to the public soon.

1 Introduction

Relational tables, as a typical form of structured data on the Web, store a vast amount of knowledge. Table understanding (Wang et al., 2012) aims to understand the semantics of tabular data as well as the associated text jointly, which further improves the evaluation results of several tasks, including table question answering (Khalid et al., 2007; Sun et al., 2016; Bogin et al., 2019), table retrieval (Zhang and Balog, 2018), and table fact verification (Chen et al., 2020; Zhang et al., 2020). Recently, inspired by the huge success of pre-trained language models (PLMs) in understanding *free-form* natural language (NL) sentences, researchers have attempted to model *structured* data using pre-training techniques. Various table pre-training models (Herzig et al., 2020; Yin et al., 2020; Yu et al., 2021; Liu et al., 2021; Cheng et al., 2021; Shi et al., 2022; Dong et al., 2022) have been proposed and made remarkable progress in learning the structured schema of tables and the alignment between the input text and the schema.

In real-world scenario, table understanding usually faces more challenging situations, in which tables are from different domains and each table contains very limited annotations. Despite its importance, few-shot (Lake et al., 2015) table understanding is rarely explored in previous works due to the following three obstacles. First, the deficiency of public well-designed benchmark datasets for few-shot table understanding makes the model evaluation inconvenient. Second, the lack of public large-scale high-quality table pre-training corpora blocks the exploration of table PLMs. Lastly, a table pre-training baseline tailored to the few-shot table understanding is also needed for better performance comparison. Especially, the lack of benchmark datasets and pre-training corpora in Chinese hinders the research on table understanding.

To fill the above gaps, in this paper, we focus on the dataset construction for table-understanding related tasks in Chinese. We first establish a few-shot table understanding benchmark dataset, **FewTUD**, with five table related tasks, by which research and exploration can be carried out extensively. Different from existing table understanding tasks, which concentrate on the information inside tables, our tasks, based on real-world scenarios, lay emphasis on the interaction between tables and the corresponding NL text (e.g., Table Fact Verification, Table QA, Table Selection, and Schema Detection), and focus on the whole content of tables (e.g., Table Classification). We collect tables with meta information and the corresponding NL text from the Web, then manually annotate the dataset for

*Both authors contributed equally to this work.

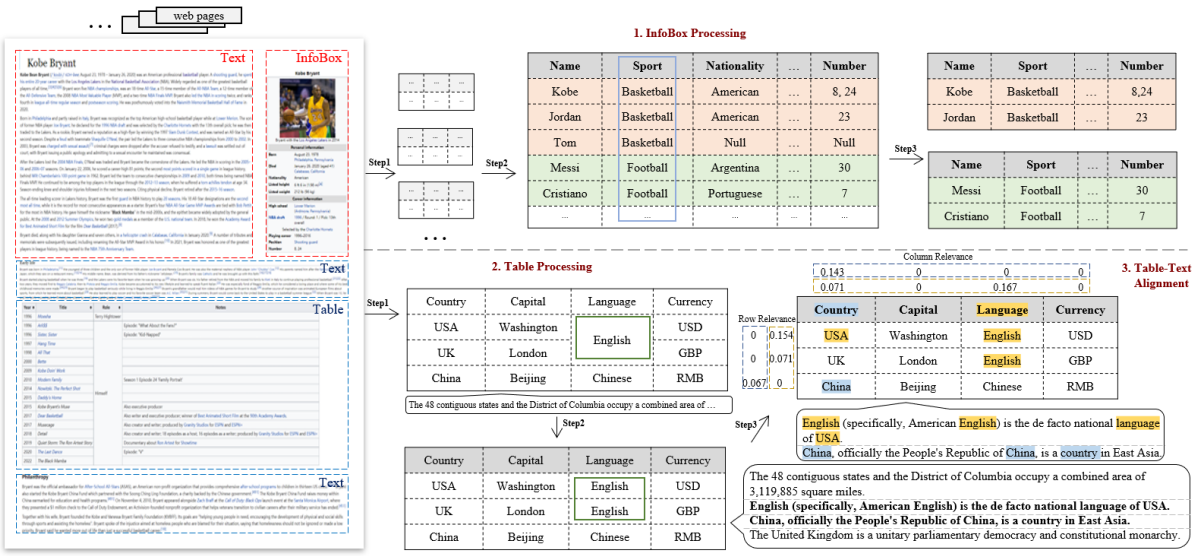

Figure 1: Data Construction of TPC-1M with three main procedures: (1) InfoBox processing, (2) table processing with column identification and cell value splitting, (3) table-text alignment and relevance score estimation.

each task, aiming to provide an evaluation criterion for few-shot table understanding.

Second, we collect a large-scale Chinese table pre-training corpus, **TPC-1M**, with more than 1 million tables and their associated NL text. How to obtain semantically relevant NL text is critical for constructing the table pre-training corpus. Previous works either obtain the associated NL text by synthesising pseudo NL text on available tables (Yu et al., 2021; Liu et al., 2021; Qin et al., 2021) or crawling surrounding NL text of tables simply based on position information (Yin et al., 2020; Herzig et al., 2020). Differently, we locate the associated NL text via semantic matching approach. We also provide the relevance score between the NL text and each row/column of the table as auxiliary supervised interaction signal, which may further facilitate the table pre-training.

Lastly, we propose a novel table pre-training model, **FewTPT**, to serve as the baseline for the few-shot table understanding benchmark. Different from previous table pre-training models (Yin et al., 2020; Herzig et al., 2020), we focus on catching the interactions between tables and NL text. Specifically, we devise two novel spatial-aware pre-training tasks, i.e., column relevance prediction (CRP) and row relevance prediction (RRP), which predict the relevant column from table schema and the relevant row from table content respectively based on the supervised interaction signals. We also employ variants of Masked Language Modeling (MLM) objective to predict the tokens in NL text, column names and cell values. Finally, we conduct extensive experiments to show the advantage of FewTPT against several strong baselines.

To summarise, our main contributions are threefold. (1) We establish a Chinese benchmark dataset for few-shot table understanding, which includes five downstream tasks: Table Fact Verification, Table Question Answering (Table QA), Table Selection, Schema Detection, and Table Classification. We hope the benchmark can be a testbed for future few-shot table understanding research in Chinese. (2) We contribute a large-scale high-quality table pre-training corpus in Chinese, which covers 1 million tables across 13 domains and their semantically-associated NL text. (3) We propose a novel table pre-training model to serve as the baseline for the benchmark, and experimental results show the competitiveness of our model.

2 Construction of TPC-1M and FewTUD

In this section, we introduce the construction of table pre-training corpus, TPC-1M, and benchmark dataset for few-shot table understanding, FewTUD.

2.1 TPC-1M Corpus

Corpus Collection. We first crawl huge amount of table pages from the web (e.g., Baidu Baike[1] and E-commerce website[2]). Web pages with the same topic field are grouped into one domain. For

[1]https://baike.baidu.com.
[2]https://www.jd.com.

	Max	Min	Mean	Median
Row	120	2	11.6	6
Column	80	2	6.5	6

Table 1: Data statistics of TPC-1M.

a Baike page, it contains a special kind of table InfoBox[3] that illustrates the properties of one entity (i.e., celebrity) and can be regarded as the single-row table. To enrich information density and table variety, we aggregate similar InfoBoxes into a multi-row table. Figure 1(1) demonstrates the process of InfoBox pre-processing. We first combine InfoBoxes with similar schemas into one big table. Then we split it into sub-tables according to the row values in a specific schema, which is determined by the lowest entropy with Maximum Entropy algorithm (Jaynes, 1982) (i.e., *Sport* in Figure 1(1)). We further filter some sparse columns and rows of these sub-tables to reduce complexity. For other tables in the web page, some are already corrupted, i.e., "<th>" (table header) is omitted, or some cells are merged. For the absence of table header tag, we train a binary classifier[4] to identify whether the first row or first column is the table header. For the merged cells, we split them into individual ones based on the position information. The above process is shown in Figure 1(2).

Finally, we need to pick out the associated NL text for each table. For previous web-crawled tables and context from English Wikipedia (Lehmberg et al., 2016), the context is mainly mined based on position information and may not be semantically related to tables. As Figure 1(3) shows, we find the NL text from the title, caption and text descriptions around a table. To locate the semantically-relevant NL text accurately, we calculate the linguistic overlap ratio with Jaccard Similarity algorithm (Niwattanakul et al., 2013) between the table and its candidate text snippets. The text snippets with the top-N similarity scores are chosen as the associated NL text. Furthermore, we assign a relevant score to each row/column based on the n-gram overlap between the row/column and the associated NL text. The relevant scores indicate how likely the row/column is mentioned in the associated NL text, which can be regarded as auxiliary supervised interaction signals between tables and the NL text.

[3]Similar to https://en.wikipedia.org/wiki/Help:Infobox.

[4]We construct the training set from well-formed tables and the classification accuracy is 95%.

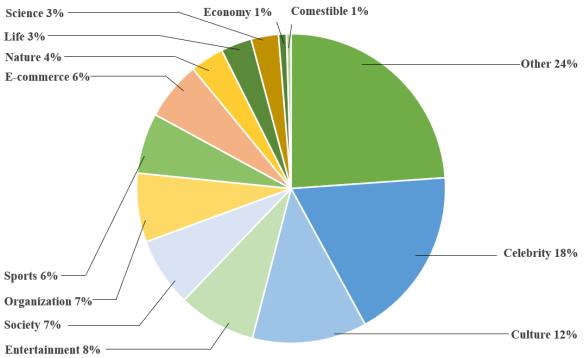

Figure 2: Domain distribution of TPC-1M.

Statistics Analysis. After pre-processing, we obtain 1,002k tables with 3,014k semantically associated NL sentences. Figure 2 shows the distribution of domains. TPC-1M covers 13 different domains, including culture, character, celebrity, sports, and e-commerce, and the statistics of row and column number are illustrated in Table 1.

2.2 FewTUD Dataset

In this sub-section, we introduce the benchmark dataset of five few-shot table understanding tasks. For each task, we first invite annotators to manually annotate the dataset[5]. To follow the setting of few-shot learning, we fix the test set as *query set*, and sample different numbers of samples (N-shot) from training set as *support set*. During experiments, we randomly sample the *support set* from training set for 10 times, then report the average performance as the final result. Table 2 shows the detailed statistics of our benchmark dataset. Examples of each benchmark dataset are also illustrated in Figure 3 for better understanding.

Table Fact Verification. Table Fact Verification is a fundamental task for NLP and can benefit many downstream applications, such as misinformation detection and fake news detection. This task aims to verify whether a NL hypothesis is entailed or refuted by the given table as knowledge. Considering there is no off-the-shelf Chinese table fact verification dataset, we then construct one based on the NL text-table pairs we collected. 20 crowdsourcing annotators were invited to label the NL text-table pairs and check the quality. If the NL text claims the fact from the corresponding table, the pair is labelled as positive; otherwise, the anno-

[5]Annotations have been cross-checked and the inter-annotator agreement score is 90% (Fleiss' Kappa score (Fleiss and Cohen, 1973)).

Task	#Table	Train	Dev	Test	N-shot
Table Fact Verification	81	4.2k	0.8k	0.8k	{2,4,6,8,10,15}
Table QA	108	1.2k	0.5k	0.5k	{1,2,3,4,5,10}
Table Selection	266	2k	1.5k	1.5k	{1,2,3,4,5,10}
Schema Detection	100	1k	1k	1k	{1,2,3,4,5,10}
Table Classification	3k	1k	1k	1k	{5,10,15,20,30,50}

Table 2: Dataset statistics of few-shot experiments.

tators need to write a positive statement manually. Then, the sampled positive statements are rewritten as negative ones by replacing the key mentions in the NL texts with other cell values from the corresponding tables. In total, 5,800 fact statements for 81 unique tables were annotated and the ratio for positive and negative sentences is 1:2. For example, suppose a given table is about the milk powder product, the positive statement is "澳大利亚进口奶粉12-36月龄幼儿适用 (Milk powder imported from Australia is suitable for children aged 12-36 months)", while the negative statement is "意大利进口奶粉12-36月龄幼儿适用(Milk powder imported from Italy is suitable for children aged 12-36 months)". The model needs to verify which statement is true based on the table information. The evaluation metric for this task is F_1 score.

Table Question Answering. Given a table and question, Table QA aims to locate the exact cell in table that can answer the question. Given a query about *"How tall is Kobe Bryant?"* and a table on athletes information, the task is designed to pick the cell value *198cm* involved in the table to answer the question. Considering that it is labor-intensive and time-consuming to collect large-scale, high-quality question-answer pairs based on tables, we set up a question generation framework following Shi et al. (2022) to obtain questions given a table and answers. Specifically, we leverage the fine-tuned T5 model (Raffel et al., 2019) to generate questions automatically, then ask the annotators to manually check these ⟨question, answer, table⟩

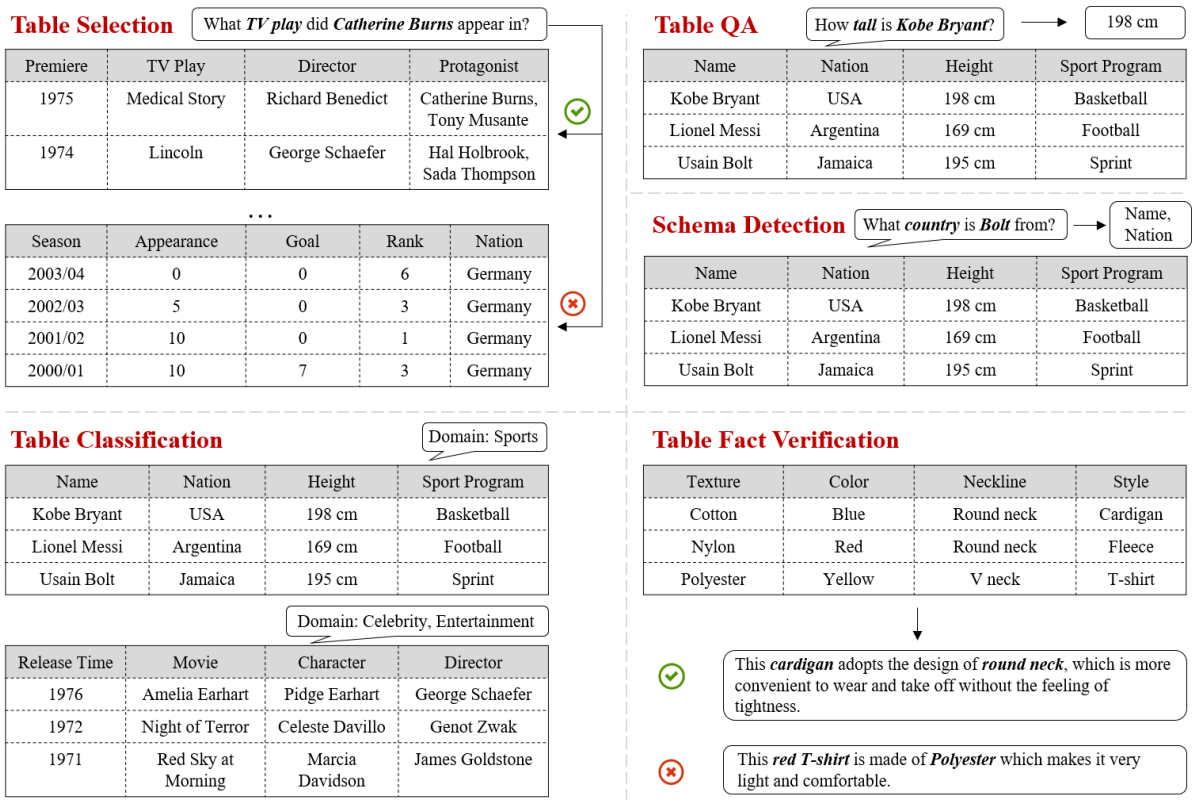

Figure 3: Examples of benchmark dataset **FewTUD**. The original data is in Chinese, we translate it into English for illustration.

triplets and rewrite semantically irrelevant or disfluent questions. Two public Chinese Machine Reading Comprehension datasets (He et al., 2018; Wang et al., 2020a) are used to train the T5 model.

Table Selection. Table selection or table retrieval is an important task as table contains valuable information to explore in various domains. This task aims to select the most relevant table from a list of candidates to answer the given query. Here, we construct the dataset for few-shot table selection similar to Table QA task. We first generate 5,000 queries for 266 different tables given the table-related context and its overlapped cell value. Then annotators are invited to re-check the generated queries and rewrite those disfluent ones. The revised query-table pair is noted as positive, while the one with replaced table (any other table) is negative. The ratio of positive and negative samples is 1:9 for test set, and 1:1 for training and validation sets. Given a query *"What TV play did Catheriner Burns appear in?"* and a list of tables, the model is expected to compute the matching score between each query-table pair and extract the most appropriate table to answer this question. $R_{10}@1$ is the evaluation metric for this ranking task (Lowe et al., 2015).

Schema Detection. As an important task for semantic parsing, schema detection bridges the gap between NL query and database schema. Given a query and table, schema detection requires to identify the column names mentioned in the query. For instance, given a query *"What country is Bolt from?"* and a table containing information on athletes, the model is required to predict the related column names of *Name, Nation* involved in the query. It is a challenging task aiming to explore the model's performance on covering semantic and structural correspondences and extracting general knowledge from structural data during table-query interaction. Here, we construct the dataset based on two public Chinese datasets (Wang et al., 2020b; Sun et al., 2020). Instead of using the whole sketch for SQL generation, we only keep the column information in SELECT, ORDER clauses, and WHERE conditions as the ground truth for schema detection.

Table Classification. Different from the above mentioned tasks that jointly learn the representations of table and text, table classification focuses on table understanding without additional text-based input. Given the table on movies and their characters' information, the task is expected to predict the multiple domain labels. Here we construct the dataset based on the domain labels we assigned to the collected tables in TPC-1M. The task can be formulated as a multi-label classification, which is designed to examine the performance of table pre-training models on structured-information understanding. For example, suppose a table contains Vivien Leigh's films, the model is expected to predict the domain labels *Celebrity, Entertainment*.

3 Table Pre-training Baseline

3.1 Model Architecture

Figure 4 illustrates the architecture of our proposed model FewTPT, which is based on the pre-trained language model BERT (Devlin et al., 2018) to encode the table and NL text and learn the structural-aware representations.

Input Embedding. FewTPT linearizes the input into a sequence of tokens by concatenating the query and table meta data by rows. For each cell in the table input, we adopt row linearization (Yin et al., 2020) to represent a cell with column name, column type and cell value together. Moreover, a [CLS] token is inserted at the beginning of whole input sequence and each cell is separated by the [SEP] symbol. Thus, for each token T_i of position i, the embedding E_T^i is defined as follows:

$$E_T^i = E_W^i + E_S^i + E_P^i + E_R^i, \tag{1}$$

where E_W, E_S, and E_P are the token embedding, segment embedding, and position embedding following Devlin et al. (2018). E_R represents row embedding inspired by Herzig et al. (2020).

Gated Cell Representation. Although row linearization method can accommodate the input of tabular data, previous methods (Yin et al., 2020) simply utilize the pooling of the column name, type and value as representation of a cell. Considering that the column name and cell value emphasize different information, it's necessary to distinguish the table column/cell separately. Here, we adopt a gated fusion mechanism to selectively integrate column name E_n^{jk}, column type E_t^{jk} and cell value E_v^{jk} to obtain the cell representation E_c^{jk}:

$$\begin{aligned} g^{jk} &= \sigma(W_g E_n^{jk} + U_g E_t^{jk} + V_g E_v^{jk} + b), \\ E_c^{jk} &= g^{jk} \odot E_n^{jk} + (1 - g^{jk}) \odot E_v^{jk} + E_t^{jk}, \end{aligned} \tag{2}$$

where W_g, U_g, V_g are learnable matrices, and jk represent the j^{th} column and k^{th} row of the table.

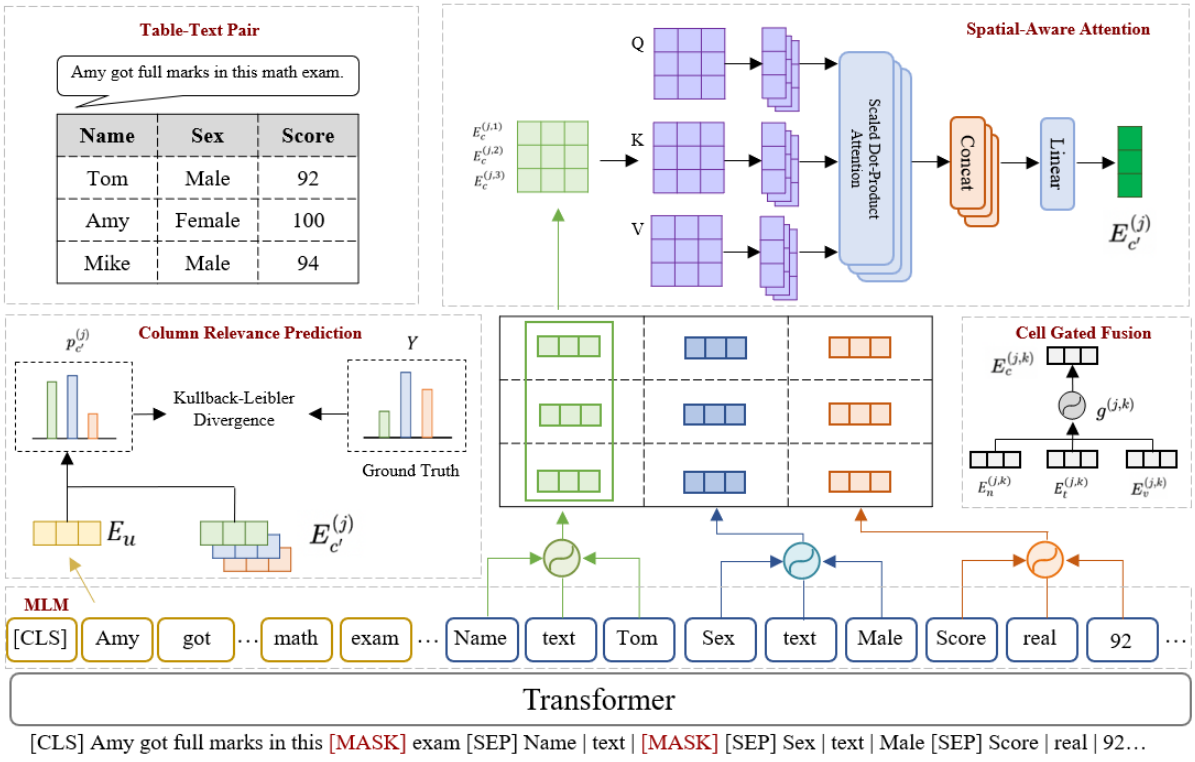

Figure 4: Model Architecture. (1) The model linearizes NL text and table into a sequence of tokens as input, and the gated fusion and row-aware attention are leveraged to get column representation (row representation can be obtained in the same way)). (2) The pre-training tasks includes MLM, Column Relevance Prediction (optimised with KL loss), and Row Relevance Prediction.

Spatial-Aware Representation. Inspired by self-attention (Vaswani et al., 2017), there are two kinds of attention mechanism in our framework, column-aware attention and row-aware attention, which attend each gated cell representation of a table both horizontally and vertically, namely spatial-aware attention. Assume that the number of columns and rows are C and M in a table, where $j \in \mathbb{R}^C, k \in \mathbb{R}^M$, the row-aware attention can be defined as:

$$
\begin{aligned}
e^{jk} &= (W_Q E_c^{jk})(W_K E_c^{jk})^T / \sqrt{d}, \\
a^{jk} &= \frac{\exp(e^{jk})}{\sum_{k=1}^{M} \exp(e^{jk})}, \\
E_{c'}^{j} &= \sum_{k=1}^{M} a^{jk} W_V E^{jk},
\end{aligned}
\tag{3}
$$

where W_Q,W_K,W_V are weight matrices for row-aware attention and d is dimension of W_Q. We perform row-aware attention here to obtain the column representation $E_{c'}^{j}$. Similarly, the column-aware attention can be defined in the same way to obtain the row representation $E_{r'}^{k}$.

Final Output. The output of FewTPT includes four parts: the representation of [CLS] token (E_{cls}), the representations for the tokens of NL text (E_u), and the interactive representations for row ($E_{r'}$) and column ($E_{c'}$) respectively.

3.2 Pre-Training Objectives

We implement three pre-training tasks, including Masked Language Modeling (MLM) and its variants to learn contextual representations for tokens in NL and table, Column Relevance Prediction (CRP) and Row Relevance Prediction (RRP) to capture the semantic interactions between the column/row and the give NL text respectively.

Masked Language Modeling. MLM is a widely-used objective for pre-training, which encourages the model to capture the contextual information of given sequence. Previous works also employ the variants of MLM for table pre-training (Herzig et al., 2020; Deng et al., 2020). Inspired by this, we devise three kinds of MLM tasks. Firstly, for tokens in NL text, we randomly mask 20% of individual sub-tokens and then recover the masked ones. Secondly, for columns names in the table, we randomly select 15% column names of the input table and require the model to predict them with column type and cell value information surrounded. Thirdly, considering cell values are basic units to record content in a table, we mask 15% of

cell values as well. Following Wang et al. (2021), we randomly select out cell strings from the table as candidates, then at each blanked position, we encourage the model to retrieve its corresponding string. We formulate all MLMs as multi-class classification and use cross entropy for training.

Row Relevance Prediction. The goal for row relevance prediction (RRP) is to predict whether a row is mentioned in the NL or not. As our table pre-training corpus provides the relevance scores X^k between NL text and each row as auxiliary supervised signals, we encourage the model to predict the relevance score for each row. For each row representation $E^k_{r'}$, we apply a fully connected layer with *sigmoid* activation function σ to obtain the probability of whether the row is mentioned in the utterance or not. Then Mean Squared Error (MSE) loss is leveraged as training objective:

$$\begin{aligned} p^k_{r'} &= \sigma(W_{r'} E^k_{r'} + E_u), \\ \mathcal{L}_{r'} &= \sum\nolimits_{k=1}^{M} ||p^k_{r'} - X^k||^2 / M. \end{aligned} \tag{4}$$

Column Relevance Prediction. Similar to RPP, given the column representation $E^j_{c'}$, the goal of column relevance prediction (CRP) is to predict the relevance score between each column and NL text. Since we also have the auxiliary supervised relevance score for each column, the predicted relevance score distribution is forced to fit that of supervised signals between text and columns:

$$p^j_{c'} = \text{softmax}(W_{c'} E^j_{c'} + E_u), \tag{5}$$

where j represents the index of different column. Thus, we take the Kullback-Leibler Divergence (KL) (Kullback and Leibler, 1951) as training objective to optimize the column relevance prediction:

$$\mathcal{L}_{c'} = \sum\nolimits_{j=1}^{C} p^j_{c'} \log \frac{p^j_{c'}}{Y^j}, \tag{6}$$

where Y^j is column-utterance relevance score obtained from the training corpus and $p^j_{c'}$ is the predicted score from our model.

4 Experiments

Here, we first introduce the experimental setup of pre-training and the comparable baselines, then we analyze the main results and also conduct ablation study to show the contribution of different modules.

4.1 Experimental Setup

Pre-Training Configuration. For pre-training setup, we train **FewTPT** with TPC-1M for 6 epochs, with the batch size of 4 on 4 Tesla V100 GPUs. Specifically, we set the learning rate as 4e-5, the number of attention heads as 12, and the weight decay and dropout are set as 0.01 and 0.1 respectively.

Comparable Models. For few-shot table understanding tasks, we adopt three strong baselines for comparison: (1) **BERT** (Devlin et al., 2018), which is the popular general-purpose PLM trained with free-form text. We linearize the table and feed the concatenated query and table into the official Chinese version[6] for fine-tuning and inference. (2) **TaBERT** (Yin et al., 2020), which is a recently proposed table pre-training model designed to jointly learn representations for NL text and (semi-)structured tables. Since the released model is in English version, we pre-train the model with TPC-1M corpus from scratch with the official source code[7]. (3) **SDCUP** (Hui et al., 2021), which proposes a schema dependency pre-training objective to impose the desired inductive bias into the learned representations for table pre-training. We fine-tune the official released model[8] directly for comparison.

N-shot	1	2	3	4	5	10
Table QA						
BERT	14.0	26.4	33.9	41.6	43.4	58.4
TaBERT	14.6	28.6	34.4	44.0	44.6	59.0
FewTPT	**15.8**	**30.6**	**35.8**	**45.4**	**46.2**	**60.6**
Schema Detection						
BERT	14.1	34.3	42.5	49.9	55.1	73.3
TaBERT	16.3	36.7	44.5	51.8	55.4	73.9
FewTPT	**16.9**	**37.5**	**44.7**	**52.2**	**56.1**	**74.2**
Table Selection						
BERT	83.2	82.4	85.1	87.5	86.2	85.6
SDCUP	81.6	82.4	84.3	84.9	85.5	86.6
TaBERT	83.8	85.6	86.7	87.9	88.6	89.5
FewTPT	**86.2**	**88.3**	**88.7**	**89.0**	**89.5**	**91.2**

Table 3: Few-shot table understanding performance on the tasks of Table QA, Schema Detection, and Table Selection.

[6]https://github.com/google-research/bert

[7]https://github.com/facebookresearch/TaBERT

[8]https://github.com/alibaba/AliceMind/tree/main/SDCUP. Considering SDCUP is tailored for the NL2SQL task, we only compare with it on two tasks to avoid modifying its model architecture too much.

4.2 Main Results

Experimental results are illustrated in Table 3 and Figure 5. For Table QA, Schema Detection and Table Selection, we evaluate all the models by ranging training samples N from {1,2,3,4,5,10}. It's observed that, (1) with the increase of N, the performance of all the models improves rapidly, indicating the size of training sample plays a vital role to all tasks. The results also demonstrate how many training samples are needed at least for each task to reach an *acceptable* performance, which is a critical issue but ignored by previous works. We argue that our experiments can be a valuable reference for table understanding applications in real-world scenario. (2) Compared to all baselines, our proposed model **FewTPT** yields substantial gains on all three tasks, and the benefits are more pronounced when N is small. It demonstrates the advantages of our proposed table pre-training method and the contribution of TPC-1M corpus. (3) **FewTPT** surpasses **TaBERT** consistently. Considering both models were pre-trained with the same corpus, it reveals the gains are from our model structure and pre-training objectives, which can fuse the information from NL text and table seamlessly and finally facilitate the downstream table understanding tasks.

Figure 5 demonstrates the experimental results on Table Fact Verification and Table Classification tasks in a more direct way. We observe the similar trend that **FewTPT** outperforms all baselines by a large margin. Especially, **FewTPT** progressively outperforms **TaBERT** by nearly **12.3%** ($N = 8$) on Table Fact Verification. We conjecture it's because this task heavily relies on the deep interactions between NL text and table content to discriminate whether the statement is correct or not, thus the proposed spatial-aware attention and row/column relevance prediction objectives are genuinely required and can benefit the table understanding. For Table Classification, the performances of **FewTPT** and **TaBERT** are comparable. Considering there is no query in the input, we guess the advantage of deep interactions between text and table in our model may be weakened.

4.3 Ablation Study

We conduct further experiments to figure out the contribution of each component, including the improved Masked Language Modeling (MLM) objective, Column Relevance Prediction (CRP), Row Relevance Prediction (RRP), and Cell Gated Fusion (CGF) mentioned in Section 3. Due to space limitation, we perform ablation study on Table QA and Schema Detection, under 2-shot setting. Figure 6 presents the experimental results by adding each component in a cumulative way, demonstrating the necessity of each component in our method.

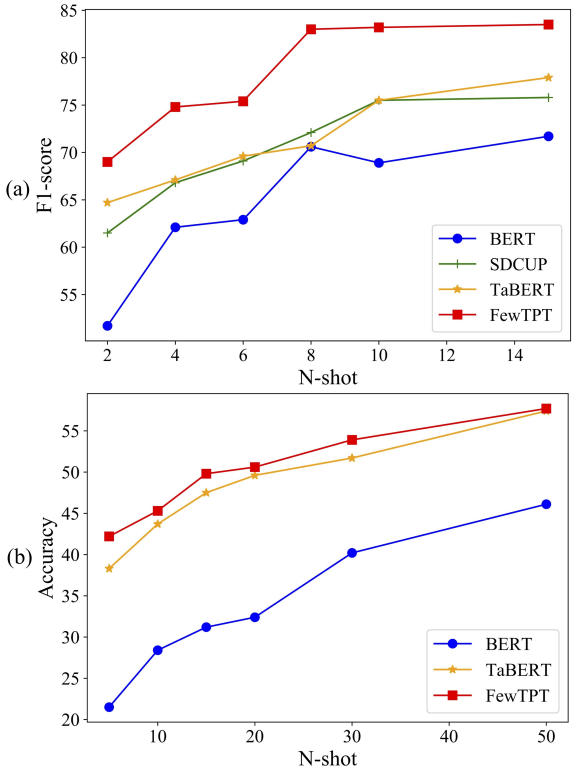

Figure 5: Few-shot model performance on the tasks of Table Fact Verification (a) and Table Classification (b).

5 Related Work

Large-Scale Tabular Corpus. Most of the recent table-text pre-training tasks utilize web-crawled tables and context from English Wikipedia (Lehmberg et al., 2016). Apart from the extreme noise contained in the tables, the context is mainly mined based on position information and may not be semantically related to tables. Some table-to-text generation task (Lewis et al., 2020) are also proposed to provide well-controlled text-table corpus with either grammar-supported data generation or powerful generative models (Nan et al., 2021; Liu et al., 2018; Parikh et al., 2020). However, the linguistic diversity of the generated data is limited. Besides, most of the current studies focus on English. Considering the increasing demand of table understanding tasks in Chinese, the deficiency of large-scale table-text corpus blocks the pre-training exploration on Chinese structured data. Therefore,

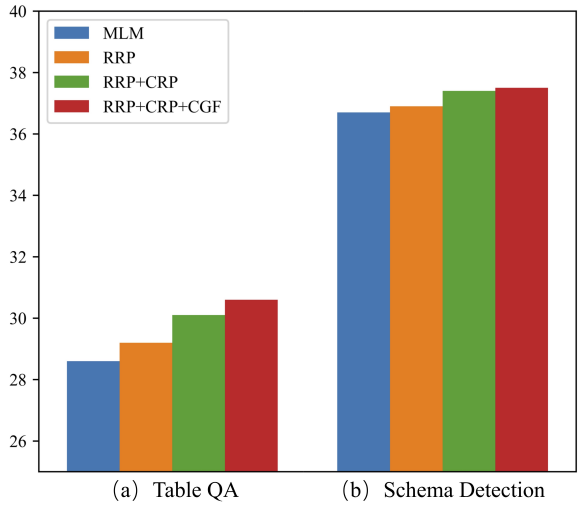

Figure 6: Ablation study of our method on Table QA and Schema Detection.

in this work, we collect a large-scale table pre-training corpus with more than 1 million tables and will release to the community soon.

Table Pre-Training. Inspired by the recent success of leveraging PLMs on tasks with huge amount of unstructured natural language (Devlin et al., 2018; Liu et al., 2019), some researches have tried to apply pre-training approaches into structured tabular data. TAPAS (Herzig et al., 2020) and TaBERT (Yin et al., 2020) introduce novel pre-training methods to learn the joint representations of table and text with a large-scale web crawled tables and their contextual natural language descriptions. The vanilla Masked Language Modeling (MLM) is adopted by either masking the tokens from input text or tokens from tables. To cultivate the alignment between utterances and table context, some researchers introduce supervised pre-training objectives by involving the logic language interaction such as SQL semantic prediction (Yu et al., 2021), SQL generation (Shi et al., 2021) and SQL execution (Liu et al., 2021). However, to ensure the quality of synthesised training data, they either design complicated generation templates manually or rely on pre-trained generative models. Whereas these generated training data is lack of variety compared with the natural utterances in real-world scenario of table-text understanding. Besides, some other researchers focus on table encoding with relational and complex structures (Iida et al., 2021; Deng et al., 2021). However, most of the unsupervised training objectives involved in above researches neglect enhancing the semantic interactions between natural languages and tables (Shi et al., 2022; Cheng et al., 2021; Dong et al., 2022). Differently, we propose a novel table pre-training method equipped by cell gated fusion, spatial-aware attention, masked language modeling, row/column relevance prediction to catch the deep semantic interactions between NL text and table.

Few-Shot Learning. Recently, lots of researches have focused on few-shot learning in NLP and propose a variety of methods including meta-learning (Kaiser et al., 2017), embedding learning (Bertinetto et al., 2016), memory-based learning (Kaiser et al., 2017). Whereas the recent flouring of PLMs (Brown et al., 2020) achieve remarkable few-shot performance solely by leveraging a natural-language prompt and a few task demonstrations as input context (Gao et al., 2021). However, few of them have paid attention to the challenge of few-shot table understanding or table-text interactions tasks where only few text-table pairs are available. Chang et al. (2020); Chen et al. (2021) explore the zero-shot text-to-sql task, and both of them illustrate the importance of leveraging the abundant table cell information and header information during training to improve table-text semantic relevance. To the best of our knowledge, there is no well-designed benchmark dataset and baseline for the few-shot table understanding. In this work, we are dedicated to fill the gap to facilitate future research.

6 Conclusion

In this paper, we focus on the few-shot table understanding problem and establish a benchmark dataset with five downstream tasks including Table Fact Verification, Table QA, Table Selection, Schema Detection, and Table Classification. We also contribute a large-scale Chinese tabular corpus which covers 1 million tables across 13 domains and the semantically-associated NL text. Finally, we provide a table pre-training method and conduct extensive experiments on the few-shot table understanding benchmark to set up the baselines. Experimental results demonstrate that catching the interactions between text and tables can improve the downstream tasks significantly. We hope the benchmark, tabular corpus, and the baselines can facilitate the future research on this field. In the future, we will explore more table structure friendly objectives to improve the pre-training.

References

Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. 2016. Learning feed-forward one-shot learners. In *Advances in neural information processing systems*, pages 523–531.

Ben Bogin, Jonathan Berant, and Matt Gardner. 2019. Representing schema structure with graph neural networks for text-to-SQL parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4560–4565, Florence, Italy. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Shuaichen Chang, Pengfei Liu, Yun Tang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Zero-shot text-to-sql learning with auxiliary task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7488–7495.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yongrui Chen, Xinnan Guo, Chaojie Wang, Jian Qiu, Guilin Qi, Meng Wang, and Huiying Li. 2021. Leveraging table content for zero-shot text-to-sql with meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3992–4000.

Zhoujun Cheng, Haoyu Dong, Fan Cheng, Ran Jia, Pengfei Wu, Shi Han, and Dongmei Zhang. 2021. Fortap: Using formulae for numerical-reasoning-aware table pretraining. *arXiv preprint arXiv:2109.07323*.

Xiang Deng, Ahmed H. Awadallah, Chris Meek, Alex Polozov, Huan Sun, and Matthew Richardson. 2021. Structure-grounded pretraining for text-to-sql. In *NAACL 2021*.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2020. Turl: table understanding through representation learning. *Proceedings of the VLDB Endowment*, 14(3):307–319.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haoyu Dong, Zhoujun Cheng, Xinyi He, Mengyu Zhou, Anda Zhou, Fan Zhou, Ao Liu, Shi Han, and Dongmei Zhang. 2022. Table pretraining: A survey on model architectures, pretraining objectives, and downstream tasks. *arXiv preprint arXiv:2201.09745*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *QA@ACL*.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Binyuan Hui, Xiang Shi, Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2021. Improving text-to-sql with schema dependency learning. *arXiv preprint arXiv:2103.04399*.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. TABBIE: Pretrained representations of tabular data. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456.

Edwin T Jaynes. 1982. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952.

Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. 2017. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*.

M. A. Khalid, V. Jijkoun, and MD Rijke. 2007. Machine learning for question answering from tabular data. In *International Conference on Database and Expert Systems Applications*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. 2016. A large public corpus of web tables containing time and context metadata.

Proceedings of the 25th International Conference Companion on World Wide Web.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Qian Liu, Bei Chen, Jiaqi Guo, Zeqi Lin, and Jianguang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL Conference*.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Bowen Qin, Lihan Wang, Binyuan Hui, Ruiying Geng, Zheng Cao, Min Yang, Jian Sun, and Yongbin Li. 2021. Sdcup: Schema dependency-enhanced curriculum pre-training for table semantic parsing. *arXiv preprint arXiv:2111.09486*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Peng Shi, Patrick Ng, Feng Nan, Henghui Zhu, Jun Wang, Jiarong Jiang, Alexander Hanbo Li, Rishav Chakravarti, Donald Weidner, Bing Xiang, and Zhiguo Wang. 2022. Generation-focused table-based intermediate pre-training for free-form question answering. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*.

Peng Shi, Patrick Ng, Zhiguo Wang, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Cicero Nogueira dos Santos, and Bing Xiang. 2021. Learning contextual representations for semantic parsing with generation-augmented pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15).

Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*. ACM.

Ningyuan Sun, Xuefeng Yang, and Yunfeng Liu. 2020. Tableqa: a large-scale chinese text-to-sql dataset for table-aware sql generation. *arXiv preprint arXiv:2006.06434*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, and Xiaochuan Wang. 2020a. Reco: A large scale chinese reading comprehension dataset on opinion. In *AAAI*.

Jingjing Wang, Haixun Wang, Zhongyuan Wang, and Kenny Qili Zhu. 2012. Understanding tables on the web. In *Conceptual Modeling - 31st International Conference ER 2012, Florence, Italy, October 15-18, 2012. Proceedings*, volume 7532 of *Lecture Notes in Computer Science*, pages 141–155. Springer.

Lijie Wang, Ao Zhang, Kun Wu, Ke Sun, Zhenghua Li, Hua Wu, Min Zhang, and Haifeng Wang. 2020b. DuSQL: A large-scale and pragmatic Chinese text-to-SQL dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6923–6935, Online. Association for Computational Linguistics.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021. Tuta: Tree-based transformers for generally structured table pre-training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1780–1790.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, bailin wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, richard socher, and Caiming Xiong. 2021. Gra{pp}a: Grammar-augmented pre-training for table semantic parsing. In *International Conference on Learning Representations*.

Hongzhi Zhang, Yingyao Wang, Sirui Wang, Xuezhi Cao, Fuzheng Zhang, and Zhongyuan Wang. 2020. Table fact verification with structure-aware transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1624–1629. Association for Computational Linguistics.

Shuo Zhang and Krisztian Balog. 2018. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1553–1562. ACM.