Tab-CQA: A Tabular Conversational Question Answering Dataset on Financial Reports

Chuang Liu[1], **Junzhuo Li**[2], and **Deyi Xiong**[1,2] *
[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] School of New Media and Communication, Tianjin University, Tianjin, China
{liuc_09,jzli,dyxiong}@tju.edu.cn

Abstract

Existing conversational question answering (CQA) datasets have been usually constructed from unstructured texts in English. In this paper, we propose Tab-CQA, a tabular CQA dataset created from Chinese financial reports that are extracted from listed companies in a wide range of different sectors in the past 30 years. From these reports, we select 2,463 tables, and manually generate 2,463 conversations with 35,494 QA pairs. Additionally, we select 4,578 tables, from which 4,578 conversations with 73,595 QA pairs are automatically created via a template-based method. With the manually- and automatically-generated conversations, Tab-CQA contains answerable and unanswerable questions. For the answerable questions, we further diversify them to cover a wide range of skills, e.g., table retrieval, fact checking, numerical reasoning, so as to accommodate real-world scenarios. We further propose two different tabular CQA models, a text-based model and an operation-based model, and evaluate them on Tab-CQA. Experiment results show that Tab-CQA is a very challenging dataset, where a huge performance gap exists between human and neural models. In order to promote further research on Chinese tabular CQA, we release the dataset as a benchmark testbed at https://github.com/tjunlp-lab/Tab-CQA.

1 Introduction

Conversational question answering (CQA) extends traditional question answering to a conversational scenario where questions and answers are usually related to conversation history. Recent years have witnessed an upsurge of interest in both dataset building for CQA and models/approaches to CQA. Previous CQA datasets have been constructed either for text span extracting tasks (Choi et al., 2018; Reddy et al., 2019; Campos et al., 2020; Saeidi et al., 2018) or SQL-style table search tasks (Iyyer et al., 2017; Yu et al., 2019b). All these datasets are in English. Public CQA datasets in other languages are either rare or not available at all, which is of course not desirable for an inclusive CQA study across different classes of languages (Joshi et al., 2020).

In addition to the language dimension in diversifying CQA tasks and datasets, data source is yet another important factor. Existing CQA datasets are usually constructed from unstructured texts. Structured or semi-structured data, like tables, are also important sources for information gathering. Furthermore, structured tables exhibit reasoning skills (e.g., more numeric reasoning) with a different distribution from those on unstructured texts.

Inspired by the aforementioned diversity in both languages and data sources for CQA, in this paper, we propose Tab-CQA, a large-scale tabular conversational question answering dataset built from Chinese financial reports. It contains 2,463 manually-generated conversations and 4,578 automatically-generated conversations, with a total of 109,089 question-answer pairs. Each dialogue consists of multiple rounds of questions and answers, which are either manually created by crowdsourced workers playing roles of students and teachers or automatically created by a template-based method.

For manually-created dialogues, the student reads a partially masked table and asks a series of questions. We train students to ask questions as naturally as possible, preserving the characteristics of natural conversations, such as ellipsis and coreference in dialogue. The teacher provides answers to the questions from the student by carefully checking the given completely visible table.

In addition to question-answer pair collection in a conversational way, we also provide annotations to manually created dialogues. The teacher annotates an answer type for each provided answer. Based on pre-annotation and analysis on financial

*Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 196–207

Dataset	# Conversations	# Questions	# Avg.Turns	Tabular	Natural language questions	Numerical reasoning	Chinese
CQA	8,399	127,000	15.2	X	✓	X	X
QuAC	13,594	98,407	7.2	X	✓	X	X
DoQA	2,437	10,917	4.48	X	✓	X	X
SQA	6,066	17,553	2.9	✓	✓	✓	X
CoSQL	2,164	15,598	5.2	✓	X	X	X
HybridQA	-	70,153	-	✓	✓	✓	X
OTT-QA	-	45,841	-	✓	✓	✓	X
TAT-QA	-	16, 552	-	✓	✓	✓	X
FinQA	-	8,281	-	✓	✓	✓	X
DROP	-	96,567	-	X	✓	✓	X
CMRC2017	-	364,295	-	X	✓	X	✓
CMRC2018	-	19,071	-	X	✓	X	✓
DuReader	-	200,000	-	X	✓	X	✓
Tab-CQA	7,041	**109,089**	**14.6**	✓	✓	✓	✓

Table 1: Tab-CQA in comparison to other relevant datasets.

tables, we divide answer types into three categories: table retrieval, fact checking and computation. In addition to the answer type annotation, the teacher also needs to provide conversation flow tags to control the flow of conversation, *i.e., good, ok, unallowable*. A tabular conversation question answering example from Tab-CQA is in the Appendix A.

All questions in Tab-CQA require reasoning across the given table and dialogue history. Even for table retrieval questions, they are created in the way that is more difficult than just span extraction. The dataset also contains unanswerable questions.

In addition to manually created QA pairs, we automatically generate 73,595 QA pairs based on predefined templates over tables. Unlike manually labeled QA pairs, automatically-generated pairs are tagged with special labels. We then propose two different CQA models: a text-based model and an operation-based model. The text-based model is to convert the table into a passage by a multi-type network for different types of answers in Tab-CQA. The operation-based model converts the table into triplets to facilitate numeric reasoning.

The contributions of the work are as follows.

- We propose Tab-CQA to diversify existing CQA datasets in language, data source and task definition. To the best of our knowledge, Tab-CQA is the first conversational question answering dataset in Chinese. And unlike other Chinese QA datasets, it focuses on understanding tables in financial reports.

- We introduce a method to build tabular conversational QA datasets, where students cannot see entire tables and ask questions that require high cognitive skills to answer, e.g., logical and numerical reasoning skill.

- We use Tab-CQA as a bechmark dataset to test two different methods depending on the form of table representation and provide a systematic error analysis for the best method. The dataset will be publicly available soon.

2 Related Work

Tab-CQA is related to datasets for text-based conversational question answering and tabular question answering. It is also partially related to datasets on numerical reasoning and machine reading comprehension in Chinese. The comparison of Tab-CQA to other related datasets is shown in Table 1.

Text-based conversational question answering datasets. Reddy et al. (2019) propose a CQA dataset CoQA. It contains 8K conversations with 127K question-answer pairs. The dataset selects passages from several domains, such as children's stories, news, and science. Answers of CoQA are mostly a short fragment or an entity. Passages are visible to both questioners and responders in CoQA.

Choi et al. (2018) present a CQA dataset QuAC that focuses mainly on information seeking. It contains 14K conversations on passages selected from Wikipedia. Unlike CoQA, only the responder can see complete passages in QuAC while questioner can only see the titles of passages. Due to this setting, QuAC may contain unanswerable questions. Hence, annotators not only annotate the answerability of questions, but also provide dialogue actions to control the flow of dialogue. Paritially inspired by this, we present only header rows/columns of extracted tables to the student in building our dataset.

Campos et al. (2020) build a domain-specific CQA dataset DoQA. The dataset contains 2.4K conversations, with 10.9K question-answer pairs.

DoQA also includes question-answer pairs in the information retrieval scenarios.

All these CQA datasets are different from Tab-CQA in that they create QA-style conversations on unstructured texts written in English.

Tabular and database-based question answering datasets. Database queries are often relatively complex. Hence, Iyyer et al. (2017) propose a SQL-style CQA dataset SQA to decompose complex queries into several simple questions, so that there is a contextual relationship between them. Their dataset uses Wikipedia tables to create 6K sequences of questions, where complex questions are decomposed into several simple questions.

Yu et al. (2019a) create a cross-domain corpus based on a conversational query system CoSQL. It collects 3K+ conversations from 200 databases covering 138 domains, containing 30K+ rounds of conversations and 10K+ annotated SQL queries. CoSQL is significantly different from ours in that it collects SQL-style queries rather than natural language questions.

Chen et al. (2020c) build a hybrid text- and table-based QA dataset HybridQA. Each question is aligned to a structured Wikipedia table and entities in the table are linked to free texts. The dataset contains 70K question-answer pairs and 13K tables, each of which is associated with an average of 44 paragraphs.

Chen et al. (2020a) present an open domain QA dataset OTT-QA built on the base of HybridQA. They re-annotate 45K questions, which require multi-step reasoning, aggregating information from tables and texts.

Zhu et al. (2021) and Chen et al. (2021) propose hybrid QA datasets, also focusing on answering questions over financial data. The two datasets contain 16,552 and 8,281 question-answer pairs, respectively. Despite the similarity to Tab-CQA in financial QA, the two datasets are in English and not in a conversational format.

Yet another dataset related to Tab-CQA is TABFACT (Chen et al., 2020b), which is not a QA dataset. The dataset focuses on table-based fact detection. Inferences are regarded positive if they match corresponding table descriptions.

All the above datasets are in English and questions/queries in these datasets are either SQL-style or uncontextually linked.

Datasets on numerical reasoning. Dua et al. (2019) propose a QA dataset DROP with numerical inference-type questions. As numbers provide important supporting information for financial statements, we create QA pairs involving numerical reasoning.

Chinese machine reading comprehension datasets. Inspired by the well-known machine reading comprehension (MRC) dataset SQuAD (Rajpurkar et al., 2016), several Chinese MRC datasets have been also proposed (Cui et al., 2016, 2018, 2019b). He et al. (2018) build a large-scale open domain QA dataset, which annotates 200K queries from search engines. Jing et al. (2019) present a bilingual MRC dataset where parallel Chinese and English texts, questions and answers are provided. Sun et al. (2020) propose a free-form multiple-Choice Chinese machine reading Comprehension dataset C3. Unfortunately, none of these Chinese MRC datasets are in a conversational setting.

3 Dataset Creation

This section elaborates how Tab-CQA is created, including details on table extraction, conversation collection and annotation.

3.1 Table Extraction

We have collected nearly 30 years of annual financial reports of Chinese companies, listed in the major segments of the Shenzhen Stock Exchange and Shanghai Stock Exchange, covering 18 industry sectors, e.g., business, trade, power, retail, real estate and so on. First, for each year, we randomly select one company from each industry sector. Second, from each selected company, we randomly choose a financial report of that company. In this way, we have collected 6,661 reports for table extraction. As all reports are in PDF formats, we use the table extraction tool PDFlux[1]. Only tables where the number of cells is large than 15 and blank cells account for less than 30% of all cells are kept. Finally, we have extracted 7,041 tables.

3.2 Conversation Collection

Manually-Generated Conversations. Once tables are extracted, we develop a conversation collection and annotation tool to collect a conversation and required annotations for each extracted table. We have 28 crowdsourced workers who can alternatively play as either a student to ask questions or a

[1] http://pdflux.com/

teacher to provide answers. For each conversation, once they choose their roles, the chosen roles will be fixed until the conversation is completed. We train all crowdsourced workers in a pre-annotation phase. Only when they are quite familiar with the data collection protocol, they are allowed to participate in the formal conversation collection stage. Half of the crowdsourced workers have financial background while the other half do not have. Therefore, our collected conversations are mixed with financially professional and nonprofessional utterances.

The collection tool has different user interfaces for the student and teacher. For the student, only the header rows, header columns and randomly selected cells of a given table are visible to him/her. Hence the student needs to ask questions step by step to understand the masked table. We encourage the student not to ask questions easily searchable from a given table. A variety of types of questions, e.g., table retrieval, muti-step reasoning, computation, numerical comparison, can be used by the student to help himself/herself to have a clearer understanding of the masked table in a conversation setting.

The teacher is able to see the entire given table. Therefore the teacher needs to first judge whether a question raised by a "partially blind" questioner is answerable according to the information in the given table. Additionally, the teacher is also required to provide annotations of answer type and dialogue action to control the flow of a conversation, which will be introduced in the next subsection.

In this way, we have obtained 2,463 conversations with 35,494 question-answer pairs.

Automatically-Generated Conversations. We use a template-based method to automatically generate QA pairs. Specifically, we define 9 operation templates over extracted tables, as show in Appendix B. Then we randomly select an operation, perform it on a set of triplets extracted from tables. Each triplet consists of the cell value from the table and its row and column names. We define the triplet as $< Row_i, Column_i, Cell_i >$, where i is the index of arguments in predefined templates (i.e., $i = 1$ or 2). We randomly selected 100 manually-generated conversations. We then selected operations that occur more than 8 times as template operations. In total, the selected operations account for 84% of the selected samples. More details are in Appendix B.

Statistics	Train	Dev	Test
Table	6548	247	246
Avg.T Tokens	771.24	770.45	761.00
Question/Answer	101,884	3,601	3,604
Avg.Q Tokens	19.58	11.52	11.55
Avg.Turns	15.56	14.58	14.58

Table 2: Overall statistics of manual annotation of TabCQA. T: table. Q: question.

In the end, we have automatically generated 4,578 conversations with 73,595 question-answer pairs.

3.3 Conversation Annotation

In order to have a deep understanding on the nature of collected answers and questions and a good control of conversation flow that allows the student and teacher to focus on a given table, our collection tool requires the teacher to do two types of conversation annotation on the teacher side. Appendix C provides details on how we control quality and diversity.

Answer Type Annotation. As not all information of a given table is visible to the student, we do not ask the student to annotate the type of questions. On the contrary, the teacher can see the complete table. Hence the teacher knows what should be given as an answer and how the answer should be found. According to the nature of answers from extracted financial tables, we roughly divide them into three types: table retrieval, fact checking and computation. For table retrieval, answers can be found directly from a given table via simple reasoning. For fact checking, answers are yes or no according to the facts in the given table. For computation, answers are not directly from the given table, but they can be obtained by numerical reasoning over numbers in the given table, such as numerical comparison (e.g., finding the maximum, minimum, larger, smaller numbers from the table), arithmetic operations, and so on. The teacher is asked to annotate each answer with one of these three answer types. In addition, if there is no answer, the teacher needs to annotate "unanswerable".

For automatically-generated questions, we annotate answer types according to Table 5, and if there is no cell value in the selected triplet, the answer type is "unanswerable".

Conversation Flow Control Annotation. In ad-

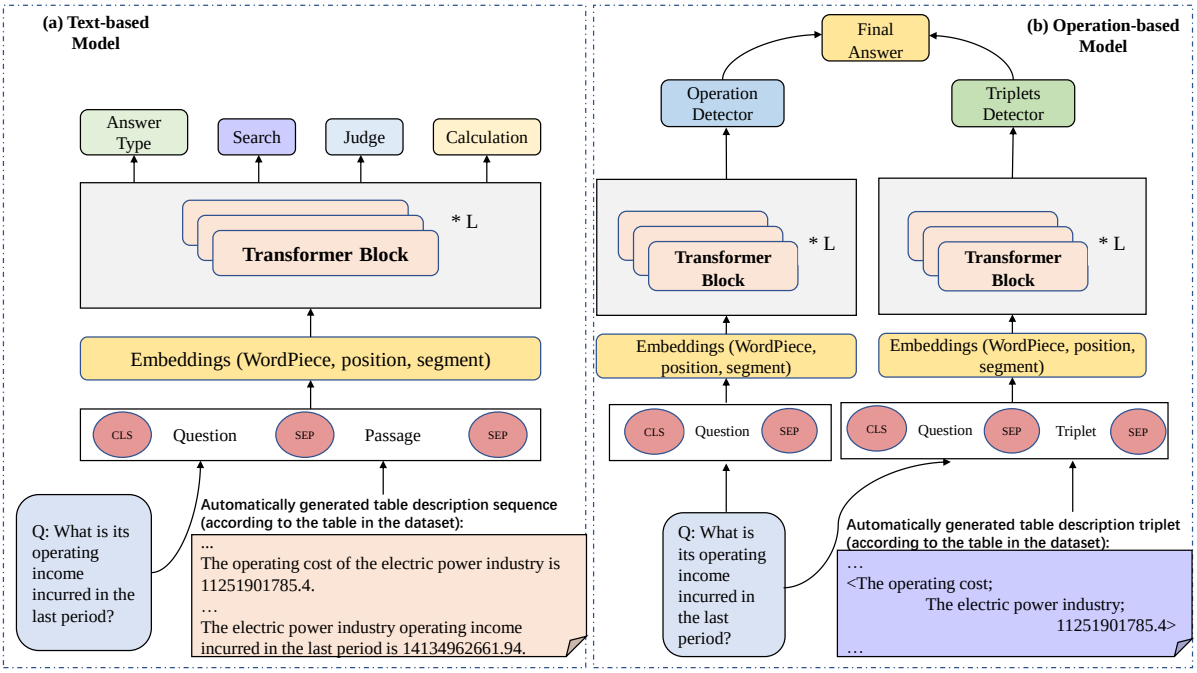

Figure 1: The text-based (a) and operation-based (b) neural models for Tab-CQA.

dition to providing answers and annotating answer types, the teacher is also in charge of conversation flow control (i.e., guiding the student to ask questions relevant to the given table). According to the relatedness of questions to the given table, the teacher will annotate each answer with a flow control tag: "good", "ok" or "unallowable". "good" suggests that the topic of the question in current conversation turn is related to the given table and questions from the same topic can be continued in future conversation turns. "ok" indicates that the current topic is ok but encouraged to be changed in future conversation turns. "unallowable" means that the current topic is not related to the given table and should be changed immediately. All the three flow control tags will be present to the student so that the student can ask appropriate questions in future conversation.

We do not perform conversation flow control annotation during the automatically-generated process.

3.4 Overall Statistics

Table 2 shows the overall statistics of Tab-CQA. From financial reports collected from Chinese listed companies in the last three decades, we randomly select 75 companies from 18 domains. We have extracted 7,041 tables and collected 109,089 question-answer pairs. The average numbers of tokens in extracted tables and questions are 770.85 and 19.01, respectively. The average number of turns in collected conversations is 15.49. We divide manually-generated data into the training, development and test set in the proportion of 8:1:1. To ensure that the development and test set is close to the real scenario, the automatically-generated data are used as a supplement to the training set. Further analyses of Tab-CQA are displayed in Appendix D.

4 Models for Tabular CQA

We propose two different models, namely text-based and operation-based model, as shown in Figure 1. For the text-based model, we convert each table into a piece of text, so the task is converted into a reading comprehension form. However, in practice this approach is difficult to effectively model numerical reasoning that is pervasive in our dataset. Therefore, we further propose an operation-based neural model that represent a table as a series of triplets.

4.1 Text-based Model

Partially inspired by MTMSN (Hu et al., 2019), we modify the output type of the text-based model to *Table Retrieval*, *Fact Check* and *Computation*. We transform a table into a passage that is a table description sequence containing each cell in the table. We then concatenate the passage and question into an input sequence with

[CLS] and [SEP] as in BERT (Devlin et al., 2019). This concatenated sequence is processed by L pretrained Transformer blocks:

$$\mathbf{H}_i = \text{TransformerBlock}\left(\mathbf{H}_{i-1}\right), \forall i \in [1, L] \tag{1}$$

We then use the contextualized token representations as the input to predict the type of answer as:

$$\mathbf{p}^{\text{type}} = \text{softmax}\left(\text{FFN}\left(\mathbf{h}^{\text{CLS}}\right)\right) \tag{2}$$

For *Table Retrieval* questions, we calculate the probability of the beginning and ending positions of the answer fragment in the passage as:

$$\begin{aligned} \mathbf{p}^{\text{start}} &= \text{softmax}\left(\mathbf{W}^S \mathbf{H}_i^{\text{start}}\right), \\ \mathbf{p}^{\text{end}} &= \text{softmax}\left(\mathbf{W}^E \mathbf{H}_i^{\text{end}}\right) \end{aligned} \tag{3}$$

For *Fact Check* questions, we consider it as a binary classification problem and calculate whether the problem description is true or not:

$$\mathbf{p}^{\text{fact check}} = \text{softmax}\left(\mathbf{W}^F \mathbf{H}_{CLS}\right) \tag{4}$$

For *Computation* questions, we assign a computational sign to all numbers in the passage, i.e., $+$, $-$, 0. We then consider it as a ternary classification problem. For example, for a problem that requires summation, the numbers involved in the problem are assigned $+$, while other unrelated numbers are assigned 0:

$$\mathbf{p}_i^{\text{computation}} = \text{softmax}\left(\mathbf{W}^C \mathbf{H}_i\right) \tag{5}$$

$i \in [1, n]$, n is the number of numbers in the passage.

4.2 Operation-based Model

This model consists of two sub-units for triplet prediction and operation prediction, respectively. For the triplet prediction unit, we consider it as a binary classification problem. We use outputs from a pretrained LM to estimate probabilities for the triplet detector:

$$\mathbf{p}^{\text{triplet}} = \text{softmax}\left(\mathbf{W}^T \mathbf{H}_i\right) \tag{6}$$

We then take questions and predicted triplets as input to the operation prediction unit and predict the desired operation. We consider it as an N-ary classification problem, where N is 9, the number of operation templates that have been defined in Table 5. The predicted probability is:

$$\mathbf{p}^{\text{operation}} = \text{softmax}\left(\mathbf{W}^O \mathbf{H}_i\right) \tag{7}$$

Model	BERT	FIN	WWM
$Text_X$	9.17 / 8.58	9.13 / 8.52	9.22 / 8.61
$Text^*_X$	16.17 / 15.62	14.29 / 14.13	16.18 / 15.70
Op_X	16.78 / 18.01	16.83 / 16.36	16.06 / 16.78
Op^*_X	19.24 / 20.32	18.29 / 18.76	19.57 / 19.43

Table 3: F1 (%) results for all models, each cell shows dev/test scores. $Text_X$ denotes the text-based model with corresponding PLM X (i.e., BERT, FIN (FinBERT), WWM (BERT-wwm)) while Op_X indicates the operation-based model with PLM X. $Text^*$ indicates that the training set contains both manually-generated and automatically-generated data. Op^* indicates that the number of positive and negative training instances for the triplet prediction module is balanced.

The final answer is obtained based on the predicted triplets and operation together. For example, if the predicted triplet contains T_1 and T_2, and the operation is 4. It means to determine whether the value of T_1 is bigger than the value of T_2.

5 Experiments

5.1 Experimental Settings

For the text-based model, we set the max sequence length, maximum query length and maximum answer length to 384, 64 and 30, respectively. The batch size was set to 10. For the operation-based model, we set the max sequence length to 32. The batch size was set to 3. All the optimizers were Adam with a learning rate of 5e-5. The number of Transformer layers for all PLMs is 12. Appendix E provides details on baseline models.

5.2 Results

We used F1 (%) to evaluate the performance of different pretrained language models on our dataset. It should be noted that for the text-based approach, automatically-generated QA pairs can be used as additional data to the training set, while for the operation-based approach, only the automatically generated QA pairs can be used as training instances because manually labeled data lack the corresponding labels. Table 3 shows the experiment results of all models. The overall F1 of all these methods are much lower than those of neural models on other CQA datasets. This may be due to two reasons: pervasive numeric reasoning questions and 47.7% of questions involve either coreference or ellipsis, which make our Tab-CQA challenging for current neural models. The operation-based model is better than the text-based model as the

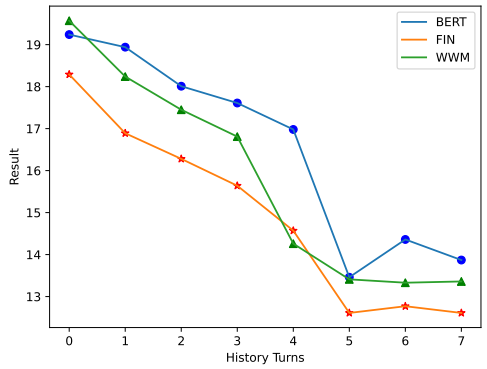

Figure 2: The results with the number of conversation histories about Op^*_X on the validation set.

Error Type	Percentage (%)
TPE	59.5
OPE	9.5
APE	8.1
OEOP	22.9

Table 4: Error analysis on the dev sets.

former is more suitable for numeric reasoning.

All Op^*_X models are better than Op_X models, suggesting that maintaining a balance of positive and negative samples is good for final performance since only a small fraction of cells in tables are related to questions in Tab-CQA.

5.3 The Impact of the Number of Dialogue Histories on Performance

Figure 2 shows the impact of conversation histories on model performance. It is important to note that the performance of the model does not increase with the number of conversation histories. A direct reason for this is the distance between the current question and the conversational history on which it relies in Tab-CQA. Valid contextual information is not available in the proximate conversation histories. When the number of conversation histories reaches a certain number, there is a slight performance increase followed by a decrease. This is because a certain number of conversation histories provide sufficient contextual information required for answering the current question. Additional conversation histories may bring noise to the model, we will investigate this issue further in the future.

5.4 Error analysis

We selected the best model (Op^*_{WWM}) on the development set to conduct an in-depth error analysis. We classify answer errors into four categories: Triplet Prediction Errors (TPE), Operation Prediction Errors (OPE), Insufficient Number of Cell Values (INCV), and Operation Error Outside of Predefined (OEOP). Specifically, TPE means that an irrelevant triplet is selected; OPE denotes that a wrong operation is predicted, e.g., an addition operation is predicted as a subtraction operation; INCV represents that correctly answering the question involves more triplets than our model setting, For example, to answer the question, "What is the number of jobs with the highest number of people in the company in 2010?", it is required to retrieve all relevant triplets and then compare them; and OEOP means that the actual operation is not predefined. For example, the correct answer to the question "Which year's operating income is more than 1 million" is the row name, even though the model has selected the correct triplet, but there is no correct operation to answer it correctly.

We randomly selected 100 QA pairs and manually check the results for each cell in Op^*_{WWM} according to the error type. Of these, 26 questions were answered correctly, and the remaining 36 questions with errors contained 44 TPEs, 7 OPEs, 6 INCVs and 17 OEOPs, as shown in Table 4.

6 Conclusions

In this paper, we have presented Tab-CQA, a tabular conversational question answering dataset built from tables randomly extracted from annual financial reports of Chinese listed companies over the past three decades in 18 industry sectors. The dataset contains 7,041 tables, of which 2,463 tables are equipped with a manually collected conversation generated by crowdsourced workers playing the roles of students and teachers, another 4,578 tables are automatically generated according to templates. We have collected 109,089 QA pairs , covering table retrieval, fact checking and computation, 47.7% of which are associated with coreference or ellipsis. We propose two models for Tab-CQA, and the experimental results indicate that the operation-based model is better than the text-based model.

Acknowledgements

The present research was partially supported by Zhejiang Lab (No. 2022KH0AB01) and Huawei. We would like to thank the anonymous reviewers for their insightful comments.

References

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and W. William Cohen. 2020a. Open question answering over tables and text.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and Yang William Wang. 2020b. Tabfact: A large-scale dataset for table-based fact verification. *ICLR*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020c. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019a. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019b. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. Dataset for the first evaluation on Chinese machine reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for Chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. 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 *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1596–1606, Hong Kong, China. Association for Computational Linguistics.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Yimin Jing, Deyi Xiong, and Zhen Yan. 2019. BiPaR: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2452–2462, Hong Kong, China. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter Lasecki, and Dragomir Radev. 2019a. CoSQL: A conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1962–1979, Hong Kong, China. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019b. SParC: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523, Florence, Italy. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. 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 3277–3287, Online. Association for Computational Linguistics.

IDX	Operation	Answer Type
1	Find()	Table Retrieval
2	Find(max(find(),find()))	Table Retrieval
3	Find(min(find(),find())	Table Retrieval
4	Bool(max(find(),find()))	Fact Check
5	Bool(min(find(),find()))	Fact Check
6	Bool(same(find(),find())	Fact Check
7	Bool(diff(find(),find())	Fact Check
8	Sum(find(),find())	Computation
9	Sub(find(),find())	Computation

Table 5: Predefined templates.

A An Example from Tab-CQA

Figure 3 is a tabular conversation question answering example from Tab-CQA. The upper part is a part of an extracted financial table while the bottom part shows a multi-round conversation with question-answer pairs on the table. Grey cells: invisible areas to students in data collection. Orange section: dialogue history. Blue section: current conversation turn. S/T denotes student/teacher. We also provide conversation flow tag and answer type (in brackets) to each answer. Better view in color.

B The Details of the Selection Procedure

We define a triplet as $< Row_i, Column_i, Cell_i >$, $i \in (1, 2)$. The automatic process of QA generation is illustrated Figure 4. We first feed the row and column names from the two triples into the entity slots in the template. Next, we select words from a predefined verb list to fill the verb slots in the template based on known operations. For example, for a $Fact\ Check$ type question, we will randomly select some comparison verbs, such as "smaller than". As shown in Figure 4, after filling the slots, a complete sentence is generated to check whether the fact "2011 operating income is less than 2012 operating income" is true.

To ensure that the questions contain contextual links to dialogue history, we make some transformations to questions in each set of dialogues. If the same entity occurs in the previous rounds of questions, it is replaced with a pronoun. If the content of both entity slots in the same triplet in the previous round of questions is the same, it is simply omitted. We create $Coreference$ and $Ellipsis$ in automatically generated QA pairs in this way.

The answer will be generated based on the $Cell$ value and specific operation will be selected when generating the question. A question is treated as unanswerable when the $Cell$ is empty.

Domains	Train	Dev	Test	Perc. (%)
Accommodation	3705	257	126	3.8
Agriculture	6666	191	386	6.6
Building	8619	498	490	8.8
Comprehensive	2199	160	174	2.3
Culture	4067	292	303	4.3
Education	1992	120	178	2.1
Electricity	6550	303	507	6.8
Environment	3441	170	146	3.5
Estate	5832	90	78	5.5
Finance	12637	263	315	12.1
Health	3419	120	73	3.3
Leasing	2434	15	-	2.2
Manufacturing	14289	477	409	13.9
Mining	5180	121	86	4.9
Science	3627	-	-	3.3
Software	4914	117	120	4.7
Transportation	5598	139	92	5.4
Wholesale	6715	268	121	6.5

Table 6: Statistics on domains.

9 operation templates are displayed in Table 5.

C The Details of Quality and Diversity Control

We use a strict quality control process to ensure the quality of Tab-CQA.

Before starting to annotate Tab-CQA, we trained 28 annotators to help them fully understand our annotation conventions and learn how to use our annotation system. Afterwards, we gave all annotators samples for pre-annotation and based on the pre-annotation results, we provided further explanations and training to ensure that the annotators could understand our goals.

For each annotation completed, we asked both QA parties to swap roles for validation, including checking whether conversations are reasonable in context, whether answers are consistent with the table, and whether calculations are correct. If any errors were found, the annotators were asked to make corrections. When all annotations were completed, we selected ten annotators with good performance to perform a second round of checking of data checking.

D Dataset Analysis

D.1 Table Distribution Over Domains

The distribution of extracted tables over these domains is displayed in the Table 6. The top 3 domains are manufacturing, finance and building.

行业名称 Industry Name	本期发生额 Amount in the Current Period		上期发生额 Amount in the Last Period	
	营业收入 Operating Income	营业成本 Operating Costs	营业收入 Operating Income	营业成本 Operating Costs
电力行业 Electric Power Industry	14252402452.91	11251901785.4	14134962661.94	11778548987.79
贸易行业 Trade Industry	6521248499.64	6479611128.88	6383848180.5	6365168039.51

S: 电力行业的营业成本是多少?
What are the operating costs of the power industry?
T: "11251901785.4" (*good*) (*Table Retrieval)*

S: 它的上期发生额的营业收入是多少?
What is its operating income incurred in the last period?
T: "14134962661.94" (*ok*) (*Table Retrieval)*

S: 娱乐行业的上期与本期发生额的营业成本之差是多少?
What is the difference between the operating costs incurred in the last period and the current period in the entertainment industry?
T: "unanswerable" (*unallowable) (Unanswerable)*

Figure 3: A tabular conversation question answering example from Tab-CQA.

D.2 Question Type Distribution

To further understand the types of questions and reasoning skills required to answer questions, we have randomly sampled 1000 questions from Tab-CQA for manual analysis. Table 7 shows the analysis results. The percentages of questions over the three types (i.e., table retrieval, fact checking, and computation) are the same as those of answers. For each type of questions, we further check if they are associated with discourse phenomena, such as co-reference (e.g., using pronouns to refer entities mentioned in previous conversation turns) and ellipsis (e.g., omitting entities from previous conversation turns). We calculate the percentages of discourse-related question types. In total, ordinary questions that are not contextually linked account for 52.3% while questions associated with coreference account for 15.3% and questions with ellipsis 32.4%.

E Settings

E.1 Baseline Models

BERT: As BERT (Devlin et al., 2019) can be used in both span extraction QA tasks (Devlin et al., 2019) and MCQ tasks (Sun et al., 2020), we used a Chinese BERT trained on Chinese texts as our first PLM.

FinBERT: FinBERT[2] is the first Chinese pre-trained language model trained on financial texts based on the BERT architecture. FinBERT has achieved significant improvements in several downstream tasks in the finance domain over baselines. Since our dataset extracts tables from financial reports, we chose FinBERT as another PLM.

BERT-wwm: BERT-wwm (Cui et al., 2019a) is a Chinese pre-trained model trained with full-word masking rather than subword masking. BERT-wwm uses a corpus from Chinese Wikipedia, which contains 24M sentences. The vocabulary size is set as 21,128. We used both BERT-wwm as our PLM too.

[2]https://github.com/valuesimplex/FinBERT

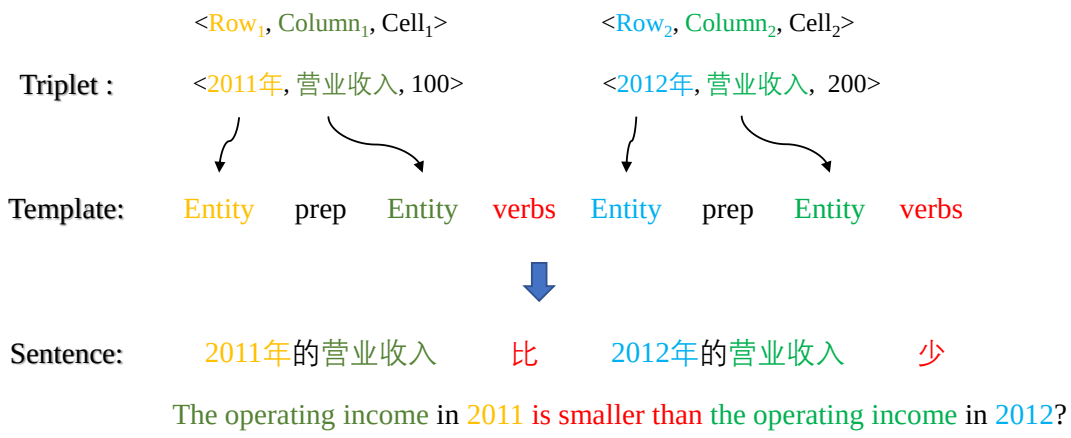

Figure 4: Question generation templates.

Question Type	Percentage (%)	Discourse	Percentage (%)	Example
Table Retrieval	81.5	**Ordinary**	57.5	机器设备的年折旧率是多少? What is the annual depreciation rate of machinery and equipment?
		Coreference	10.5	它的上期金额是多少? What was its prior period amount?
		Ellipsis	32.0	比例是多少? What is the ratio?
Fact Checking	10.3	**Ordinary**	38.5	营业税比城市维护建设税的本期数多吗? Is sales tax more than the current amount of city maintenance and construction tax?
		Coreference	43.3	它比期末数坏账准备大吗? Is it larger than the ending number of bad debt provision?
		Ellipsis	18.2	比2017年的多吗? Is it more than in 2017?
Computation	8.2	**Ordinary**	17.1	本期增加最高的和最低的和是多少? What is the sum of the highest and lowest increase for the period?
		Coreference	29.3	它们的和是多少? What is the sum of them?
		Ellipsis	53.6	小了多少? How much smaller?

Table 7: Question type distribution of Tab-CQA.