Enhancing Financial Table and Text Question Answering with Tabular Graph and Numerical Reasoning

Supplimental Material

	% Data	% Error (Train) Original	Corrected
Sum	2.3	11.6	6.1
Average	6.8	6.3	3.2
Multiply	0.2	20.0	16.7
Division	3.7	12.0	3.3
Difference	16.0	4.6	0.9
Change ratio	9.9	1.5	0.6

Table 1: Percentages of errors in the training set before and after error correction.

	% Data	% Error (Dev) Original	Corrected
Sum	3.4	10.3	3.6
Average	8.5	6.4	0.7
Multiply	0.2	25.0	25.0
Division	3.7	6.5	1.6
Difference	14.4	5.9	1.7
Change ratio	9.3	9.7	0.0

Table 2: Percentages of errors in the development set before and after error correction.

1 Data Preparation

During the dataset development process, the annotators wrote the answers' derivations for automatic label creations. However, while most derivations follow similar patterns, making them easy to process, some do not. We discuss this problem in Section 4.1 of the paper; here, we provide additional detail on a set of derivations that our algorithm could not produce the annotated answer, and we found that it was due to errors in the automatic labeling process. Tables 1 and 2 show the percentages of labeling errors found in the training and development sets and the manual corrections we made. Note that we could not correct all mislabeling for specific reasons, mainly related to the operations that TagOp and our model cannot handle, such as comparison.

2 Graph-based Tabular Evidence Extraction

This section explains how our algorithm locates the table headers. It starts from the top row downwards, first checking for numbers and units. If there are scales (e.g., thousand) but no numbers, it considers that row part of the header. It also identifies the row as a header if there are words and no dollar signs. Otherwise, if the row satisfies at least one of the following conditions, it designates that row as a non-header: (1) There are numbers and no years and months; (2) The row is empty; (3) There is a cell with only "-".

3 Hyperparameter Settings

We conducted our experiments with several different configurations. Some hyperparameters apply across all settings, and some are specific to data sizes and models (Table 3), particularly the learning rate, which we adjusted based on the number of training steps. The following hyperparameters are the same regardless of the data size: dropout rate = 0.1, batch size = 16, epochs = 50, max norm of the gradients = 0.5, warmup ratio = 0.1.

Data Size	RoBERTa Large	RoBERTa Base	Distil BERT
1%	1e-4	5e-5	1e-4
2.5%	5e-5	1e-4	2e-4
5%	1e-4	2e-4	3e-4
10%	2e-4	4e-4	5e-4
25%	2e-4	4e-4	5e-4
50%	3e-4	5e-4	7e-4
All	1e-3	3e-3	5e-3

Table 3: Learning rates specific to data sizes and models.