Experiments in Improving MT2Net Performance for Financial QA

Petar Duric, Meet Gandhi, Juhi Kamdar, Matreyi Pitale

George Mason University | Volgenau School of Engineering | Fairfax, VA

Introduction

Models such as MT2Net by Zhao, et al represent the state of the art in financial QA. The performance of these SOTA models is still significantly behind human experts, and therefore not deployable. However, MT2Net does not utilize domain-specific knowledge of any kind. It also relies on transformations of the tabular input that remove data from context and potentially distorts its meaning. Because input examples are often too large for typical QA networks to handle altogether, SOTA models use the structure proposed by Chen, et al to break the problem into two parts, fact retrieval and reasoning. The reasoning module generates an executable program to produce the answer to the question; however, this segment of the model achieves generally excellent performance when given the correct facts. Therefore, fact retrieval is the weak link for the overall method.

We propose using language models with domain- and format-specific understanding, FinBERT and TaBERT, to overcome the limitations of the MT2Net architecture, and show preliminary results for the use of FinBERT here on the MultiHierTT dataset proposed by Zhao, et al that includes hierarchically organized tables, unlike prior benchmarks..

Objective

Our objective is to achieve performance on the fact retrieval subproblem better than MT2Net's fact retriever. To do this, we replace the RoBERTa LM with FinBERT, and then with both FinBERT and TaBERT (shown at right) to utilize domain- and format-specific knowledge in the hopes of improving fact retrieval.

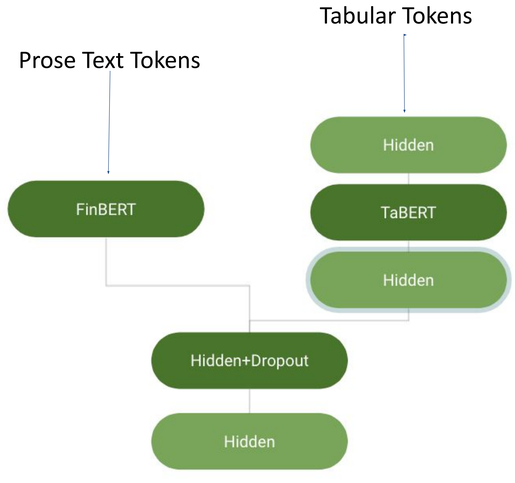
Formally speaking, given a question Q, prose evidence E, tabular evidence T, and a set of facts F found in E and T, we want to maximize the likelihood:

$$P(F|Q,E,T) = \sum P(f \in F|Q,E,T)$$

Methods

Our contribution is to modify the fact retrieval module in two distinct ways to utilize the domain specific knowledge of FinBERT and the format specific, tabular understanding of TaBERT. First, we replace RoBERTa in the MT2Net architecture with FinBERT, keeping the input procedure the same. Second, for a separate experiment, we devise a method to utilize TaBERT and FinBERT in parallel. We utilize FinBERT straightforwardly on the prose text elements of the input example, and we utilize TaBERT on the tabular evidence. In order to ensure that the classification layers that follow the language models can properly classify encodings from both TaBERT and FinBERT, we include a hidden layer to project TaBERT encodings into an output space consistent with FinBERT's encodings. Because TaBERT is not necessarily capable of properly understanding the hierarchically organized tables that predominate in the MultiHierTT dataset, we also include a hidden layer to project the encodings produced by TaBERT's tokenizer into an input space that is more manageable by TaBERT.

Dual-BERT Fact Retriever



Results

The MultiHierTT dataset includes 7830 training, 1566 testing, and 1044 dev examples. Each example may or may not contain hierarchical tables, and a single example may have more than one table. Facts relevant to the question may exist in more than one table and the prose evidence. To extract the top-10 facts, we use an Nvidia-amd gpu with 5 epochs of model training. We allocate 100gb space and finetune the BERT LMs.

For the purposes of evaluating only the fact retriever, we reimplemented the fact retriever only. To get to our main objective we divided our task in two main parts. First we reimplemented Zhao, et al's methodology in PyTorch and got the following results:

	Acc	F1	Loss	Recall
Training	14.3	12.5	9.643	49.3
Validation	14.3	12.5	9.643	50.4

Next, we replaced RoBERTa in Zhao, et al's implementation with FinBERT. We compare the performance of the MT2Net using FinBERT to the basic, RoBERTa-based method at right. Test labels are not available in the dataset, and test performance is evaluated via a leaderboard system we could not access due to time constraints.

Model	F1	Loss	Recall
Roberta-base validation	45.4	2.21	90.76
Finbert Validation	44.89	2.33	89.86

Conclusion

Since FinBERT is already trained on financial data, we expected FinBERT to perform better than RoBERTa on the FinQA task. However, our preliminary results do not bear this out. One potential problem is that FinBERT is not capable of understanding tabular data, and that this is the main difficulty for LMs in this task. Incorporating TaBERT to handle tabular evidence may enhance the overal model significantly for this reason. Nonetheless, further experimentation is required for conclusive results.

Citations

1. Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa- pers), pages 6588–6600, Dublin, Ireland. Associa- tion for Computational Linguistics.
2. Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.
3. Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Se- bastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data.
4. 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.