TAPASGO: Transfer Learning towards a German-Language Tabular Question Answering Model

Dominik Kowieski, Michael Hellwig, Thomas Feilhauer
Josef Ressel Centre for Robust Decision Making
Research Centre Business Informatics
Vorarlberg University of Applied Sciences
Hochschulstraße 1, 6850 Dornbirn, Austria
{dominik.kowieski, michael.hellwig, thomas.feilhauer}@fhv.at

Abstract

Processing tabular data holds significant importance across various domains and applications. This study investigates the performance and limitations of fine-tuned models for tabular data analysis, specifically focusing on using fine-tuning mechanics on an English model towards a potential German model. The validation of the effectiveness of the transfer learning approach compares the performance of the fine-tuned German model and of the original English model on test data from the German training set. A potential shortcut that translates the German test data into English serves for comparison. Results reveal that the fine-tuned model outperforms the original model significantly, demonstrating the effectiveness of transfer learning even for a limited amount of training data. One also observes that the English model can effectively process translated German tabular data, albeit with a slight accuracy drop compared to fine-tuning. The model evaluation extends to real-world data extracted from the sustainability reports of a financial institution. The fine-tuned model proves superior in extracting knowledge from these training-unrelated tables, indicating its potential applicability in practical scenarios. This paper also releases the first manually annotated dataset for German Table Question Answering and the related annotation tool.

Keywords: Natural Language Processing, Information Extraction, Tabular Question Answering, German Language

1. Introduction

The impressive results of the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin and *et al.*, 2019) led to many subsequent improvements and variations on this architecture (Wang et al., 2022). This paper focuses particularly on the task of Question Answering (QA) (Wang, 2022). QA algorithms analyze natural language queries for relevant answers to user provided questions. This is done by looking through large amounts of data. It involves identifying key phrases or instances within a query, determining their semantics, and then searching for matches in the available data sources.

While recent NLP models have achieved remarkable results (OpenAI, 2023) for closed-style QA tasks, they often struggle with tabular data due to their limited ability to deal with structured data as well as a lack of contextual information within the tables (Jin et al., 2022). Hence, Tabular Question Answering (TQA) developed to address these limitations. TQA refers to a specific variant of the QA problem, where answers are looked up in structured tables rather than textual documents. One of the latest developments for TQA is the TAPAS (**Ta**ble **Par**s**e**r) model (Herzig et al., 2020) that emerged from pre-training on millions of data tables from English Wikipedia. The model leverages graph attention mechanisms to capture relationships between table cells. This way, it provides accurate answers to complex questions. Based on BERT, it utilizes relative position embeddings and specific token types to encode tabular structure. TAPAS has been fine-tuned on several data sets, e.g., Wiki Table Questions (Pasupat and Liang, 2015). The TAPAS model exhibits a remarkable performance when processing well-structured table data. Yet, it is still limited to rather small input sizes (≤ 512 table cells) (Google Research, 2020). Besides these limitations, TAPAS is still one of the better-performing TQA models, even when being compared to more recent similar pre-trained models like TAPEX (Zheng et al., 2023). Especially, when handling vertical tabular data, it still represents one of the state-of-the-art algorithms in TQA (Etezadi and Shamsfard, 2022; Yang et al., 2023). As already stated in the domain of oridinary QA by (Höffner et al., 2016), multilingual QA is being seen as one of the biggest challenges in the domain. This attribute holds even more truth in the domain of TQA as the problem of lacking high-quality TQA datasets other than English is even bigger. Hence, to the best of our knowledge, TAPAS models are only available for large-resource languages like English or Chinese TQA tasks and to date, no German version exists. For pre-trained models, this limitation is regarded to be resolvable by fine-tuning on a precise application domain, i.e., a custom task, or specific language, respectively. Such fine-tuning approaches

LREC-COLING 2024, pages 15579–15584
20-25 May, 2024.

for domain adaptation are commonly referred to as transfer learning (Weiss et al., 2016; Bozinovski, 2020). They often can be achieved with minimal amounts of training data. Fine-tuned models typically perform well on tasks in the intended domain. Transfer learning, as demonstrated in other NLP domains like speech recognition, has shown that utilizing knowledge from related languages such as English to German can lead to satisfying results already. This approach reduces the time and resources required to train a robust model, as proven by (Kunze et al., 2017). Given our limited resources, we employ a similar strategy in TQA.

This paper shows the potential of a specific transfer learning approach that aims to increase the usability of TAPAS for German TQA tasks. The evaluation of the model reveals that fine-tuning an English TQA model with German data is suitable to considerably increase the accuracy for German use cases. The comparison with other work-arounds demonstrates the effectiveness of the transfer learning approach. The results introduce opportunities for future research into different types of structured data and more complex question scenarios in German. On top, they shed light on the potential of using translation tools as an alternative to the time-consuming task of fine-tuning. By releasing the manually annotated TQA dataset and the corresponding annotation tool, future research and work should be facilitated, encouraging collaboration and accelerating advancements in this domain.

2. Motivation & Objective

Research on ESG risk assessment (European Banking Authority (EBA), 2021) quickly encounters the challenge of data availability. There are different approaches to closing the current data gaps (Network for Greening the Financial System (NGFS), 2022). One is the extraction of additional information from the written sustainability reports of obligated companies. QA models appear to be suitable for this purpose, but the structure, size, level of detail, and terminology of such sustainability reports vary greatly. This renders conventional QA models, that process continuous text, unfit for this kind of task. For instance, when extracting CO_2 emissions caused by a company, one company might only refer to the annual absolute amount emitted, while another divides it into different scopes, or does not reference it at all.

One common feature of such documents, however, is tabular summaries. Tables are structured data by nature, simplifying the task of retrieving precise information. Yet, the application of conventional QA tools does not provide very satisfactory results, e.g., due to the absence of contextual information. This is where TQA in the guise of the TAPAS model comes into play. To the best of our knowledge, no German TAPAS variant was trained, and early and naive attempts to use an English model for German tables resulted in sobering results. Additionally, we were not able to find any existing German TQA dataset, hence the motivation for creating our own dataset. Though, training a complete self-developed TAPAS model from scratch, with pre-training for the tabular structure and training on the TQA task, would prove to be difficult with limited resources. Especially since the pre-training of the English TAPAS model alone is based on 6.2 million different tables extracted from the English Wikipedia (Herzig et al., 2020).

This study intends to adapt the original model's learned parameters, initially trained for English tables, towards the nuances and structure of German tables. The transfer learning approach is experimentally validated by comparison of the fine-tuned model's performance with that of the original TAPAS model (with and without translation of the test data). Further, all TQA variants compete on unknown-domain bank data, i.e., tables extracted from the sustainability report of a regional bank that were neither part of the training/test data set. This way, we intend to verify whether potential overfitting occurs.

3. Model Fine-Tuning: TAPASGO

Refining the TAPAS model via transfer learning requires a training set of German tabular data with corresponding question and answer (Q&A) pairs. As no training data for TQA in German is available, an intuitive and streamlined interfaces is developed to facilitate and outsource the manual data labeling effort (Kowieski and Hellwig, 2023). It uses freely accessible governmental German data tables (DESTATIS, 2013). This self-developed tool allowed to process 236 different tables and create 1035 Q&A pairs for model fine-tuning. The data include important information for the training like the question-text (e.g. "How many male workers aged between 55 and 65 were employed at their workplace in the year 2020?", the answer-text (e.g. "3596896"), the answer-coordinates (e.g. "[(5, 3)]") inside the table as well as the table (a csv file transformed to a dataframe) itself. Notice, for simplicity and clarity the fine-tuned TAPAS model is referred to as TAPASGO (for **TAPAS G**erman **O**ffshoot). Figure 1 shows the broad training and validation workflow of the fine-tuning and validation process.

The fine-tuning approach uses the pre-trained English *Google/TAPAS-base* model. Leveraging the knowledge of the English language (concerning structural data) encoded in the model's parameters, transfer learning is performed on the self-created German tabular data set. Thus, the En-

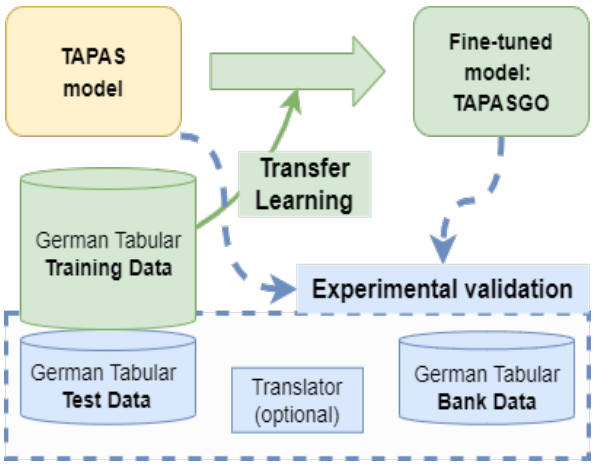

Figure 1: Illustration of the transfer learning and validation process of the TAPASGO development.

glish model extends towards the knowledge of tabular German Q&A structure. As the TAPAS model uses an English tokenizer to transform text into respective token sequences, dealing with the German Q&A data set requires a different tokenization. Hence, the vocabulary file of the German *bert-base-german-cased* model is used to create a custom German tokenizer.

From the self-created data set of 1035 Q&A pairs, only 947 can be utilized as training data because the other questions refer to tables with a too large input size for the model to be executed. From this pool of training data, 80% (757) of the data are for training and 20% (190) for testing. The TAPAS model was fine-tuned using the AdamW-optimizer with a learning-rate of 5e-5. The training ran for 100 epochs. The batch-size for the training data is 13. Other model parameters and the whole architecture were adopted from the original TAPAS-base model (Herzig et al., 2020). The logarithmic loss of the additional training can be observed to continuously decrease on the left hand side of Figure 2. After a sharp drop for the first 50 epochs, one realizes a deceleration of the loss reduction. For this reason, and regarding the rather small amount of training data, training beyond 100 epochs was not considered as the progressively slower decrease in loss demands substantial resources to approximate the theoretical optimum. Further, the right hand side of Figure 2 indicates already a good model accuracy after this number of epochs. The denotation accuracy indicates the precision of the model's predictions on a data set (Liu et al., 2022). It is defined as the ratio of correct predictions compared to the ground-truth. A prediction is considered correct, if the predicted tabular cell coordinates match the expected cell coordinates with the exact answer. On the tabular German test data set, the TAPASGO model realizes a denotation accuracy of 93.16% on the test data set. Further, the improvement of the model precision during training over 100 epochs is monitored and displayed on the right-hand side in Figure 2. The accuracy demonstrates significant improvement during the initial half of the training epochs. After-wards, it slows down and shows some fluctuations which might be an indicator for overfitting.

4. Model Validation

This section presents the experiments performed to validate the TAPASGO model as well as the respective results in comparison to the performance of the original English TQA-base model TAPAS.

4.1. Evaluation on Test Data

First, the TAPASGO performance is compared to two variants of the original TAPAS model. Representing an English TQA model, the first TAPAS variant simply employs the standard tokenizer (English). The second TAPAS variant differs in applying the German tokenizer which is also used in TAPASGO. The comparison is carried out on the test data which correspond to 20% of the self-generated German tabular data set. The results are summarized in Table 1.

The original TAPAS variant reaches a denotation accuracy of 45,26%. This remarkable accuracy of the English model is explained by its structural knowledge which enables TAPAS to even exploit similarities of non-English keywords and turn those into correct answers. Trying to integrate the customized German tokenizer employed in TAPASGO, results in a denotation accuracy of 8,42%. Using tokens in a different language, the corresponding performance loss is not unexpected. Note, that this kind of tokenization was tried for validation purposes only and is not used in the remainder of this paper. Ultimately, one observes that the TAPASGO model significantly outperforms both TAPAS variants with a denotation accuracy of 93.16% on the German TQA task. This indicates that the transfer learning approach fits the model quite well to the German data set.

4.2. Use on Unknown-Domain Data

Until now, test and training data originated from the same source of self-generated German tabu-

Model	Tokenizer	Precision
TAPAS	English	45%
TAPAS	German	8%
TAPASGO	**German**	**93%**

Table 1: Denotation accuracy (precision) of TAPAS and TAPASGO on the test data set.

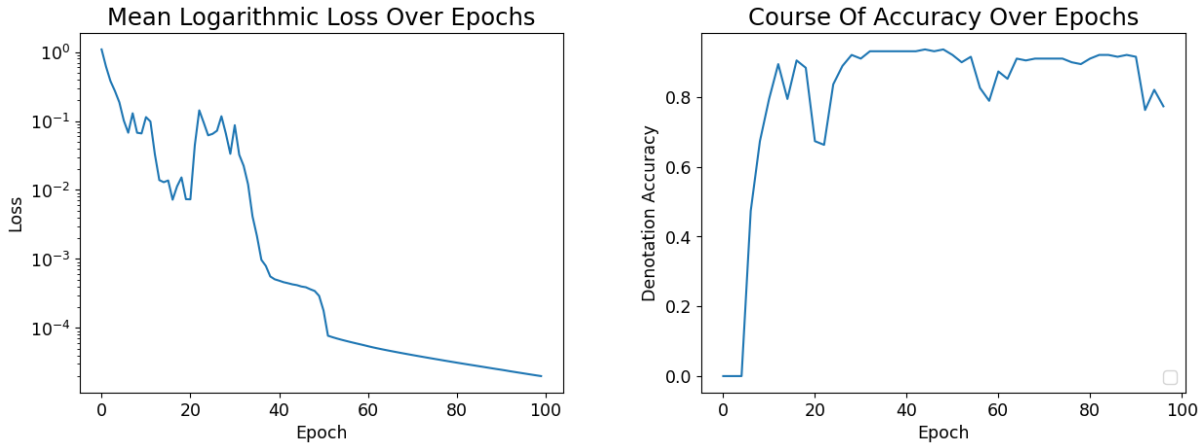

Figure 2: The logarithmic loss (left) as well as the denotation accuracy (right) over the 100 transfer learning epochs of the transition towards the TAPASGO model.

lar data. That is, the test set tends to be of a similar structure as the training set. To this end, as second step, the model performances are assessed on unknown-domain data to investigate their generalizability to real-world tasks. The data set consists of tables extracted from openly-available sustainability reports of a regional bank and are denoted bank data for simplicity.

Extraction of the tables out of the source PDF-files is performed by the Python tool `camelot-py`. Interestingly, the application of both models to the raw tables yields a denotation accuracy of 0%. This is due to the non-standard table formats of the sustainability report. Using different multi-column styles even in a single source document represents a major structural difference from the training data of the original TAPAS model.

Though, introducing an additional preprocessing step to fix the table irregularities resolves this issue, and both models are able to increase in precision. Table 2 displays the respective values for the bank data set. Applying the data to the original English TAPAS-base model, it exhibits a denotation accuracy of 21% while the fine-tuned TAPASCO model reaches a denotation accuracy of 39%. Although this is a deterioration compared to the test accuracy that could be an indicator for slight overfitting with respect to the test-data, TAPASGO is still able to maintain approximately twice the accuracy of the TAPAS model. Yet, the rather low precision illustrates the need of a richer and more diverse German tabular training data set in order to obtain precision values relevant for practical usage. A similar finding was noted in the research conducted by (Zheng et al., 2023), which observed a decrease in TAPAS performance when dealing with more diverse data, particularly when tables are no longer structured in a clear vertical format. This decline is likely due to the fact that pre-trained models such as TAPAS are usually being trained and fine-tuned primarily on tailored and well-structured vertical tables.

4.3. Translation as a Shortcut

Finally, TAPASGO is compared to the performance of the English TAPAS-base model after application of a preceding translation step. Translation can be regarded a convenient shortcut to the data creation and fine-tuning effort necessary for building the TAPASGO model. For this experiment, translation of the German test and bank data (tables as well as Q&A data) is carried out via the `DeepL Python API` (DeepL, 2023). The translated inputs are fed directly into the original TAPAS model with the corresponding English tokenizer. No further training or processing is applied. The resulting accuracy values are displayed in Table 2 as TAPAS*. By applying automated translation to the test data, English TAPAS model reaches a denotation accuracy of 43,16%. This represents a slight drop of about two percentage points which most probably is explained with non-unique translations of descriptive keywords in the table entries. Regarding the unknown-domain bank data, preceding translation (TAPAS*) realizes a denotation accuracy of 36%. While the increase of about 15 percentage points is in line with expectations, the performance is still below that of the TAPASGO model (39%). Taking into account the effort that went into the training of the `Deepl` and the TAPAS model, this supports the investment of resources for the transfer learning approach and the development of a German TQA variant like TAPASGO.

5. Discussion

This study addressed the potential of fine-tuning an English TQA model *Google/TAPAS-base-finetuned-SQA* for another language (i.e., Ger-

Model	Data Set	Precision
TAPAS	Test	45%
TAPAS*	Test	43%
TAPASGO	**Test**	**93%**
TAPAS	Bank	21%
TAPAS*	Bank	36%
TAPASGO	**Bank**	**39%**

Table 2: Precision of different model and data set combinations. TAPAS with preceding translation is indicated by the asterisk (*) after the model name.

man). Both, the fine-tuned TAPASGO model, and the German training data, are openly available (Kowieski and Hellwig, 2023). The experiments provided in Section 4 substantiate the successful application of transfer learning approaches for specific tasks with a limited amount of training data.

The reported denotation accuracy of the initial English TAPAS model (Herzig et al., 2020) ranges from 49% for the WikiTQ data set to 88% (fully supervised) for the WikiSQL data set. This indicates that the precision is substantially dependent on the structure of the evaluation data set used. Interestingly, the TAPAS model retains much of its precision even when applied to the German-language test data set (about 43%). This ability can be explained by the exploitation of learned table structures and is due to the quite homogeneous structure of the German test data set.

The fine-tuned TAPASGO model yields superior performance in all considered settings. In retrospect, this justifies the effort invested in generation of the German tabular training data set, The remarkable precision of 93% on the test data shows how well these language models can be adapted to individual tasks. Yet, the obtained precision (39%) on unknown-domain data must be regarded too low for the intended practical purpose of enhancing the ESG data situation. As also the precision of the TAPAS model is degrading, some of the problems might be attributed to the data quality of the tables. Hence, the denotation accuracy can almost certainly be improved by the targeted construction of a more extensive training data set including information of historical sustainability reports. Further, the training data would benefit from the integration of more diversely structured tables and corresponding Q&A pairs.

The advantage of TAPASGO over the use of an preceding translation step further justifies the transfer learning approach. This is particularly relevant as it eliminates the risk of translation errors which are difficult to trace and can have a severe negative impact on the TAPAS performance.

6. Outlook

Future work will focus on improving the availability and the diversity of German-language TQA data in order to address the limitations of this study. Extending the evaluation of both models to a larger set of evaluation data allows for a more comprehensive evaluation of their performance and generalizability. In this regard, the model can be made more robust by using way more heterogeneous table data with different quality in both the pre-training and fine-tuning stage. Moreover, the presence of a larger amount of training data is expected to support the development of a more accurate and more robust TAPASGO model. Additionally, creating an independent German TQA model that incorporates expensive pre-training may lead to enhanced results.

Acknowledgement

The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

Bibliographical References

Stevo Bozinovski. 2020. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica (Slovenia)*, 44.

DeepL. 2023. DeepL API Documentation.

Ferderal Office of Statistics Germany DESTATIS. 2013. German tabular data.

Jacob Devlin and *et al.* 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Romina Etezadi and Mehrnoush Shamsfard. 2022. The state of the art in open domain complex question answering: a survey. *Applied Intelligence*, 53.

European Banking Authority (EBA). 2021. Report on management and supvervision of ESG risks for credit institutions and investment firms EBA/REP/2021/18.

Google Research. 2020. Tapas limitation issue.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training.

Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. 2016. Survey on challenges of question answering in the semantic web. *Semantic Web*, 8.

Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. 2022. A survey on table question answering: Recent advances.

Dominik Kowieski and Michael Hellwig. 2023. TAPASGO: Resources Repository. On Github.

Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. 2017. Transfer learning for speech recognition on a budget.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. Tapex: Table pre-training via learning a neural sql executor.

Network for Greening the Financial System (NGFS). 2022. Final Report on Bridging Data Gaps.

OpenAI. 2023. Gpt-4 technical report.

Panupong Pasupat and Percy Liang. 2015. Compositional Semantic Parsing on Semi-structured Tables. *arXiv preprint arXiv:1508.00305*.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-Trained Language Models and Their Applications. *Engineering*.

Zhen Wang. 2022. Modern Question Answering Datasets and Benchmarks: A Survey.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Peng Yang, Wenjun Li, Guangzhen Zhao, and Xianyu Zha. 2023. Row-based hierarchical graph network for multi-hop question answering over textual and tabular data. *J. Supercomput.*, 79(9):9795–9818.

Mingyu Zheng, Yang Hao, Wenbin Jiang, Zheng Lin, Yajuan Lyu, QiaoQiao She, and Weiping Wang. 2023. IM-TQA: A Chinese table question answering dataset with implicit and multi-type table structures. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5074–5094, Toronto, Canada. Association for Computational Linguistics.