BOUN at SemEval-2021 Task 9: Text Augmentation Techniques for Fact Verification in Tabular Data

Abdullatif Köksal
Department of Computer Engineering
Boğaziçi University
abdullatif.koksal@boun.edu.tr

Yusuf Yüksel
Department of Computer Engineering
Boğaziçi University
yusuf.yuksel@boun.edu.tr

Bekir Yıldırım
Department of Computer Engineering
Boğaziçi University
bekir.yildirim@boun.edu.tr

Arzucan Özgür
Department of Computer Engineering
Boğaziçi University
arzucan.ozgur@boun.edu.tr

Abstract

In this paper, we present our text augmentation based approach for the Table Statement Support Subtask (Phase A) of SemEval-2021 Task 9. We experiment with different text augmentation techniques such as back translation and synonym swapping using Word2Vec and WordNet. We show that text augmentation techniques lead to 2.5% improvement in F1 on the test set. Further, we investigate the impact of domain adaptation and joint learning on fact verification in tabular data by utilizing the SemTabFacts and TabFact datasets. We observe that joint learning improves the F1 scores on the SemTabFacts and TabFact test sets by 3.31% and 0.77%, respectively.

1 Introduction

Recognizing Textual Entailment (RTE) (Dagan et al., 2005) is one of the core NLP problems for understanding the semantic relations between words and sentences, which is useful for other tasks including Question Answering (Abacha and Demner-Fushman, 2019), Text Summarization (Lloret et al., 2008), and Text Classification (Yin et al., 2019). For the RTE task, datasets of various sizes (Dagan et al., 2005; Bowman et al., 2015) and from different domains (Romanov and Shivade, 2018) have been introduced. However, these works and datasets are solely focused on textual data without considering structured data such as tables.

Recently, question answering (Iyyer et al., 2017) and textual entailment datasets (Wenhu Chen and Wang, 2020; Wang et al., 2021) for tabular data have been introduced. SemEval-2021 Task 9 addresses the problem of statement verification (Phase A) and evidence finding (Phase B) using tables from scientific articles (Wang et al., 2021). The shared task also introduced a new dataset, namely the SemTabFacts dataset, an example from which is provided in Figure 1. The goal of Phase A (Table Statement Verification) of the shared task is to determine whether a statement is entailed, refuted, or unknown given a table and its description (if available). For example, given the table and its description in Figure 1, the first two statements

Distance statistics between buildings of ancient buildings and modern buildings to the main water channel (unit: meter).

Index	Ancient building	Modern building
AVERAGE	174.095	273.917
STDEV.S	58.780	190.928
MIN	6.763	4.868
MAX	321.608	912.368
MEDIAN	173.010	243.885

Statement	Label
There are 2 types of building - Ancient building and Modern building.	Entailed
All the values of Ancient building is less than Modern building except MIN value.	Entailed
The value of Modern building is is lesser than Ancient building in AVERAGE. [1]	Refuted

Figure 1: Sample table, description, and statements from SemTabFacts.

[1]The grammatical error exists in the given dataset.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 431–437
Bangkok, Thailand (online), August 5–6, 2021.

are entailed, whereas the third statement is refuted. This example demonstrates that there are various challenges such as understanding numerical operations and comparisons as well as textual entailment.

Transformers architecture (Vaswani et al., 2017) enabled the pretraining of large language models, which achieve significant improvement in numerous NLP tasks (Wang et al., 2018, 2019). Recent works have also focused on pretraining language models for tabular data by introducing new embedding layers and objective functions, as well as large-scale augmented data to better represent numerical values and rankings (Herzig et al., 2020; Eisenschlos et al., 2020).

Data augmentation is a way to enrich training data to improve the supervised training scheme and is widely used in computer vision (Perez and Wang, 2017) and speech recognition (Park et al., 2019). Different text augmentation techniques such as back translation, synonym replacement, and text editing have been investigated for various tasks including text classification (Wei and Zou, 2019) and natural language inference (Min et al., 2020).

In this study, we aim at investigating the impact of text augmentation on the statement verification task from tables. We implement various text augmentation techniques based on WordNet (Miller, 1998), Word2Vec (Mikolov et al., 2013), and Back Translation (Yu et al., 2018) to enrich the statement variety in the SemTabFacts dataset. We finetune a recently introduced pretrained transformer architecture, the TAPAS model (Eisenschlos et al., 2020), for our approach. In addition, we investigate the domain adaptation and joint learning capabilities of two tabular fact verification datasets: SemTabFacts and TabFact. Promising results are achieved on the SemTabFacts test dataset.

2 Datasets

We use two different table-based fact verification datasets for the experiments: SemTabFacts (Wang et al., 2021) and TabFact (Wenhu Chen and Wang, 2020). We compare SemTabFacts and TabFact in terms of the average size of the tables, average word length of the statements, and the number of examples for each class in Table 1. We only report the statistics for the training sets, since the development and test sets have similar distributions with the training sets in both datasets. There is almost an order of magnitude difference between the datasets in terms of the number of tables and statements. Furthermore, we observe that the average table size and average statement length in terms of words are greater in TabFact than SemTabFacts.

	SemTabFacts	**TabFact**
# Tables	981	13,182
# Statements	4,506	92,283
# Entailed	2,818	50,820
# Refuted	1,688	41,463
Avg. Row Size	9.0±8.0	13.5±8.6
Avg. Column Size	5.3±2.9	6.4±1.7
Avg. Statement Length (Words)	11.5±7.1	13.2±4.5

Table 1: Comparative statistics of SemTabFacts and TabFact.

SemTabFacts (Wang et al., 2021): This dataset consists of tables from articles published in Elsevier, which are available on ScienceDirect. After filtering complicated examples, five entailed and five refuted statements about these tables are generated by high-quality crowd-sourcing. These statements are further verified by additional crowdsource workers, especially for filtering out ungrammatical sentences. To increase the quality level, Wang et al. (2021) further verified the statements in the development and test sets. The SemTabFacts dataset also contains automatically generated statements and unknown classes in the development and test sets for the fact verification and evidence finding tasks. In this study, we target two-way (Entailed / Refuted) classification without automatically generated statements for the fact verification task.

SemTabFacts releases tables and statements in XML format. We convert these tables into CSV format to properly use in our models. Due to cells with multirow and multicolumn features in XML, we could not accurately convert all tables into CSV, which might affect our models' overall performance. We manually checked the XML to CSV conversion of 50 tables. We identified three errors related to multirow and multicolumn features, and one error that causes a missing column.

TabFact (Wenhu Chen and Wang, 2020): This dataset crawls tables from Wikipedia articles following previous works on table question answering (Pasupat and Liang, 2015; Zhong et al., 2017).

Designed input parameters during experimental set-up.

Experimental bottle ID	Material type	Material amount, g
A	Inoculum	200.2
B	Inoculum	200.4
1A	Inoculum + Olive cake	200.4
1B	Inoculum + Olive cake	199.7

Augmentation Methods	Sentence
Original	199.7 is the lowest Inoculum compare to all others.
WordNet	199.7 is the small Inoculum compare to all others.
Word2Vec	199.7 is the lowest Inoculum comparisons to all others.
Back Translation	199.7, compared to others is the most low inoculum.

Figure 2: Generated sentences by different text augmentation methods for the same statement. The table for the original statement is given above with some modifications.

Complicated tables including multirows, multicolumns, and latex symbols, and large tables with more than 50 rows or 10 columns were filtered out. Amazon Mechanical Turk was used to generate simple and complex statements about tables. The Mechanical Turk workers also filtered out poor statements that have grammatical errors or vague claims. Finally, annotator agreement scores were computed by having the same set of statements labeled by another set of Mechanical Turk workers.

3 Methods

3.1 TAPAS

Deep transformers models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have achieved significant improvement in different NLP tasks as seen in the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks. However, it is not straightforward to benefit from these models for structured data formats such as tables or graphs. TAPAS (Herzig et al., 2020) introduces different objectives such as cell selection and aggregation prediction, and new additional embeddings such as column/row id and rank id over BERT's architecture, which are more suitable for complex numerical operations and comparisons in tables. The TAPAS model has been designed by focusing on the task of question answering over tables (Herzig et al., 2020). However, TAPAS fails to handle complex compositional structures like multiple aggregations and large tables due to the maximum length limit of the tokenizer.

To overcome the problems in (Herzig et al., 2020), recently, Eisenschlos et al. (2020) introduced new mechanisms such as table pruning to make TAPAS work with large tables without memory errors. Furthermore, two augmentation methods for statements were presented (Eisenschlos et al., 2020). The first one is based on creating counterfactual statements by replacing entity mentions with other entities on entailed examples to populate negative samples. The second one is based on a synthetic data generation method to populate statements with complex numerical operations.

In this study, we use a TAPAS model from HuggingFace's Transformers library (Wolf et al., 2019). This model is pretrained on Masked Language Model and additional intermediate pretraining steps as discussed in (Eisenschlos et al., 2020). In addition, it is finetuned on the TabFact dataset (Wenhu Chen and Wang, 2020). We further finetuned this model on SemTabFacts (Wang et al., 2021) with additional augmentation steps by utilizing WordNet, Word2Vec, and back translation.

3.2 Text Augmentation

3.2.1 WordNet

WordNet (Miller, 1998) is a lexical database that groups words into adverbs, adjectives, nouns, and verbs, and shows the relations between them such as hyponymy, antonymy, and synonymy. In this work, we focus on swap-based WordNet augmentation that changes words by their WordNet synonyms. The implementation is done by the TextAttack (Morris et al., 2020) library. As shown in Figure 2, the word *lowest* is changed to *small* by synonym swapping.

3.2.2 Word2Vec

Word2Vec (Mikolov et al., 2013) is a technique to find dense word embeddings by shallow networks.

Training Set	Dev	Test
SemTabFacts	0.7661	0.7044
SemTabFacts+WN	0.7791	**0.7294**
SemTabFacts+W2V	0.7486	0.7201
SemTabFacts+BT	0.7725	0.6941
SemTabFacts+WN+W2V	0.7648	0.7147
SemTabFacts+WN+BT	**0.7869**	0.7217
SemTabFacts+W2V+BT	0.7614	0.7202
SemTabFacts+W2V+WN+BT	0.7484	0.7101

Table 2: F1 scores of different augmentation techniques on SemTabFacts. WN, W2V, and BT represent WordNet, Word2Vec, and Back Translation, respectively.

It helps to represent syntactic and semantic features of words by a dense vector. Due to the low dimensional space, similar words and synonyms have closer word embeddings. We make use of this feature of Word2Vec to replace words with their Word2Vec synonyms by TextAttack library. For example, this augmentation technique changes the word *compare* to *comparisons* as shown in Figure 2. While WordNet augmentation preserves the part of speech tags of the words, Word2Vec augmentation may distort the part of speech tags and may produce ungrammatical sentences.

3.2.3 Back Translation

The back translation technique paraphrases a given sentence from a source language by translating it into another target language and then translates it back into the source language. It was first introduced as a data augmentation mechanism for the reading comprehension task (Yu et al., 2018), where significant improvement was observed by back translation augmentation. Recent machine translation systems (Sennrich et al., 2016) are robust to back translation mechanism and tend to produce the same sentence. To overcome this issue, we used two different versions of the same system. First, we translated the statements in English to Turkish by Google Translate [2]. Then, we translated the Turkish statements into English by the GOOGLETRANSLATE function in Google Sheets. The back translation method paraphrases the sentence and unlike the WordNet and Word2Vec approaches, it may change word order in addition to the words, as illustrated in Figure 2.

[2] https://translate.google.com

Training Set	SemTabFacts		TabFact	
	Dev	Test	Dev	Test
SemTabFacts	0.7661	0.7044	0.7435	0.7471
TabFact	0.7284	0.7019	**0.8200**	0.8178
SemTabFacts +TabFact	**0.7992**	**0.7335**	0.8167	**0.8255**

Table 3: F1 scores of domain adaptation and joint learning capabilities of SemTabFacts and TabFact.

4 Experimentation and Results

We conduct two different experimental setups to compare our results. In both experiments, we finetune all layers of a pretrained TAPAS model and its classifier head. First, we finetune the TAPAS model on the SemTabFacts dataset with all combinations of different augmentation techniques. Second, instead of using augmentation techniques, we finetune the TAPAS model on TabFact only and then on SemTabFacts and TabFact jointly and compare the results on the test sets of TabFact and SemTabFacts. In all these experiments, we use the AdamW (Loshchilov and Hutter, 2018) optimizer with a $5e-5$ learning rate and 0.01 weight decay. We set batch size as 8 with 2 accumulation steps, and the number of steps used for linear warm-up is 100 in SemTabFacts training and 2000 in TabFact and joint training. We finetune this model over 10 epochs and decide the best model based on the development set. We use the official evaluation metric, which is the macro-average of F1 scores over the tables.

In the augmentation steps, we include new augmented statements for each statement and augmentation method to the training data of SemTabFacts. The original versions of the development and test sets are used without any augmentation. We observe that different augmentation techniques in SemTabFacts can improve F1 scores on the test set as shown in Table 2. The best model for the development set of SemTabFacts is the model with WordNet and back translation augmentations. Besides, all augmentation techniques, except back translation, improve the test F1 score over the base model without augmentation. Finally, we observe that WordNet augmentation increases the test F1 score by 2.5% over the base model without augmentation.

In Table 3, we investigate the domain adaptation and joint learning capabilities of SemTabFacts and

TabFact. We have finetuned three separate models, with SemTabFacts training data, TabFact training data, and SemTabFacts and TabFact training data. We evaluate these finetuned models on the development and test sets of the datasets. The original versions of the training, development, and test sets are used in these experiments without any additional augmentation. The model trained with TabFact and SemTabFacts data achieved the highest F1 scores on the test sets of both datasets. The joint model improves the F1 score by 3.31% and 0.77% on the SemTabFacts and TabFact test sets, respectively. Further, we observe that we can achieve similar scores on the SemTabFacts test set when the model is trained on the SemTabFacts training data or on the TabFact training data.

We further analyzed the errors in terms of table sizes (number of rows x number of columns) and length of the statements. However, our results indicate that there is no significant difference in F1 scores for different table sizes and different lengths of statements. Our models have similar performance for small and large tables as well as for short and long statements.

5 Conclusion

In this work, we described our models for the Table Statement Support Subtask (Phase A) of SemEval-2021 Task 9. Our base model relies on the recently introduced pretrained transformer architecture for tabular data, TAPAS. We proposed three different augmentation techniques which are based on WordNet, Word2Vec, and Back Translation. We showed that all combinations of these augmentation techniques except Back Translation perform better on the test set than methods without augmentation. Furthermore, we investigated the domain adaptation and joint learning capabilities of SemTabFacts and TabFact. We showed that our best model in terms of development and test F1 for SemTabFacts occurs when we trained TAPAS jointly on the SemTabFacts and TabFact datasets. Additionally, we illustrated that the joint model achieves better results on the TabFact test set than the model trained only on the TabFact training dataset. As future work, we plan to focus on better preprocessing the SemTabFacts dataset and more diverse augmentation techniques by integrating perplexity scores of augmented statements.

Acknowledgments

TUBITAK-BIDEB 2211-A Scholarship Program (to A.K.), and TUBA-GEBIP Award of the Turkish Science Academy (to A.O.) are gratefully acknowledged.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):1–23.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

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.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

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.

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.

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*.

Elena Lloret, Oscar Ferrández, Rafael Munoz, and Manuel Palomar. 2008. A text summarization approach under the influence of textual entailment. In *NLPCS*, pages 22–31.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, volume 26, pages 3111–3119.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *Proc. Interspeech 2019*, pages 2613–2617.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

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 *NIPS*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Nancy Xin Ru Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 Task 9: A fact verification and evidence finding dataset for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of SemEval*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V

Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.