Sattiy at SemEval-2021 Task 9: An Ensemble Solution for Statement Verification and Evidence Finding with Tables

Xiaoyi Ruan, Meizhi Jin, Jian Ma, Haiqin Yang[§]**, Lianxin Jiang,**
Yang Mo and **Mengyuan Zhou**
Ping An Life Insurance Co., Ltd.
Shenzhen, Guangdong province, China
{RUANXIAOYI687, EX-JINMEIZHI001, MAJIAN446, JIANGLIANXIN769, MOYANG853}@pingan.com.cn
[§] the corresponding author, email: hqyang@ieee.org

Abstract

Question answering from semi-structured tables can be seen as a semantic parsing task and is significant and practical for pushing the boundary of natural language understanding. Existing research mainly focuses on understanding contents from unstructured evidence, e.g., news, natural language sentences, and documents. The task of verification from structured evidence, such as tables, charts, and databases, is still less explored. This paper describes sattiy team's system in SemEval-2021 task 9: Statement Verification and Evidence Finding with Tables (SEM-TAB-FACT). This competition aims to verify statements and to find evidence from tables for scientific articles and to promote the proper interpretation of the surrounding article. In this paper, we exploited ensemble models of pre-trained language models over tables, TaPas and TaBERT, for Task A and adjust the result based on some rules extracted for Task B. Finally, in the leaderboard, we attain the F1 scores of 0.8496 and 0.7732 in Task A for the 2-way and 3-way evaluation, respectively, and the F1 score of 0.4856 in Task B.

1 Introduction

Semantic parsing is one of the most important tasks in natural language processing. It not only needs to understand the meaning of natural language statements, but also needs to map them to meaningful executable queries, such as logical forms, SQL queries, and Python code (Pan et al., 2019; Lei et al., 2020; Zhu et al., 2021). Question answering from semi-structured tables is usually seen as a semantic parsing task (Pasupat and Liang, 2015), where questions are translated into logical forms that can be executed against the table to retrieve the correct denotation (Zhong et al., 2017).

Practically, it is significant in natural language understanding to verify whether a textual hypothesis is entailed or refuted by evidence (Benthem, 2008; D. et al., 1978). The verification problem has been extensively studied in different natural language tasks, such as natural language inference (NLI) (Bowman et al., 2015), claim verification (Hanselowski et al., 2018), recognizing of textual entailment (RTE) (Dagan et al., 2005), and multi-model language reasoning (NLVR/NLVR2) (Suhr et al., 2018). However, existing research mainly focuses on verifying hypothesis from unstructured evidence, e.g., news, natural language sentences and documents. Research of verification under structured evidence, such as tables, charts, and databases, is still in the exploratory stage.

This year, SemEval-2021 Task 9: Statement Verification and Evidence Finding with Tables (SEM-TAB-FACT), aims to verify statements and find evidence from tables in scientific articles (Wang et al., 2021a). It is an important task targeting at promoting proper interpretation of the surrounding article.

The competition tries to explore table understanding from two tasks:

– Task A - Table Statement Support: The task aims to determine whether a statement is fully supported, refuted, or unknown to a given table.
– Task B - Relevant Cell Selection: given a statement and a table, the task is to determine which cells in the table provide evidence for supporting or refuting the statement.

The competition contains the following challenges:

– In task A, there is no training data for the "Unknown" category and the number of tables is small in the training set.
– The lexical expression of the table may be different from that in the statement.
– The table structure is complex and diverse. A table may contain missing values while the the supporting evidence may be resided in cells across

several rows or columns.

– It is difficult to understand the statements. For example, some statements express totally different semantics meaning with only one different word. This difficulty makes it even harder to find the evidence cells from tables.

To overcome these challenges, we incorporate several key technologies in our implementation:

– developing a systematic way to generate data from the "Unknown" category;
– including additional data corpus to enrich the training data;
– exploiting existing state-of-the-art pre-trained language models over tables, TaBERT (Yin et al., 2020) and TaPas (Yin et al., 2020), and ensembling them into a powerful one;
– aligning contents in tables and statements while constructing manual rules for tackling Task B.

The test shows that our implementation can increase the performance according and finally, in the leaderboard, we attain the F1 scores of 0.8496 and 0.7732 in Task A for the 2-way and 3-way evaluation, respectively, and the F1 score of 0.4856 in Task B.

The rest of this paper is organized as follows: In Sec. 2, we briefly depict related work to our implementation. In Sec. 3, we detail our proposed system. In Sec. 4, we present the experimental setup and analyze the results. Finally, we conclude our work in Sec. 5.

2   Related Work

Recently, pre-trained language models (PLMs), e.g., BERT (Devlin et al., 2019), XLNET (Yang et al., 2019), and RoBERTa (Liu et al., 2019), have witnessed the burgeoning of promoting various downstream NLP tasks, such as reading comprehension, named entity recognition and text classification (Li et al., 2020; Lei et al., 2021). However, the current pretrained language models are basically trained on the general text. They are not fit for some tasks, e.g., Text-to-SQL, Table-to-Text, which need to encode the structured data, because the data in the structured table also needs to be encoded at the same time. Directly applying the existing PLMs may face the problem of inconsistency between the encoded text from the table and the pretrained text.

TaBERT(Yin et al., 2020) is a newly proposed pretrained model built on top of BERT and jointly learns contextual representations for utterances and the structured schema of database (DB) tables. This model views the verification task completely as an NLI problem by linearizing a table as a premise sentence and applies PLMs to encode both the table and statements into distributed representation for classification. This model excels at linguistic reasoning like paraphrasing and inference but lacks symbolic reasoning skills. Intuitively, encoding more table contents, e.g., type information and content snapshots, relevant to the input utterance could potentially help answer questions that involve reasoning over information across multiple rows in the table because they can provide more hints about the meaning of a column. TaPas (Herzig et al., 2020) is another newly proposed pretrained question answering model over tables implemented on BERT to avoid generating the logical forms. The model can fine-tune on semantic parsing datasets, only using weak supervision, with an end-to-end differentiable recipe.

Another stream of work on evidence finding with table is the rule-based approaches. Most evidence cells can be extracted by rules. For example, if a row head or column head appears in the statement, we infer this row or col support this statement. Although rule-based approaches suffer from the low recall issue, they exhibit high precision and can be applied to adjust the result for ensemble.

3   System Overview

We elaborate the task and present our system in the following.

3.1   Data Description and Tasks

In Task A, the original dataset is a set of XML files, where each XML file represents a table and contains multiple sections:

– *document* section represents the whole document;
– *table* section determines the unique ID of the document;
– *caption* section is a brief description of the table;
– *legend* section is a detailed description of the table;
– *multiple row* sections describe the contents of each row of the table; and
– *statements* section provides several factual statements.

This task aims to determine if a statement is entailed or refuted by the given table, or whether, as is in some cases, this cannot be determined from

		Tables	Label Distribution		Tokens in Statements	Tokens in Tables
			Entailed/Refuted/Unknown		Max./Min./Avg.	Max./Min./Avg.
Train	Task 9	981	2,818/1,688/0		88/3/11	302/1/10
	Tabfact	16,573	63,962/54,313/0		57/4/14	127/5/13
	Augm.	17,554	66,780/56,001/61,436		88/3/12	302/1/11
Dev.	—	52	250/213/93		53/4/13	115/2/11
Test	—	52	—		82/4/12	69/2/11

Table 1: Data statistics.

the table. The competition also provides two kinds of evaluations for the task:

- 3-way F1 score evaluation: a standard precision/recall evaluation (3-way) is computed to evaluates whether each table is correctly classified into one of the three types in {Entailed, Refuted, Unknown}. It is to test whether the classification algorithm understands cases where there is insufficient information to make a determination.

- 2-way F1 score evaluation: the F1 score is computed to evaluate the performance when the statements with the "unknown" ground truth label are removed. The metric will also penalize misclassifying Refuted/Entailed statement as unknown.

In the evaluation, the score for all statements in each table is first averaged and then averaged across all tables to get the final F1 score.

In Task B, the raw dataset is a subset of task A, where unknown statements are excluded. The goal is to determine for each cell and each statement, if the *cell is within the minimum set of cells* needed to provide evidence for the statement "relevant" or "irrelevant". For some statements, there may be multiple minimal sets of cells that can be used to determine statement entailment or refusal. In such cases, the ground truth will contain all of the versions. The evaluation will calculate the recall and precision for each cell, with "relevant" cells as the positive category. The evaluation is conducted similarly as that in Task A.

3.2 Data Augmentation

There are mainly two critical issues in Task A. First, the number of the tables is small. We then include more external data, the TabFact dataset (Chen et al., 2020) to improve the generalization of our proposed system. Second and more critically, "unknown" statements do not exist in the training set but may appear in the test set. To allow our sys-

tem to output the "unknown" category, we construct additional "unknown" statements to enrich the training set. More specifically, we randomly select some statements from other tables and assign them to the "unknown" category for the current table. In order to keep balance on the labels, the number of selected statements from other tables is set to half of the statements in the current table. Details about the data statistics can be referred to Table 1.

3.3 Model Ensemble for Task A

Figure 1 outlines the overall structure of our system, which is an ensemble of two main pretrained models on table-based data, TaBERT and TaPas, or two variants of TaBERT and four variants of TaPas. It is worth noting that the input of all models are the same. That is, given a statement and a table, the input is started with the sentence token, [CLS], followed by the sequence of the tokens in the statement, the segmentation token ([SEP]), and the sequence of the tokens in the flattened table. All the tokens in the statement and the table are extracted by wordpiece as in BERT and related NLP tasks (Devlin et al., 2019; Yang, 2019; Yang et al., 2021; Yang and Shen, 2021; Wang et al., 2021b). The flattened table means that we borrow the implementation in TaBERT by only extracting the most likely content snapshot as detailed in Sec. 3.4. The obtained tokens' embeddings are then fed into six strong baselines, i.e., two variants of TaBERT and four variants of TaPas, to attain the classification scores for the corresponding labels. The classification scores are then concatenated and fed into a vote layer, i.e., a fully-connected network, to yield the final prediction.

3.4 Content Snapshot

In order to pin point the important rows and avoid excessively encode input from the table, we borrow the idea of content snapshot in TaBERT (Yin et al.,
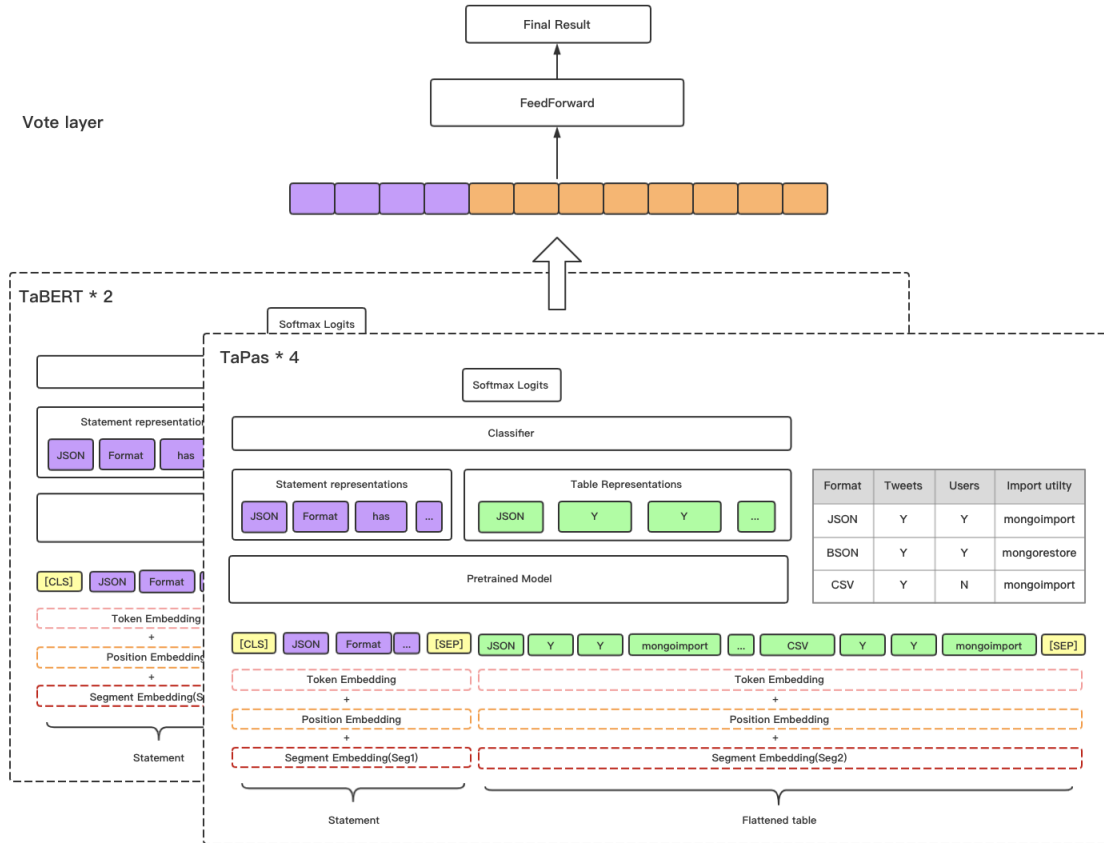
Figure 1: Ensemble architecture

Models	Original data	+TabFact	+Augm.
TaBERT_1	0.7446/0.6580	0.7634/0.6837	0.8003/0.7502
TaBERT_3	0.7689/0.6792	**0.7952**/0.7008	**0.8241/0.7653**
TaPas_TFIMLR	0.7502/0.6637	0.7859/0.6799	0.8102/0.7649
TaPas_WSIMLR	0.7498/0.6522	0.7852/0.7005	0.8024/0.7577
TaPas_IMLR	0.7538/0.6358	0.7799/0.6890	0.7908/0.7396
TaPas_WSMLR	**0.7695/0.6875**	0.7904/**0.7058**	0.8156/0.7609

Table 2: Comparison of strong baselines in Task A for 2-way and 3-way evaluation.

Models	+TabFact	+Augm.	+Rule
TaBERT_1	0.4025	0.4159	0.4605
TaBERT_3	0.4158	0.4305	0.4685
TaPas_TFIMLR	0.4253	0.4299	0.4597
TaPas_WSIMLR	0.4199	0.4208	0.4682
TaPas_IMLR	0.4006	0.4102	0.4467
TaPas_WSMLR	**0.4258**	**0.4386**	**0.4708**

Table 3: Comparison of strong baselines in Task B.

2020) to encode only a few rows that are most relevant to the statement. We create the content snapshot of K rows based on the following simple strategy. First, we count the number of rows of each table and find their median, say R. If the number of rows in the current table is less than or equal to R, then K is set to the total number of rows in the current table and the content snapshot is the entire content of the current table. If the number of rows in the current table is greater than R, we set $K = R$ and select the top-K row with the highest overlap rate between the statement and each row of n-grams as the candidate rows.

3.5 Rule Construction for Task B

For Task B, we apply the same model trained in Task A to find whether the current table supports the statement. If yes, we label all cells as entailed. Otherwise, we first align the word expression in tables and statements while building the corresponding rules to adjust the model prediction. That is, we change uppercase to lowercase and transform all abbreviations into the full name in statements, cells, col heads and row heads. We also conduct stemming on all words. For example, "definition" and "defined" is transformed to "define". After that, we collect all words in a statement into a word bag and determine the supporting relation based on the following rules: 1) If a word in the word bag appears in a row head, we then infer that cells in the whole column supports the statement; 2)If a word appears in the first column of the table, we then infer that cells in the whole row supports the statement; 3) If a word appears in both a row head and a cell in the first column of a table, we then infer that the cell corresponding to the row and column supports the statement; 4) If a word appears in a cell, we then infer that this cell supports the statement.

4 Experiments

In the following, we present the strong baselines and the results with analysis.

We have tried different combinations of TaBERT and TaPas pre-trained models and choose the following 6 best baselines: 1) TaBERT_1: the pre-trained TaBERT with $K = 1$; 2) TaBERT_3: the pre-trained TaBERT with $K = 3$; 3) TaPas_TFIMLR: the pre-trained large TaPas downloaded from tapas_tabfact_inter_masklm_large_reset.zip; 4) TaPas_WSIMLR: the pre-trained large TaPas downloaded from tapas_wikisql_sqa_inter_masklm_large_reset.zip; 5) TaPas_IMLR: the pre-trained large TaPas downloaded from tapas_inter_masklm_large_reset.zip; 6) TaPas_WSMLR: the pre-trained large TaPas downloaded from tapas_wikisql_sqa_masklm_large_reset.zip. Our proposed system is funetuned on the above models for the original training data, the original data with the TabFact data, and the augmentation data. We also tune the hyper parameters to fit a better result in the local test dataset.

Table 2 reports the evaluation results of Task A on the development set when funetuning the above six strong baselines on different training data. The results show that the TaPas_WSMLR attains the best performance on the original data. The best performance is further improved from 0.7695 to 0.7952 for 2-way evaluation and from 0.6875 to 0.7058 for 3-way evaluation, respectively, by including the TabFact data. The performance is further improved to 0.8241 for 2-way evaluation and 0.7653 for 3-way evaluation, respectively, by adding the augmentation data. Finally, we apply the voting mechanism to ensemble the results and achieve the F1 scores of 0.8496 and 0.7732 on the test set, respectively.

Table 3 reports the results of Task B on the development set when funetuning the above six strong baselines on different training data. The results show that the TaPas_WSMLR attains the best performance among all six strong baselines and the perform increases from 0.4258 after adding the TabFact data, to 0.4386 after adding the augmentation data, and to 0.4708, additional 7.3% improvement after adding the manual rules. We conjecture that TaPas_WSMLR can provide more complementary information for solving the task. Finally, we ensemble the results by the voting mechanism and achieve the F1 score of 0.4856 on the test set.

In sum, results in Table 2 and Table 3 confirm the effectiveness of our proposed system by including more training data and the manual rules.

5 Conclusion

In this paper, we present the implementation of our ensemble system to solve the problem of SemEval 2021 Task 9. To include more training data and resolve the issue of lacking data from the "Unknown" category in the training set, we include external corpus, the TabFact dataset, and specially construct the augmented data for the "Unknown" category. Content snapshot is also applied to reduce the encoding effort. Six pre-trained language models over tables are funetuned on the TabFact dataset and the augmented data with content snapshot tables to evaluate the corresponding performance. An ensemble mechanism is applied to get the final result. Moreover, data alignment and manual rule determination are applied to solve Task B. Finally, our system attains the F1 score of 0.8496 and 0.7732 in Task A for 2-way and 3-way evaluation, respectively, while getting the F1 score of 0.4856 in Task B.

References

Jfak Van Benthem. 2008. A brief history of natural logic. *London*, volume 55(2):339–346(8).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

D., C., P., Suppe, and Frederick. 1978. The structure of scientific theories. *The American Journal of Psychology*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *CoRR*, abs/1809.01479.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4320–4333. Association for Computational Linguistics.

Wenqiang Lei, Yisong Miao, Runpeng Xie, Bonnie Webber, Meichun Liu, Tat-Seng Chua, and Nancy F Chen. 2021. Have we solved the hard problem? it's not easy! contextual lexical contrast as a means to probe neural coherence. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wenqiang Lei, Weixin Wang, Zhixin Ma, Tian Gan, Wei Lu, Min-Yen Kan, and Tat-Seng Chua. 2020. Re-examining the role of schema linking in text-to-sql. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6943–6954.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652.

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. *CoRR*, abs/1907.11692.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1470–1480. The Association for Computer Linguistics.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *CoRR*, abs/1811.00491.

Nancy Xin Ru Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021a. Semeval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts). In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*.

Xinyi Wang, Haiqin Yang, Liang Zhao, Yang Mo, and Jianping Shen. 2021b. Refbert: Compressing bert by referencing to pre-computed representations. In *IJCNN*.

Haiqin Yang. 2019. BERT meets chinese word segmentation. *CoRR*, abs/1909.09292.

Haiqin Yang and Jianping Shen. 2021. Emotion dynamics modeling via bert. In *IJCNN*.

Haiqin Yang, Xiaoyuan Yao, Yiqun Duan, Jianping Shen, Jie Zhong, and Kun Zhang. 2021. Progressive open-domain response generation with multiple controllable attributes. In *IJCAI*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8413–8426. Association for Computational Linguistics.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. *arXiv e-prints*, page arXiv:2005.08314.

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

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv: 2101.00774*.