TOWARDS CONTROLLED TABLE-TO-TEXT GENERATION WITH SCIENTIFIC REASONING

*Zhixin Guo[1], Jianping Zhou[1], Jiexing Qi[1], Mingxuan Yan[1], Ziwei He[1],
Guanjie Zheng[1], Zhouhan Lin[1], Xinbing Wang[1], Chenghu Zhou[1,2]*

[1] Shanghai Jiao Tong University, Shanghai, China
[2] IGSNRR, Chinese Academy of Sciences, Beijing, China

ABSTRACT

The sheer volume of scientific experimental results and complex technical statements, often presented in tabular formats, presents a formidable barrier to individuals acquiring preferred information. The realms of scientific reasoning and content generation that adhere to user preferences encounter distinct challenges. In this work, we present a new task for generating fluent and logical descriptions that match user preferences over scientific tabular data, aiming to automate scientific document analysis. To facilitate research in this direction, we construct a new challenging dataset CTRLSciTab consisting of table-description pairs extracted from the scientific literature, with highlighted cells and corresponding domain-specific knowledge base. We evaluated popular pre-trained language models to establish a baseline and proposed a novel architecture outperforming competing approaches. The results showed that large models struggle to produce accurate content that aligns with user preferences. As the first of its kind, our work should motivate further research in scientific domains. [1]

Index Terms— Table-to-text Generation, Scientific Reasoning, Controlled Generation

1. INTRODUCTION

Table-to-text generation is a significant research problem involving the generation of analytical descriptions from tabular data. Recently, pre-trained language model (PLM)-based approaches have demonstrated impressive improvements in text fluency and fidelity compared to traditional template-based methods, as seen in the popular table-to-text challenges, WikiBio [1], LogicNLG [2], RottoWire [3], and ToTTo [4]. However, their success is highly dependent on pre-training using a large corpus of open-domain text, such as the two billion words of Wikipedia [5].

Scientific domains that require advanced expertise present a significant challenge for PLM-based NLG systems in generating descriptions with scientific reasoning. Scientific reasoning is a cognitive process characterized by systematic, logical, and evidence-based reasoning to understand the natural world, develop hypotheses, design experiments, evaluate evidence, and draw conclusions based on empirical data and observations [7], [8], [9]. In the context of scientific natural language processing (NLP), scientific reasoning represents a significant bottleneck, relying primarily on domain-specific knowledge tailored to specific scientific phenomena.

Guanjie Zheng is the corresponding author.

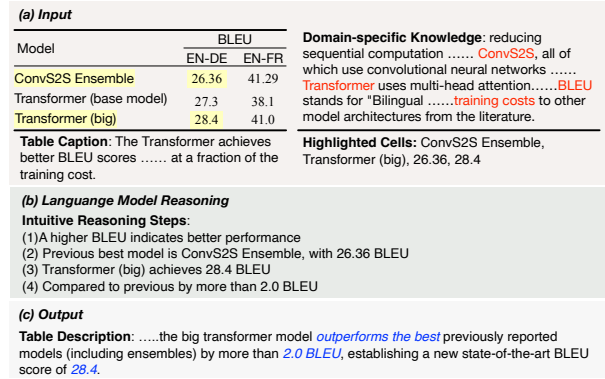[1] We release our code at `https://github.com/sjtugzx/CTRLSciTab`.

Fig. 1. An illustration of controlled table-to-text generation incorporating explicit scientific reasoning stages. (a) represents the input information, (b) illustrates the inherent reasoning processes of language models, and (c) displays the resultant descriptions. Yellow highlights user preferences; red relates to tabular knowledge, and blue indicates scientific reasoning content. Potential scientific reasoning steps are outlined at the bottom. The original table is adapted from [6].

While existing challenges such as SciGen [10] and numeric-NLG [11] address table-to-text generation related to scientific content, they predominantly focus on numerical reasoning. We argue that scientific reasoning in the context of scientific NLP presents a unique and comprehensive challenge for table-to-text generation. Moreover, with the increasing use of PLM-based natural language generation (NLG) systems in real-world applications, the focus has shifted from the generation of generic content to the generation of customized content that is aligned with user preferences. However, the currently popular methods of table-to-text tasks may not consistently match user preferences due to the broad nature of the generated output.

In this work, we propose a new task of controlled table-to-text generation with scientific reasoning, which aims to generate analytical descriptions that are consistent with domain-specific knowledge and align with user preferences. To facilitate research on this task, we present CTRLSciTab, which consists of 8,967 scientific table-description pairs with an external domain-specific knowledge base and highlighted cells. CTRLSciTab represents a challenging table-to-text generation task with two unique features: (1) all pairs are from the scientific literature, requiring scientific reasoning, and (2) all descriptions are aligned with user preferences.

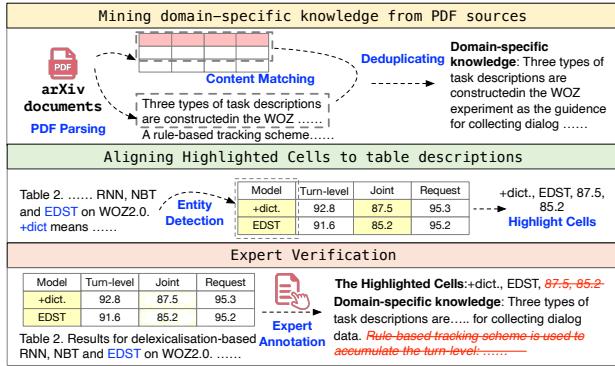We conducted extensive experiments on popular PLM-based

Fig. 2. Overview of CTRLSciTab construction steps, including mining domain-specific knowledge from PDF source, aligning highlighted cells to table descriptions, and expert verification.

models, revealing their poor performance in scientific domains requiring advanced expertise, a persistent challenge for the NLP community. To address this, we propose a retrieval-based pre-trained model trained on our dataset, which shows a better performance. However, evaluations, both automatic and human, indicate that the generated sentences may sometimes contain incorrect and hallucinatory content, highlighting this data set's potential as a valuable benchmark for evaluating controlled table-to-text generation with scientific reasoning.

2. THE CTRLSCITAB DATASET

Data Preparation. CTRLSciTab, a broad dataset, consolidates table-to-text pairs from the [10] and [11] datasets, which originally obtained these pairs from scientific literature with a focus on numerical reasoning. To cater to controlled generation coupled with scientific reasoning, we embarked on a retrieval process for original scientific articles from arXiv.org using tabular metadata. Subsequently, we instituted a procedure to extract sentences associated with tables, thereby constructing a domain-specific knowledge base for scientific reasoning. We also highlighted cells mentioned within table descriptions, serving as prompts instructing PLMs to mirror user preferences. As depicted in Figure 2, the extraction we first transform articles into XML-encoded files. A greedy algorithm is then engaged for content matching, aligning parsed content with tables, and preserving entity-referencing sentences. Deduplication removes any overlap with table descriptions, and entity detection identifies pertinent entities within the descriptions.

Annotation Procedure. Potential inaccuracies in the annotated data, such as irrelevant highlighted cells, are rectified to ensure quality. Expert annotators, who are computer science undergraduates, verify the automatically annotated data. The tasks involve refining domain-specific knowledge and annotating highlighted cells. Annotators eliminate sentences unrelated to tabular data or descriptions, retaining content explicitly stated or logically inferred from the data or descriptions. To quantify the annotation agreement, we randomly selected 100 samples, and two expert annotators achieved a 66.7% agreement rate on the highlighted cells and a 70.6% agreement rate on domain-specific knowledge, thereby underscoring the necessity of domain-specific knowledge.

Data Analysis. CTRLSciTab is a large data set that contains 8,967 controlled table-description pairs with rich domain-specific knowledge in various scientific domains. Each table in CTRLSc-

Data set	Pairs	Cell	Domain	Scientific Reasoning	Controlled
WikiBIO	400K	17	Open	✗	✗
ToTTo	136K	3	Open	✗	✓
LogicNLG	37K	91	Open	✗	✗
SciGen	53K	55	Scientific	✗	✗
numericNLG	1.3K	46	Scientific	✗	✗
CTRLSciTab	9.0K	52	Scientific	✓	✓

Table 1. Comparison of CTRLSciTab and previous table-to-text generation data sets. The pairs represent the number of annotated structure data in each data set. Cell denotes the average number of total cells.

iTab consists of 52 cells on average, with a description of 34 words and highlighted cells as controlled preferences which occupy about 20% of the cells. To support common sense generation, we construct an average of 20 domain-specific knowledge sentences per table.

In the context of table-to-text generation datasets, CTRLSciTab introduces a distinct perspective on content selection and surface realization that distinguishes it from other existing datasets. Prior to the advent of neural approaches, generation systems typically separated content selection (what to say) from surface realization (how to say it) [12]. Traditional generation datasets focused predominantly on a single phase, simplifying the overarching complexity of the task. ToTTo [13] was seminal in presenting a controlled table-to-text generation task with an emphasis on content selection. However, its design does not adequately address the needs of scientific scenarios that require expertise. Some recent datasets have proposed incorporating complex reasoning, such as numerical [11], [10] and logical reasoning [14], into the task by framing it as a summarization problem focused on surface realization. A deeper exploration of the chain of thought [15] in Large Language Models (LLMs) has led to remarkable progress in such generation tasks. CTRLSciTab uniquely challenges generation systems by imposing controlled constraints on content selection while requiring sophisticated scientific reasoning for surface realization.

3. TASK DEFINITION

This task aims to generate natural language descriptions that are both fluent and accurate, incorporating domain-specific knowledge while remaining consistent with the tabular data and user preferences. The input consists of structured data, highlighted cells, and domain-specific knowledge, denoted as $D = (T, H, B)$. Here, T signifies a linearized table, with $T = \{t_1, \cdots, t_{|n|}\}$. Each tabular data, t_i, consists of an attribute-value pair, where a_i and v_i can take values such as strings, numbers, phrases, or sentences. The highlighted cells are akin to t_i and are denoted by $H = \{h_1, \cdots, h_{|n|}\}$, acting as prompts reflecting user preferences. Furthermore, $B = \{b_i, \cdots, b_{|m|}\}$ represents domain-specific knowledge, with each b_i corresponding to a sentence associated with the tabular data. The expected output is an analytical description aligned with user preferences and incorporating domain-specific knowledge R.

4. METHODS

The abundance of supporting domain-specific knowledge for each table, with an average of over 20 sentences, makes it challenging to leverage using popular PLMs. To address this, we propose a retriever-generator framework, CTRLSciTabNet, as our initial approach to the problem. The overall framework of our method is
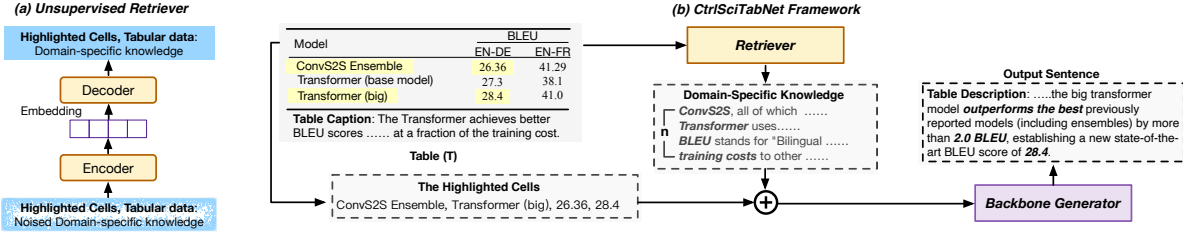
Fig. 3. An illustration of the CTRLSciTabNet structure: (a) depicts the architecture of our unsupervised retriever; (b) outlines the two-step operation of CTRLSciTabNet, which involves a retriever selecting the top-n related domain-relevant sentences, followed by a pre-trained language model, the generator, utilizing this data alongside tabular inputs and highlighted cells.

illustrated in Figure 3. The retriever retrieves supporting domain-specific knowledge based on the table contents, which is then used by the generator to produce descriptions according to the highlighted cells.

Retriever. In this study, as delineated in Figure 3 (a), we introduce an unsupervised sentence embedding technique under controlled conditions. Drawing inspiration from [16], our approach seeks to derive embeddings of domain-specific sentences represented by B through the process of reconstructing them from a perturbed corpus, denoted as $\tilde{B}$. Further refining the embedding process, we impose constraints H and T, signifying user preferences and the original tabular data, respectively. The conditional distribution $P(\tilde{B}|B, T, H)$. The training goal of the retriever is defined by Equation 1:

$$
\begin{aligned}
J_{SDAE}(\theta) &= E_{x \sim D}[\log P_\theta(x|\tilde{x})] \\
&= E_{x \sim D}[\sum_{t=1}^{L} \log P_\theta(x|\tilde{x})] \\
&= E_{x \sim [B:T:H]}[\sum_{t=1}^{l_{|B|}+l_{|T|}+l_{|H|}} \log \frac{exp(h_t^T e_t)}{\sum_{i=1}^{N} exp(h_t^T e_i)})]
\end{aligned}
\tag{1}
$$

where $L = l_{|B|} + l_{|T|} + +l_{|H|}$ denotes the total length of input tokens, including the highlighted cells, table contents, and domain-specific knowledge.

Generator. We aim to develop a sentence generator that can produce coherent and sensible descriptions based on tabular data and the highlighted cells while incorporating domain-specific knowledge according to user preferences. To achieve this goal, we employ BART-base [17] and T5-small [18] for the sentence generator. The learning objective of our sentence generator is to minimize the cross-entropy loss. Specifically, given the tabular data T, highlighted cells H, domain-specific knowledge B, and corresponding table descriptions R, the learning objective is defined as Equation 2:

$$
L_{LM} = -\sum_{i=1}^{|R|} \log P_G(R_i|R_{<i}; E([H:T:B]))
\tag{2}
$$

where E represents the encoder.

5. EXPERIMENTS

Baselines. To evaluate the effectiveness of our approach, we compare it to several popular strategies by replacing the retriever and sequence generator as baselines. Specifically, we consider the following methods:

Baselines	BLEU↑	METEOR↑	BertScore↑	BLEURT↑
TF-IDF + DG (Bart)	1.60	0.09	0.75	-0.93
Retriever + DG (Bart)	2.00	0.09	0.78	-1.00
Retriever + DG (GPT-3.5)	4.76	0.20	0.84	-0.51
CTRLSciTabNet (Bart)	**16.90**	**0.34**	**0.87**	**-0.32**
CTRLSciTabNet (T5)	6.60	0.29	0.85	-0.68
DG w/o BKG (GPT-3.5)	3.07	0.17	0.82	-0.65
Bart w/o BKG	14.90	0.31	**0.87**	-0.38
T5 w/o BKG	2.10	0.18	0.82	-0.87

Table 2. Automatic evaluation results of all methods. The symbol ↑ indicates that a higher score represents better performance. The best result is **bold**, and the second best is underlined within each metric. W/o BKG denotes the model without using domain-specific knowledge. DG denotes direct generate from the pre-trained model without fine-tuning

TF-IDF + Direct Generation (Bart-base). We implement TF-IDF [19], a commonly utilized retrieval method, to access domain-specific knowledge. The architectural design remains consistent with our framework, utilizing the Bart-base model without fine-tuning for the generative tasks.

Retriever + Direct Generation (Bart-base). To assess the efficacy of our proposed retrieval method, we substitute the TF-IDF retriever with our approach, maintaining all other settings.

Retriever + Direct Generation (GPT-3.5). We conduct an additional investigation using our dataset to further explore the functionality of GPT-3.5 [20].

Automatic Evaluation. To evaluate the performance of our model, we use five automatic metrics: BLEU [7], METEOR [8], BertScore [9], and BLEURT [10]. BLEU and METEOR are widely used metrics that measure the informativeness of generated text. BertScore and BLEURT are pre-trained metrics to measure the similarity between the generated descriptions and the reference sentence. Table 2 delivers a thorough assessment of all tested systems using metrics defined in §5. To evaluate the retrievers' performance, we maintain our model's architecture and retrieve the top-3 related domain-specific knowledge sentences using our proposed and TF-IDF-based retrievers. We then generate the corresponding descriptions directly from the retrieved sentences. Our proposed retriever outperforms the TF-IDF-based retriever based on the automatic evaluation metrics. Subsequently, we use the same retriever results as input for all generator models, which allows us to evaluate the performance of the generator part. It is worth noting that the currently popular large-scale GPT-3.5 still does not perform well on tasks that require domain-specific knowledge for scientific reasoning.

In addition, we provide a breakdown of the performance of us-

Model	Fluency↑	Faithfulness↑	Recall↑	Valid Facts↑
Bart w BKG	**4.19**	**0.39**	**0.40**	**0.88**
Bart w/o BKG	<u>4.14</u>	0.31	<u>0.38</u>	<u>0.80</u>
T5 w BKG	3.74	<u>0.33</u>	0.33	0.50
T5 w/o BKG	2.60	<u>0.30</u>	0.24	0.47

Table 3. Performance of human evaluation result of CTRLSciTab-Net using the BART-base and T5-small models based on several metrics, including fluency, faithfulness, recall of covered highlighted cells, and valid factors. ↑ denotes a higher score representing a better performance. The best models are **bold**, and the second best ones are <u>underlined</u> within each evaluation. W/o BKG denotes the model without the use of domain-specific knowledge. W BKG denotes the model trained with domain-specific knowledge.

ing domain-specific knowledge. When we use only the tabular data as input, the performance for all PLMs (GPT-3.5, Bart-base, and T5-small) significantly lags behind the results obtained when external domain-specific knowledge is incorporated. These results demonstrate the importance and necessity of domain-specific knowledge as expertise for scientific reasoning.

Except for BertScore, the performance of all baseline systems is significantly poor in the remaining metrics. Our human evaluation in §5 reveals that the majority of the generated descriptions are of poor quality. Additionally, existing automatic evaluation metrics fail to evaluate the coherence of the highlighted cells and the generated descriptions for controlled table-to-text generation. Furthermore, CTRLSciTabNet, based on the Bart-base model, outperforms the other models on most of the automatic evaluation metrics. These results highlight the inadequacy of automatic evaluation metrics for this challenge.

Human Evaluation. Existing automatic metrics have difficulty assessing the coherence between highlighted cells and generated sentences, requiring human evaluation using four different criteria: fluency, fidelity, recall, and valid facts. Fluency, measured on a five-point Likert scale, assesses the grammatical correctness and naturalness of sentences. Faithfulness measures sentence accuracy against table data, with annotators determining the proportion of faithful content. Recall assesses sentence relevance to highlighted cells, with all sentence parts aligned with cells deemed relevant. Annotators determine the proportion of such facts within the generated descriptions. Finally, Valid Facts determines the proportion of unique and accurate content within generated sentences relative to reference sentences.

Table 3 shows the results of human evaluation, highlighting the superior performance of models using external domain-specific knowledge in terms of fluency, fidelity, recall, and valid facts. The data suggest that advanced domain knowledge is essential for producing scientifically sound results in the scientific domain. Our proposed method, CTRLSciTabNet (Bart-base w BKG), consistently outperforms all evaluated metrics, in line with the results of automatic evaluation metrics.

However, the poor fidelity score indicates a significant gap within the CTRLSciTab dataset that needs to be addressed. In addition, the poor recall of highlighted cells implies that the evaluated models fall short in performing controlled generation independent of scientific reasoning. Both automatic and human evaluation results support CTRLSciTab as a practical testbed for controlled table-to-text generation with scientific reasoning. In particular, CTRLSciTabNet (Bart-base w BKG) shows the best performance across all metrics (sign test with a p-value <0.05).

Model	Positive	Negative	Total
DTN	82.7	86.5	84.6
DNN	84.5	86.0	85.7
RMNN	85.1	87.1	86.1

Table Caption: Comparison of Accuracy (in%) in CN14

Table Description: The CN14 dataset is designed for answering commonsense questions like Is a camel capable of journeying across desert. The proposed NAM models answer this question by calculating the association probability Pr(E2|E1) where E1=\ {camel,capable of\} and E2=journey across desert. We can see that both NAM methods outperform NTN in this task, and the DNN and RMNN models obtain similar performance.

CTRLSciTabNet (Bart-base): As shown in the table, the proposed NAM model significantly outperforms the NTN in both the positive and the negative measures. This results indicate the effectiveness of our proposed model.

CTRLSciTabNet (Bart-base W/o BKG): The ecosystem is in excellent The ecosystem is in excellent The ecosystem is in excellent The ecosystem is in excellent.

Fig. 4. Case study of CTRLSciTabNet. Contents in yellow cells indicate the highlighted cells. W/o BKG denotes the model without the use of domain-specific knowledge. Green text indicates the correct statements supported by the tabular data, and red text indicates the incorrect statements.

Case Study. Figure 4 shows an output generated by the Bart-base model, providing a deeper insight into how our framework works. The text colored in green indicates facts confirmed by tabular data, while the text colored in red indicates incorrect facts. These results suggest that the model trained with domain-specific knowledge outperforms its counterpart without such knowledge, in particular showing significant improvements in fluency and fidelity. In contrast, the model without domain-specific knowledge tends to hallucinate and generate content that is inconsistent with the original tabular data. However, despite these improvements, both models show inadequate performance in capturing highlighted content.

6. CONCLUSION

We introduce a rigorous task, controlled table-to-text generation with scientific reasoning, to assess the machine's ability to produce analytical descriptions that satisfy domainspecific knowledge and user preferences. To support research in this area, we present a novel dataset, CTRLSciTab, which is characterized by two features: (1) its table-description pairs from the scientific literature require domain-specific knowledge for scientific reasoning, and (2) all descriptions are consistent with user preferences. To evaluate the effectiveness of CTRLSciTab, we employ different baselines and perform comprehensive automatic and human evaluations. Our study reveals key findings: (1) existing evaluation metrics inadequately measure controlled table-to-text generation with scientific reasoning; (2) the dominant GPT-3.5 model struggles with tasks requiring advanced expertise; (3) several challenges are associated with the CTRLSciTab dataset. Despite these issues, the initial results lay the groundwork for future research on pre-training tasks for complex, realistic domains. Overall, we position CTRLSciTab as a valuable asset to the research community and expect it to inspire further exploration in this area.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui, "Table-to-text generation by structure-aware seq2seq learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang, "Logical natural language generation from open-domain tables," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7929–7942.

[3] Sam Wiseman, Stuart M Shieber, and Alexander M Rush, "Challenges in data-to-document generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2253–2263.

[4] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das, "Totto: A controlled table-to-text generation dataset," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1173–1186.

[5] Ludovic Denoyer and Patrick Gallinari, "The wikipedia xml corpus," in *ACM SIGIR Forum*. ACM New York, NY, USA, 2006, vol. 40, pp. 64–69.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] Christian D Schunn and John R Anderson, "The generality/specificity of expertise in scientific reasoning," *Cognitive science*, vol. 23, no. 3, pp. 337–370, 1999.

[8] Lei Bao, Tianfan Cai, Kathy Koenig, Kai Fang, Jing Han, Jing Wang, Qing Liu, Lin Ding, Lili Cui, Ying Luo, et al., "Learning and scientific reasoning," *Science*, vol. 323, no. 5914, pp. 586–587, 2009.

[9] Corinne Zimmerman, "The development of scientific reasoning skills," *Developmental review*, vol. 20, no. 1, pp. 99–149, 2000.

[10] Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych, "Scigen: a dataset for reasoning-aware text generation from scientific tables," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[11] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura, "Towards table-to-text generation with numerical reasoning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1451–1465.

[12] Ehud Reiter and Robert Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.

[13] Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das, "Totto: A controlled table-to-text generation dataset," *arXiv preprint arXiv:2004.14373*, 2020.

[14] Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang, "Logical natural language generation from open-domain tables," *arXiv preprint arXiv:2004.10404*, 2020.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[16] Kexin Wang, Nils Reimers, and Iryna Gurevych, "Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 671–688.

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer.," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[19] Akiko Aizawa, "An information-theoretic perspective of tf–idf measures q," *Information Processing and Management*, vol. 39, pp. 45–65, 2003.

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.