Table to Text Generation: A Survey of Datasets and Techniques

Ronak Upasham and **Pushpak Bhattacharyya**
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay
ronakupasham@cse.iitb.ac.in

Abstract

Table-to-text generation is a rapidly evolving task in natural language generation (NLG) that focuses on converting structured tabular data into coherent, fluent, and contextually appropriate natural language text. This task plays a crucial role in making complex, structured information more accessible and interpretable to humans, enabling applications in domains such as financial reporting, sports analytics, scientific documentation, healthcare, and journalism. In recent years, the field has witnessed substantial progress, propelled by the development of large-scale datasets and the application of neural architectures, especially transformer-based models. This survey presents a comprehensive overview of modern table-to-text generation, with a focus on datasets and the state-of-the-art techniques designed to tackle challenges such as content selection, factual consistency, numerical and logical reasoning, and stylistic variation. We categorize existing approaches, analyze their strengths and limitations, and highlight the unique properties of benchmark datasets that have shaped research in this area. Finally, we discuss emerging trends and future research directions aimed at improving the quality, interpretability, and controllability of generated text from structured tables.

1 Introduction

Table-to-text generation is a fundamental task in natural language generation (NLG) that involves converting structured tabular data into fluent and coherent natural language text. It serves as a critical bridge between machine-readable formats and human-consumable narratives, enabling users to interact with structured data through summaries, insights, and explanations. This task finds broad applications across various domains, including financial reporting, sports analytics, weather forecasting, scientific documentation, medical records, and public policy reporting. The ability to generate meaningful textual descriptions from structured data is essential for democratizing access to information, enhancing data communication, and supporting automated decision-making systems.

Historically, early systems for table-to-text generation relied on rule-based and template-driven approaches, where handcrafted rules governed content selection and surface realization. While these systems provided high precision in narrow domains, they lacked flexibility, scalability, and adaptability across diverse data schemas. The advent of neural models, particularly sequence-to-sequence architectures, significantly transformed the field. Models such as the Transformer and its variants have demonstrated remarkable success in learning complex generation patterns, enabling more fluent and domain-adaptive outputs. These models, especially when combined with large-scale pretraining, have opened new avenues for zero-shot and few-shot generation across previously unseen tables and domains.

In recent years, the field has witnessed a surge in benchmark datasets and modeling innovations. Datasets such as LogicNLG, SciGen, FinQA, and WikiTableT have expanded the scope of the task by incorporating elements such as numerical and logical reasoning, long-form generation, and multi-domain coverage. Alongside, contemporary techniques have aimed to improve content planning, factual consistency, controllability, and stylistic variation in generated texts. Despite these advances, challenges such as hallucination, insufficient reasoning capabilities, and generalization to unseen schemas persist. This survey provides a comprehensive overview of the evolution of table-to-text generation, with a focus on modern datasets and techniques, while also contextualizing them within the broader historical development of the field.

2 Motivation

The vast amount of structured tabular data generated across various domains presents a significant challenge in terms of accessibility, interpretability, and usability. While tables serve as an efficient format for organizing and storing structured information, they often require domain expertise to analyze and interpret effectively. Table-to-Text generation plays a crucial role in bridging this gap by transforming complex tabular data into coherent, human-readable narratives. This capability has far-reaching implications across multiple domains, making information more accessible to laypersons, enhancing data-driven decision-making, and improving the overall communication of structured data.

One of the most impactful applications of T2T generation is in the domain of scientific reporting, where large datasets are used to present experimental results, statistical findings, and research insights. Automating the generation of textual summaries from tables in research papers and technical reports can significantly reduce the cognitive load on readers while ensuring that critical information is effectively conveyed. Similarly, financial documents such as balance sheets, income statements, and cash flow reports contain intricate numerical data that require expert interpretation. T2T generation can be employed to produce natural language summaries highlighting key financial trends, risks, and performance indicators, making financial reports more comprehensible to stakeholders and investors.

In the field of sports analytics, tables summarizing match statistics, player performances, and historical trends can be automatically transformed into engaging textual reports. This enables media outlets, analysts, and fans to quickly grasp match outcomes, strategic insights, and player comparisons without manually analyzing raw statistics. Similarly, weather tables that contain temperature trends, precipitation levels, and forecast probabilities can be converted into easily understandable weather reports, aiding in effective communication with the general public.

T2T generation is also highly relevant in news reporting, where structured datasets are often used to summarize economic trends, stock market movements, election results, and demographic statistics. Automating the transformation of structured data into news articles allows for more efficient reporting while reducing manual effort. Likewise, demographic tables, such as population census data, can be converted into detailed textual reports that highlight key patterns in population growth, age distributions, and regional disparities.

In the healthcare and medical domain, patient records, test results, and epidemiological data are often represented in tabular form. Converting such structured medical data into textual summaries can assist doctors in making informed decisions, improve patient understanding of their health conditions, and facilitate better communication in clinical settings. Additionally, legal documents, which often contain structured data on case proceedings, regulatory compliance, and legal precedents, can benefit from T2T techniques by generating comprehensive summaries that improve accessibility for both legal professionals and the general public.

Beyond simplifying complex data for laypersons, T2T generation plays a vital role in extracting insights from structured data. By automatically generating text that highlights trends, anomalies, and correlations within tables, these systems can assist in data-driven decision-making across multiple industries. Moreover, improved natural language descriptions enhance data communication, enabling businesses, researchers, and policymakers to present structured information in a format that is easily digestible by a broader audience.

Given the increasing reliance on structured data across industries, the ability to transform tabular information into meaningful text holds immense potential. This survey aims to explore the latest advancements in Table-to-Text generation, shedding light on the methodologies that drive this transformation and the challenges that remain to be addressed.

3 Datasets

A fundamental aspect of T2T generation research is the availability of high-quality datasets that enable model training, evaluation, and benchmarking. These datasets vary in complexity, domain coverage, and data structure, ranging from structured tables with annotated textual descriptions to datasets that require complex reasoning and logical inference. Over the years, several datasets have been introduced to support advancements in T2T generation, covering diverse applications such as question answering, data summarization, and analytical reporting.

Recent datasets have focused on improving generalization across unseen tables, handling diverse schema representations, and minimizing hallucinations in generated text. Additionally, many datasets incorporate real-world structured data, ensuring that models are trained on realistic and practical tabular information. In this section, we review key datasets that have been widely used for Table-to-Text generation, highlighting their unique characteristics and contributions to the field.

3.1 WikiSQL

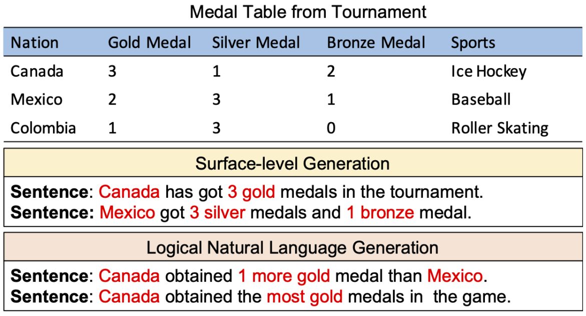
WikiSQL (Zhong et al., 2017) is a large-scale dataset designed for semantic parsing and structured data-to-text generation, consisting of 80,654 hand-annotated examples derived from 24,241 tables extracted from Wikipedia. It provides a unique challenge for models as it requires generalization not only across different queries but also across unseen table schemas. Each example in WikiSQL includes a table, an SQL query, and a natural language question that corresponds to the SQL query.

Unlike traditional T2T datasets that focus primarily on textual descriptions, WikiSQL emphasizes structured query understanding and text generation from logical forms. The dataset was created through a crowd-sourcing approach on Amazon Mechanical Turk, ensuring linguistic diversity in question formulations. The dataset's scale and complexity make it a valuable benchmark for evaluating models on structured data understanding and natural language generation.

A distinctive feature of WikiSQL is its broad coverage of table structures and domains, extracted from real-world web data. The dataset challenges models to generate coherent responses based on tabular content while maintaining syntactic and semantic consistency. Its focus on SQL-based logical forms has also made it a widely used resource in tasks involving structured data question answering and schema generalization.

3.2 SciGen

SciGen (Moosavi et al., 2021) is a dataset designed to evaluate reasoning-aware data-to-text generation, specifically targeting the description of scientific tables. Unlike traditional datasets that primarily focus on surface-level realization of tabular content, SciGen requires models to perform arithmetic reasoning over table values. The dataset consists of tables extracted from scientific articles in the computer science domain, paired with corresponding textual descriptions that incorporate numerical computations such as *argMax*, *argMin*, comparison, and subtraction.



Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level Generation

Sentence: Canada has got 3 gold medals in the tournament.
Sentence: Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation

Sentence: Canada obtained 1 more gold medal than Mexico.
Sentence: Canada obtained the most gold medals in the game.

Figure 1: An example from the LogicNLG dataset.

A key characteristic of SciGen is that the tables predominantly contain numerical data, and their textual descriptions involve reasoning beyond simple restatement. This makes SciGen the first dataset that explicitly assesses the arithmetic reasoning capabilities of generation models when dealing with complex input structures. The ability to generate scientifically coherent text from experimental results or numerical tables is crucial for applications such as automated scientific writing and specialized chatbots that can interpret and explain structured data.

SciGen was created by selecting tables from scientific papers where the descriptions were derived through arithmetic operations on table values. This ensures that models trained on SciGen must develop the capability to generate text that goes beyond surface-level synthesis, making it an important benchmark for evaluating reasoning-based table-to-text generation. The dataset plays a crucial role in advancing the field by encouraging research on generating factually grounded and logically consistent text from structured numerical data.

3.3 LogicNLG

LogicNLG (Chen et al., 2020b) is a dataset designed to advance logical inference in table-to-text generation. Unlike traditional datasets that primarily focus on surface-level realizations of table content, LogicNLG emphasizes diversified logical reasoning, including mathematical operations (e.g., max, min, sum), comparison operations (e.g., same, different), and counting-based inferences (e.g., total, only).

The dataset is constructed based on TabFact, a table-based fact-checking dataset that contains logical inferences in annotated statements. Specifically, LogicNLG extracts positive statements (i.e., statements entailed by table knowledge) from TabFact's complex channel, where sentences require logical inference. The dataset consists of 28,450 training instances, 4,260 validation instances, and 4,305 test instances, covering 7,392 open-domain tables sourced from Wikipedia. Each table is paired with five distinct examples incorporating diverse types of logical reasoning.

Each generated sentence requires some form of logical reasoning while minimizing reliance on domain-specific knowledge. This open-domain setting ensures that models cannot rely on predefined inference rules, pushing for better generalization. The dataset primarily consists of short sentences with an average length of 11 words and simple syntactic structures, isolating logical inference as the primary challenge rather than linguistic complexity. The tables originate from various domains such as sports, politics, and entertainment, making rule-based approaches infeasible and requiring models to generalize across different table structures.

LogicNLG provides a robust benchmark for evaluating the logical inference capabilities of table-to-text generation models. It is particularly useful for applications requiring numerically and logically consistent text generation, such as automated analytics reporting and fact-based summarization.

3.4 NumericNLG

NumericNLG (Suadaa et al., 2021)is a dataset designed for numerical table-to-text generation, with a strong emphasis on numerical reasoning and rich inference. Unlike standard Table-to-Text datasets, NumericNLG focuses on generating scientifically coherent text that involves complex reasoning over numerical values present in tables. The dataset consists of table-paragraph pairs, where the textual descriptions are naturally written by experts in scientific papers, ensuring high linguistic quality and domain relevance.

NumericNLG was constructed by extracting numerical tables of experimental results from scientific papers available on the ACL Anthology. The corresponding textual descriptions were collected from the source files using automated extraction methods, ensuring alignment between the table content and the generated text. The dataset provides a unique challenge as the descriptions require deeper inference beyond simple surface realization, making it an essential benchmark for evaluating models on numerical reasoning and structured data understanding.

By leveraging real-world scientific writing, NumericNLG enables research on generating high-quality, numerically accurate descriptions of tables. This dataset is particularly valuable for applications such as automated scientific reporting, where models must not only summarize data but also provide insightful interpretations based on the numerical values presented.

3.5 Logic2Text

Logic2Text (Alonso and Agirre, 2024) is a large-scale dataset designed to enhance logical reasoning in table-to-text generation. It comprises 10,753 descriptions involving common logical types, each paired with an underlying logical form. Unlike previous datasets, Logic2Text introduces logical forms with diverse graph structures and free-schema representations, posing significant challenges for models in terms of semantic understanding.

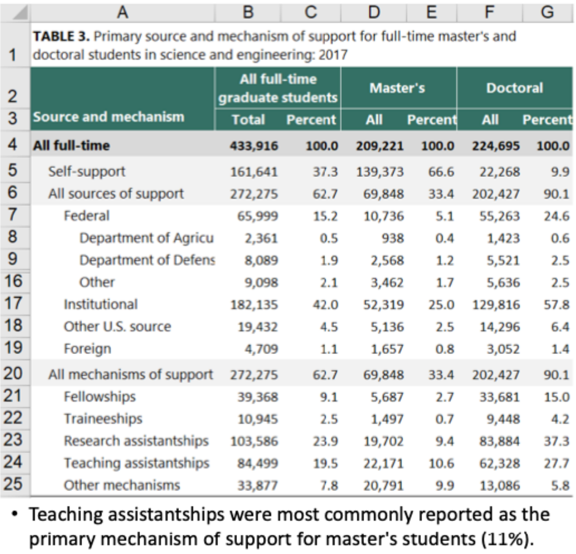
The dataset consists of 5,600 open-domain tables and 10,800 manually annotated (logical form, description) pairs. It is sourced from WikiTables, a collection of open-domain tables crawled from Wikipedia. Over-complicated tables are filtered out, retaining only those with fewer than 20 rows and 10 columns.

The dataset provides natural and informative descriptions, with logical forms that achieve 100% execution correctness. The dataset includes seven coarse-grained logic types commonly used for describing multi-row tables: *count, superlative, comparative, aggregation, majority, unique,* and *ordinal.* A Python-like program represents the logical forms, making them easily convertible to other formal representations.

Each description in Logic2Text involves exactly one type of logic, aligning with human tendencies to describe tabular data using clear and concise logical statements rather than overly complex reasoning chains. This makes Logic2Text an ideal benchmark for evaluating a model's ability to generate text based on structured logical inference.

3.6 HiTab

HiTab (Cheng et al., 2021) is a dataset designed to study natural language generation (NLG) over hierarchical tables. Unlike conventional flat tables, hierarchical tables present a complex structure with multi-level headers and implicit relationships, making table-to-text generation particularly challenging. The dataset is constructed from statistical re-



	A	B	C	D	E	F	G
1	**TABLE 3.** Primary source and mechanism of support for full-time master's and doctoral students in science and engineering: 2017						
2		All full-time graduate students		Master's		Doctoral	
3	Source and mechanism	Total	Percent	All	Percent	All	Percent
4	**All full-time**	**433,916**	**100.0**	**209,221**	**100.0**	**224,695**	**100.0**
5	Self-support	161,641	37.3	139,373	66.6	22,268	9.9
6	All sources of support	272,275	62.7	69,848	33.4	202,427	90.1
7	Federal	65,999	15.2	10,736	5.1	55,263	24.6
8	Department of Agricu	2,361	0.5	938	0.4	1,423	0.6
9	Department of Defens	8,089	1.9	2,568	1.2	5,521	2.5
16	Other	9,098	2.1	3,462	1.7	5,636	2.5
17	Institutional	182,135	42.0	52,319	25.0	129,816	57.8
18	Other U.S. source	19,432	4.5	5,136	2.5	14,296	6.4
19	Foreign	4,709	1.1	1,657	0.8	3,052	1.4
20	All mechanisms of support	272,275	62.7	69,848	33.4	202,427	90.1
21	Fellowships	39,368	9.1	5,687	2.7	33,681	15.0
22	Traineeships	10,945	2.5	1,497	0.7	9,448	4.2
23	Research assistantships	103,586	23.9	19,702	9.4	83,884	37.3
24	Teaching assistantships	84,499	19.5	22,171	10.6	62,328	27.7
25	Other mechanisms	33,877	7.8	20,791	9.9	13,086	5.8

- Teaching assistantships were most commonly reported as the primary mechanism of support for master's students (11%).

Figure 2: An example from the HiTab dataset.

ports and Wikipedia pages, ensuring that the textual descriptions are meaningful and naturally occurring rather than artificially created. To facilitate a deeper understanding of numerical reasoning, HiTab also provides fine-grained annotations of quantity and entity alignment.

A key challenge in NLG from hierarchical tables arises from hierarchical indexing, where table headers span multiple levels both horizontally and vertically. Generating textual descriptions from such tables requires models to correctly interpret multi-dimensional indexing and accurately reference numerical values. Additionally, hierarchical tables often contain implicit calculation relationships, such as aggregated totals and proportions, which are not explicitly marked. This necessitates the development of models capable of inferring hidden numerical dependencies and generating coherent descriptions.

Another major challenge involves identifying implicit semantic relationships among entities. Hierarchical structures frequently encode cross-row and cross-column dependencies without clear indications, requiring the model to infer logical connections between related data points. The presence of virtual entities further complicates generation, as descriptions may need to synthesize multiple hierarchical elements into a single coherent narrative. HiTab serves as a benchmark for studying these challenges, pushing advancements in table-to-text generation with complex structured data.

3.7 WikiTableT

WikiTableT (Chen et al., 2020a) is a large-scale dataset designed for multi-sentence data-to-text generation, where Wikipedia sections are generated based on corresponding tabular data and metadata. Unlike traditional datasets that focus on either multi-domain, single-sentence generation or single-domain, long-form generation, WikiTableT bridges this gap by covering a diverse range of topics with varying levels of generation flexibility. The dataset pairs structured tabular data from Wikipedia infoboxes, Wikidata tables, and section-specific data sources with textual descriptions, enabling the study of structured text generation in a realistic setting.

One of the key challenges posed by WikiTableT is handling different styles of generation. Some instances require flexible text generation, where models must construct coherent narratives incorporating background knowledge beyond the table. For example, generating a fictional character biography from comic book data requires linking entities within the section data while ensuring consistency with article metadata. Other instances align more closely with traditional data-to-text tasks, where the table contains all necessary information, and the goal is to produce a faithful textual representation without additional context.

The dataset is constructed through a combination of automated data extraction and filtering to maintain high quality. It provides a rich benchmark for evaluating models on complex table-to-text generation scenarios, including structured content planning, entity linking, and coherence in multi-sentence generation. By offering a vast number of instances across diverse domains, WikiTableT advances research in scalable and flexible natural language generation from structured data.

3.8 LoTNLG

LoTNLG (Zhao et al., 2023) is a dataset designed to enhance table-to-text generation by conditioning models on specific logical reasoning types. Traditional table-to-text models, such as those applied to the LogicNLG dataset, often generate insights that are biased towards a limited set of logical reasoning operations. For instance, GPT-3.5 tends to prioritize numerical comparisons while overlooking other relevant insights present in the table. This lack of diversity in generated statements restricts the informativeness of data insight generation, as users typically seek multiple perspectives on tabular data. LoTNLG addresses this limitation by explicitly guiding models to produce statements using a broader range of logical reasoning operations.

To construct LoTNLG, nine common logical reasoning operations were predefined, including aggregation, negation, superlative, count, comparative, ordinal, unique, universal quantification, and surface-level operations. Each statement from the LOGICNLG test set was annotated with up to two logical reasoning types, ensuring high-quality labeling through a multi-stage annotation process. This approach allows LoTNLG to serve as a benchmark for evaluating a model's ability to generate diverse and logically grounded insights from structured data.

By conditioning text generation on specific logical reasoning operations, LoTNLG encourages the development of models that can generate more comprehensive and varied insights. This structured approach enhances the ability of table-to-text systems to provide richer and more informative textual summaries, making it particularly useful for applications requiring in-depth data interpretation and analysis.

3.9 FinQA

FinQA (Chen et al., 2021) is a dataset designed to facilitate deep question-answering over financial data, with a focus on automating financial document analysis. Unlike general-domain QA tasks, financial data analysis requires complex numerical reasoning and an understanding of heterogeneous tabular and textual representations. FinQA consists of 8,281 expert-annotated question-answer pairs extracted from the earnings reports of S&P 500 companies. Each QA pair is supplemented with a detailed reasoning program that ensures explainability by explicitly outlining the steps required to derive the correct answer. This makes FinQA particularly relevant for table-to-text generation tasks that involve analytical summarization of financial statements.

The dataset is constructed using earnings reports from the FinTabNet corpus, which provides annotated tables from financial documents spanning 1999 to 2019. Since not all financial tables are suitable for numerical reasoning tasks, a filtering process is applied to exclude complex or unstructured tables. Only tables with at most one description header and a manageable number of rows are retained, ensuring that the dataset remains focused on structured reasoning tasks. By maintaining a well-defined scope, FinQA enables models to generate insightful textual summaries that integrate both numerical computations and explanatory content.

As an application for table-to-text generation, FinQA presents unique challenges in generating analytical summaries that require accurate financial calculations and reasoning. The dataset's focus on real-world financial documents makes it highly valuable for developing models that can generate coherent and factually grounded narratives from tabular data. This capability is essential for financial analysts, business intelligence systems, and automated report generation tools that seek to extract meaningful insights from structured financial records.

3.10 TabFact

TabFact (Chen et al., 2019) is a large-scale dataset designed for fact verification using semi-structured tables as evidence, containing 16K Wikipedia tables and 118K manually annotated natural language statements classified as either *ENTAILED* or *REFUTED*. Unlike traditional question-answering datasets, TabFact requires both linguistic and symbolic reasoning to verify the factual consistency of statements with tabular data. Linguistic reasoning involves understanding paraphrased expressions and implicit meanings, while symbolic reasoning requires performing logical and arithmetic operations over table structures. This dual requirement makes TabFact an important resource for evaluating models that generate or validate textual statements from tabular data.

The dataset construction involved filtering complex web tables from WikiTables to retain those with a manageable number of rows and columns, ensuring a balance between linguistic complexity and structural clarity. Statements were crowdsourced via Amazon Mechanical Turk, where annotators generated entailed and refuted claims based on the tables. To maintain high annotation quality, a structured pipeline was employed, including positive instance selection, negative statement rewriting, and verification steps. The resulting dataset captures diverse reasoning patterns, with each table corresponding to multiple statements that challenge models to discern factual correctness.

TabFact is highly relevant for table-to-text generation, particularly in tasks requiring the generation of factually accurate and verifiable textual summaries. Since table-based text generation often risks producing hallucinated content, the integration of a fact verification mechanism trained on TabFact can improve the reliability of generated text. Additionally, the dataset's structured nature allows for the development of models that not only generate coherent descriptions from tables but also verify their factual correctness, making it valuable for applications such as automated report generation and fact-checking systems.

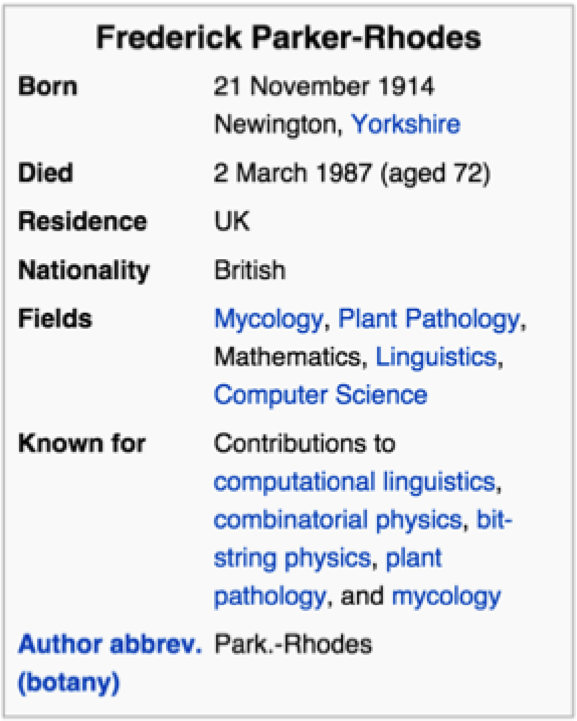
4 Techniques

In this section, we discuss key techniques employed in the task of table-to-text generation, focusing on how different architectures and modeling paradigms have evolved over time to tackle challenges such as content selection, fluency, factual correctness, and reasoning over structured inputs. Broadly, these techniques can be categorized into two major approaches: traditional neural architectures trained specifically for this task, and the more recent use of large language models (LLMs) through prompting or fine-tuning. We present each category in its own subsection, highlighting representative works, their core contributions, and limitations.

4.1 Neural Approaches

Prior to the widespread adoption of large language models, table-to-text generation was predominantly addressed using neural encoder-decoder architectures. These models, often trained end-to-end on parallel table-text datasets, focused on representing structured inputs through various encoding schemes such as flat sequences, hierarchical embeddings, or graph representations. Challenges in vocabulary generalization, content selection, and rare entity handling led to the development of specialized mechanisms such as pointer networks, copy actions, and dual-level conditioning. In this subsection, we review representative works that laid the foundation for modern table-to-text generation by innovating on neural modeling techniques and dataset scale.

One of the earliest influential works in neural table-to-text generation was proposed by (Lebret et al., 2016), which focused on generating biographical sentences from Wikipedia infoboxes as illustrated in Figure 3. The model leveraged both global and local conditioning to generate the first sentence of a biography based on structured attributes such as birthdate, occupation, and nationality. Global conditioning enabled the model to capture high-level semantic information, e.g., identifying whether the person was an athlete or actor, while local conditioning helped align generated words with specific fields in the input table.



Frederick Parker-Rhodes

Born	21 November 1914 Newington, Yorkshire
Died	2 March 1987 (aged 72)
Residence	UK
Nationality	British
Fields	Mycology, Plant Pathology, Mathematics, Linguistics, Computer Science
Known for	Contributions to computational linguistics, combinatorial physics, bit-string physics, plant pathology, and mycology
Author abbrev. (botany)	Park.-Rhodes

Figure 3: An example of Wikipedia Infobox table.

To handle the large vocabulary size typical of Wikipedia-scale data, the model incorporated a copy mechanism, allowing it to select out-of-vocabulary tokens directly from the table. This hybrid vocabulary approach addressed data sparsity and improved factual alignment. Moreover, the model embedded words differently depending on their field types and token positions, capturing structural nuances of tabular inputs. Compared to traditional count-based language models, the proposed method achieved a significant improvement of nearly 15 BLEU points.

This work is notable for its scalability, training on over 700k infobox samples with a vocabulary exceeding 400k words, and for its emphasis on conditioning mechanisms that later became standard components in neural generation models. Despite its effectiveness, the paper acknowledged limitations in handling factual consistency and multi-sentence generation, identifying these as avenues for future improvement. The authors also highlighted the need for evaluation metrics beyond surface-level overlap (e.g., BLEU) to assess factual accuracy more directly.

To jointly address the tasks of content selection and surface realization, (Mei et al., 2015) introduced an end-to-end encoder-aligner-decoder architecture that employs a novel coarse-to-fine alignment mechanism. The model is designed to generate natural language summaries from overdetermined databases, such as weather records or sportscasting logs, where only a subset of the input records are relevant for generating a coherent and informative description. This makes the task particularly suitable for scenarios where structured tables contain more data than what should be verbalized, mirroring real-world data-to-text settings.

The proposed architecture uses bidirectional LSTMs to encode the full set of records and then leverages a two-stage aligner, first, a coarse preselector to shortlist relevant records, followed by a fine-grained refiner to perform token-level alignment during decoding. This alignment mechanism allows the model to dynamically attend to the most salient entries at each generation step. Crucially, the system achieves state-of-the-art performance on the WEATHERGOV dataset, improving BLEU scores by 59% without any hand-crafted templates or linguistic resources. The model also generalizes well to new domains, such as the ROBOCUP dataset, indicating its robustness and adaptability. This work is foundational in demonstrating the feasibility of fully neural architectures for selective generation and highlighting the importance of modular attention mechanisms.

Addressing the need to model both the content and structural layout of tables, (Liu et al., 2018) proposed a structure-aware sequence-to-sequence model that combines field-gating mechanisms with dual-level attention. The model is specifically designed for generating biographical descriptions from Wikipedia infoboxes, utilizing the WIKIBIO dataset of over 700k biographies. The key insight in their approach is to distinguish between local and global addressing: local addressing determines which words within a field-value pair to focus on, while global addressing guides the model toward the most relevant fields to include in the output.

To implement this, the model uses a field-gated encoder where each LSTM cell's memory state is modulated by the associated field embedding, allowing the encoder to retain structural context. On the decoding side, a dual attention mechanism, comprising word-level attention for local alignment and field-level attention for global focus, is used to align the generated tokens with both word and field representations. This design enables the model to better capture hierarchical relationships and ordering variations within tables, which are often ignored in flat encodings. The authors show significant improvements over previous baselines on the WIKIBIO dataset and provide attention visualizations that support the model's interpretability. Their method exemplifies how structural priors can be deeply integrated into neural architectures to enhance generation quality.

4.2 LLM-based Approaches

With the advent of large pretrained language models (LLMs), table-to-text generation has witnessed a shift toward architectures that unify structured and unstructured data processing through transformer-based encoders. Unlike traditional neural approaches that often rely on explicit alignment, copying mechanisms, or handcrafted encodings, these newer models are trained on vast corpora of text and tabular data, allowing them to generalize across domains, aggregation tasks, and reasoning styles. Many of these models do not require full logical forms or schema-level supervision, enabling scalability and simplification of the training pipeline. In this subsection, we highlight key LLM-based approaches that have advanced the field through pretraining on large-scale table-text pairs, weak supervision, and differentiable reasoning frameworks.

TAPAS (Herzig et al., 2020) represents a major step forward in using pretrained transformer architectures for reasoning over tables. Unlike traditional semantic parsers that convert natural language questions into logical forms, TAPAS eliminates the intermediate formal step altogether. It directly maps a question-table pair to an answer by selecting relevant table cells and optionally applying aggregation functions such as COUNT, SUM, or AVERAGE. This enables TAPAS to handle a wider variety of question types while simplifying the overall architecture and training pipeline.

The model builds on BERT by introducing table-specific embeddings that capture the tabular layout, column headers, row positions, and segment associations. During pretraining, TAPAS leverages a masked language modeling objective applied jointly to Wikipedia tables and accompanying text, learning to encode both formats into a unified representation. For fine-tuning, TAPAS uses a differentiable, weakly supervised training

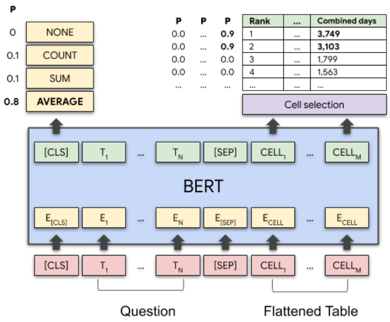


Figure 4: Architecture of the TAPAS model.

strategy: when answers involve direct selection, the model is trained to highlight the correct cells, and when aggregation is involved, it regresses to the gold answer by modeling an expected value over aggregation functions.

TAPAS achieves state-of-the-art or competitive performance on three major datasets: SQA, WIKISQL, and WIKITQ. Notably, it improves accuracy on SQA from 55.1 to 67.2 and achieves effective transfer learning from WIKISQL to WIKITQ. By avoiding symbolic programs and adopting a table-aware pretraining strategy, TAPAS illustrates how LLMs can be effectively adapted for table-based question answering and factual text generation tasks, setting a precedent for subsequent models that further blur the line between structured and unstructured data handling.

TaBERT (Yin et al., 2020) addresses the challenge of reasoning jointly over free-form natural language and structured tabular data by extending BERT with mechanisms specifically designed for tabular structure. Unlike standard BERT models trained only on natural language, TaBERT is pretrained on 26 million tables paired with corresponding English text, enabling it to encode both textual and tabular modalities into a shared representation space.

To represent structured tables within the transformer framework, TaBERT linearizes tables and introduces a novel *content snapshot* mechanism, which selects a semantically relevant subset of table rows for encoding based on the input utterance. This helps mitigate computational costs and prevents the model from being overwhelmed by large tables. Further, a *vertical self-attention* module is used to enable interactions across different rows while maintaining the structured integrity of the table.

TaBERT is designed to be a plug-and-play encoder for semantic parsing systems, providing contextual embeddings for both utterance tokens and table schema components such as columns and cell values. By aligning the representations of natural language queries with structured tables through task-agnostic pretraining, TaBERT offers a generalized and reusable model for table-based language understanding and generation tasks.

In the work of (Zhao et al., 2023), authors present a comprehensive evaluation of several large language models (LLMs), including GPT-4, LLaMA-2, and Vicuna, on table-to-text generation across two real-world information-seeking settings: (1) *data insight generation* and (2) *query-based text generation*. These scenarios mimic how users often interact with structured data to derive summaries or find answers.

The authors introduce two new benchmark datasets: **LOTNLG** (for insight-driven logical summarization) and **F2WTQ** (for free-form question answering on Wikipedia tables), complementing existing ones like LOGICNLG and FeTaQA. These datasets are designed to test the models' abilities in both logical reasoning and factual language generation.

Their findings highlight that GPT-4 consistently outperforms other LLMs in generating fluent and factually faithful text from tables. It also shows strong performance as a reference-free evaluator using chain-of-thought prompting and as a feedback generator for post-editing model outputs. In contrast, open-source LLMs like LLaMA-2 and Vicuna lag significantly behind, particularly on logically complex or multi-row reasoning tasks.

Importantly, the study also explores LLMs' potential to aid other systems by providing natural language feedback. GPT-4, when prompted appropriately, can identify factual inconsistencies in generated text and offer corrected versions, making it a valuable tool not just for generation but also for evaluation and improvement of other models.

This work marks a key step in understanding how LLMs can be integrated into practical systems for real-world table understanding and summarization, highlighting their strengths and limitations in terms of factuality, fluency, and reasoning.

5 Challenges

Despite significant advancements in table-to-text generation, the task remains fraught with several persistent and emerging challenges. These challenges span linguistic, structural, and modeling dimensions, and limit the deployment of robust, general-purpose systems.

1. Hallucination and Factual Consistency. One of the most critical issues in table-to-text generation is *hallucination*, where models generate text that is fluent but factually incorrect or unsupported by the input table. This is particularly problematic in high-stakes domains like finance or healthcare. Models often struggle to restrict themselves to only the content present in the table, leading to spurious or exaggerated claims.

2. Subjectivity and Stylistic Control. While early approaches aimed for factual summarization, recent trends explore incorporating subjectivity, such as emotional tone, opinions, or persuasive elements, into the generated text. However, balancing subjectivity with factuality remains an open problem. Controlling the tone, sentiment, or viewpoint in a systematic and interpretable way is still underdeveloped.

3. Domain Generalization and Adaptability. Many models are trained and evaluated on domain-specific datasets (e.g., weather, sports, biographies) and fail to generalize effectively to unseen domains. The strong reliance on structural patterns seen during training often limits adaptability. Pretrained LLMs offer some improvements, but they may still lack robustness when faced with tables from new domains or with schema variability.

4. Long-Table and Multi-Row Reasoning. Real-world tables can be large, with hundreds of rows and multiple interrelated records. Generating coherent summaries that require reasoning over multiple rows, identifying trends, or computing aggregates remains a major challenge. Memory and attention bottlenecks in current models often force truncation or sampling strategies that sacrifice context.

5. Alignment and Content Selection. Determining which parts of a table to include in the output, especially when the table is over-complete, is a non-trivial problem. Jointly learning *content selection* and *surface realization* in a controllable and explainable manner is still an open area of research. Misalignment between selected content and the generated text is a frequent source of errors.

6. Evaluation Metrics. Standard automatic metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or METEOR (Banerjee and Lavie, 2005) are insufficient for evaluating factuality, coherence, or usefulness in table-to-text generation. While recent studies explore reference-free metrics and human-in-the-loop evaluations, developing reliable and domain-agnostic evaluation frameworks remains a key challenge.

7. Data Scarcity and Annotation Bottlenecks. Training effective models often requires large-scale paired datasets of tables and high-quality textual summaries, which are expensive and time-consuming to curate. Weak supervision and synthetic data generation are partial solutions but come with trade-offs in quality and realism.

6 Conclusion

Table-to-text generation stands at the intersection of structured data understanding and natural language generation, offering immense potential for data summarization, user assistance, and information accessibility. Over the past decade, the field has evolved from early rule-based systems to neural models with selective generation and structure-aware mechanisms, and most recently to large language models capable of zero-shot generalization.

This survey has traced the trajectory of major approaches, spanning classical pipeline methods, sequence-to-sequence architectures, structure-aware enhancements, and LLM-based paradigms. Alongside the methodological innovations, we have also highlighted the use of diverse datasets that evaluate various aspects such as factuality, fluency, reasoning, and subjectivity.

However, the task remains open-ended and challenging. Persistent issues such as hallucination, poor content selection, and the need for human-like reasoning continue to hinder progress. Furthermore, as real-world applications demand more personalized, subjective, and context-aware text, future work must focus on controllable generation, robust evaluation, and better domain adaptation.

In conclusion, while current models show promising results in constrained settings, building trustworthy, generalizable, and user-centric table-to-text generation systems requires addressing both foundational challenges and emerging expectations, and there is a large scope for future work in this domain.

References

Iñigo Alonso and Eneko Agirre. 2024. Automatic logical forms improve fidelity in table-to-text generation. *Expert Systems with Applications*, 238:121869.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2020a. Wikitablet: A large-scale data-to-text dataset for generating wikipedia article sections. *arXiv preprint arXiv:2012.14919*.

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

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2021. Hitab: A hierarchical table dataset for question answering and natural language generation. *arXiv preprint arXiv:2108.06712*.

Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Hongyuan Mei, Mohit Bansal, and Matthew R Walter. 2015. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. *arXiv preprint arXiv:1509.00838*.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. 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)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. 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)*, pages 1451–1465.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175.

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