Error Analysis of ToTTo Table-to-Text Neural NLG Models

Barkavi Sundararajan and **Somayajulu Sripada** and **Ehud Reiter**
Department of Computing Science, University of Aberdeen
{b.sundararajan.21, yaji.sripada, e.reiter}@abdn.ac.uk

Abstract

We report error analysis of outputs from four Table-to-Text generation models fine-tuned on ToTTo, an open-domain English language dataset. We carried out a manual error annotation of a subset of outputs (a total of 3,016 sentences) belonging to the topic of *Politics* generated by these four models. Our error annotation focused on eight categories of errors. The error analysis shows that more than 46% of sentences from each of the four models have been error-free. It uncovered some of the specific classes of errors; for example, *WORD* errors (mostly verbs and prepositions) are the dominant errors in all four models and are the most complex ones among other errors. *NAME* (mostly nouns) and *NUMBER* errors are slightly higher in two of the GeM benchmark models, whereas *DATE_DIMENSION* and *OTHER* categories of errors are more common in our Table-to-Text model. This in-depth error analysis is currently guiding us in improving our Table-to-Text model.

1 Introduction

End-to-end neural Table-to-Text models produce outputs that suffer from hallucination (output texts contain parts that are not supported by input data). This may be because these models learn the noise from complex examples during the training process and produce more errors than rule-based systems (Rebuffel et al., 2021). The automatic metrics such as BLEU and ROUGE do not uncover common classes of errors and are therefore less helpful to improve the models (Gehrmann et al., 2021a). The accuracy evaluation shared task by Thomson and Reiter (2021) using the gold standard methodology proposed by Thomson and Reiter (2020) was successful in identifying errors that are difficult to detect using automatic metrics (Gehrmann et al., 2022).

In this paper, we performed a detailed error analysis, adopting the Thomson and Reiter (2020) methodology on four Table-to-Text model outputs (trained on the ToTTo dataset) to identify and group the errors these models make in the output text. We created one of these Table-to-Text models by fine-tuning a t5-base Text-to-Text model with the ToTTo dataset using BLEU as a validation metric with the standard cross-entropy objective function, and we will be applying error corrections to this model in our future work. The other three model outputs came from GEM benchmark Table-to-Text models fine-tuned from t5-small, t5-base, and t5-large Text-to-Text models (Gehrmann et al., 2021a). Previous research studies for error analysis predominantly focused on Machine Translation (MT) systems using a simple framework by (Stymne and Ahrenberg, 2012) and extensively using Multidimensional Quality Metrics (MQM) framework by Freitag et al. (2021), and on other NLG tasks (Cai et al., 2020; Thomson and Reiter, 2020). Other annotation methods have been used to check for errors in the ToTTo dataset. Yin and Wan (2022) use a method based on MQM, assigning multiple labels to individual sentences. We take a different approach, annotating at the more granular level of token spans, i.e., words or phrases.

2 Table-to-Text

Table-to-Text generation is an important and challenging task in Natural Language Generation (NLG), which focuses on producing a factual, meaningful, and fluent output from structured tabular data. Most domains (viz. journalism, medical diagnosis, sports broadcasting and weather reports) are data-rich, and the information required for critical decision-making in these domains comes from the dataset, which is better represented as textual narratives than represented in structures such as indexes, tables and key-value pairs (Rebuffel et al., 2019).

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 456 - 470

Input Table data from ToTTo (includes metadata such as Title, Highlighted cells in yellow and their headers):

Page Title: List of ambassadors of the United States to Germany

Section Title: Heads of the U.S. Embassy at Bonn (1955–1999)

Name and Title	Presentation of Credentials	Termination of Mission
James B. Conant, Ambassador	May 14, 1955	February 19, 1957
David K. E. Bruce, Ambassador	April 17, 1957	October 29, 1959
Walter C. Dowling, Ambassador	December 3, 1959	April 21, 1963
George C. McGhee, Ambassador	May 18, 1963	May 21, 1968

Input Source/Linearized representation of the above Input Table data in Text format:

<page_title> List of ambassadors of the United States to Germany </page_title> <section_title> Heads of the U.S. Embassy at Bonn (1955–1999) </section_title> <table> <cell> David K. E. Bruce, Ambassador <col_header> Name and Title </col_header> </cell> <cell> April 17, 1957 <col_header> Presentation of Credentials </col_header> </cell> <cell> October 29, 1959 <col_header> Termination of Mission </col_header> </cell> </table>

Output/Reference Text faithful to the above Input Table data in Text format:

David K. E. Bruce served as the United States Ambassador to Germany from April 17, 1957 to October 29, 1959.

Table 1: Input Table sample from ToTTo (Parikh et al., 2020), Linearized representation of the Input Table data and Reference Text

2.1 ToTTo

Table 1 shows an example from a controlled Table-to-Text generation dataset (ToTTo) (Parikh et al., 2020) where a subset of the cells from the Wikipedia tables are taken as Input and paired with a relevant sentence description from the same Wikipedia page. This dataset was created using crowdsourcing to mark relevant cells from the table (shown in yellow) along with their corresponding row and column headers as inputs (removing the need for the content selection sub-task followed in the rule-based NLG systems).

As part of the Input Table data in Table 1, Page Title, Section Title and Section Text (if available) are called metadata. The ToTTo Table-to-Text task is to fine-tune neural NLG models to auto-generate output texts that describe the highlighted table cells along with their metadata faithfully and are similar (similarity measured using both automatic metrics as well as human judgement) to the reference text(s), example, the Reference Text in Table 1. Training Table-to-Text models with the more controlled ToTTo training dataset is expected to generate high-quality outputs (Parikh et al., 2020) because it focuses on addressing a simplified task instead of end-to-end Table-to-Text.

The ToTTo dataset covers a diverse distribution of topics such as Sports, Politics, Entertainment, Literature, Performing Arts, Broadcasting and so on. This dataset helps to understand how the generations differ for each domain and accordingly identify any pattern of errors made by the models. This level of insight would be helpful to improve our Table-to-Text model.

The ToTTo dataset has three splits based on the 83,141 unique Wikipedia tables: i. **Train** with 120,761 samples, ii. **Validation** with 7,700 samples and iii. **Test** with 7,700 samples[1].

The validation and test split samples are further categorised into overlap and non-overlap. Overlap split refers to the data (i.e. set of header values) already seen in training samples. Non-overlap split refers to the set of header values not seen in the training split and increases the generalization challenge (Parikh et al., 2020). In the validation split, we have 3784 overlap and 3916 non-overlap samples as further discussed in section 4.1[2].

2.2 Linearized Representation of the Input Table data

The relevant contents from the Input Table data as mentioned in section 2.1 are converted to the linearized representation i.e., metadata and highlighted texts with headers as mentioned in Table 1 for each of the training samples. The Reference Text is already 'in the text' format. The pre-trained transformer model we used takes the source-reference (input-output) pairs in a Text-to-Text format. Hence, the preprocessed Input Table data in Text format (linearized representation) and

[1]Since the output of each sample in ToTTo generates only one sentence as output, we used the term *sentence(s)* and *sample(s)* interchangeably in this paper.

[2]The test split reference outputs are not open-source and not considered in our error analysis.

its corresponding Reference Text, as shown in Table 1 for one sample, has been applied for all the training samples (120,761).

3 Table-To-Text Models included in our Error Analysis

3.1 Our Model

The Text-to-Text Transfer Transformer (T5) model pre-trained on Colossal Clean Crawled Corpus (C4) (Raffel et al., 2019) is taken to fine-tune our model with the ToTTo dataset. The T5 model is said to outperform GPT-2 and BERT models and is robust to handle out-of-domain inputs (Kale and Rastogi, 2020). The linearized representation of the Input Table data and its Reference Text pair, as shown in Table 1 and elaborated in section 2.2, are used for fine-tuning our Table-to-Text task.

The **M1: BLEU** model is a Table-to-Text model created by us by fine-tuning t5-base Text-to-Text models (220 million parameters) with the ToTTo dataset using BLEU as validation metric with the standard cross entropy objective function. The input or the encoder's maximum length of our model is 512 tokens to align with the limit of the pre-trained models. It is fine-tuned with a constant learning rate of 0.0001 and a beam size of 10 to generate the target text with at most 128 tokens (i.e., the decoder's maximum length). The batch size used for this M1: BLEU model is 2 and trained on a commodity server with GeForce RTX 2080 Ti with 11G memory using Single-precision Floating-point format (FP32). It took approximately seven days to train this model for 180,800 training steps.

3.2 GeM Benchmark Models

The error analysis also uses outputs from three GeM benchmark (Gehrmann et al., 2021a) Table-to-Text models that are fine-tuned from t5-small (GM2), t5-base (GM3) and t5-large (GM4) Text-to-Text models. These three variants of pre-trained t5 models come in different sizes. **GM2: t5-small** is pre-trained with 60 million parameters, **GM3: t5-base** is pre-trained with 220 million parameters and **GM4: t5-large** is pre-trained with 770 million parameters. Other specific fine-tuning or configuration details of these three benchmark models are unknown. In contrast, since we know these fine-tuning and configuration details for our model as described in section 3.1, the error analysis reported in this paper could be exploited to improve our model in future.

4 Evaluation and Results

4.1 Metric Based Evaluation

The best practice for evaluation choices (Gehrmann et al., 2021b) is to use a combination of metrics from at least two different categories. Hence, the scores of four Table-to-Text model outputs are computed using different types of metrics such as BLEURT (Sellam et al., 2020) and BERTScore (Zhang* et al., 2020) for **semantic** measure, BLEU (Papineni et al., 2002), ROUGE-2 (Ganesan, 2015) and METEOR (Banerjee and Lavie, 2005) for **lexical** measure and PARENT metric (Dhingra et al., 2019) that is relevant for **Table-to-Text** systems.

Table 9 and Table 11 in Appendix A shows the metric scores for the overall validation set of ToTTo (7,700 samples). The overall scores imply that GM3: t5-base (benchmark model) has the best scores for BLEU, BLEURT, ROUGE2, BERTScore and METEOR. Whereas our model M1: BLEU has the best score for PARENT (overall).

Table 10 and Table 12 in Appendix B shows the metric scores for the Politics domain (754 out of the 7,700 samples) in the validation set. The best scores slightly differ for this domain, where GM3: t5-base (benchmark model) scored well only for the overlap samples of BLEU, ROUGE2, BERTScore and METEOR. Whereas GM4: t5-large (benchmark model) has a better score for the non-overlap (challenging samples) for BLEU, BLEURT, BERTScore and METEOR. Our model M1: BLEU scored well for PARENT (both in overall and non-overlap samples).

These scores do not provide complete guidance on the actual performance of the neural models and cannot measure factual accuracy. To verify the performance of the system, we carried out a manual error analysis by focusing on eight categories of errors for the *Politics* domain (because this topic covers only 4% of the ToTTo data which is easier to error annotate) as detailed in section 4.2.

4.2 Human Evaluation

Performing a human evaluation in Amazon Mechanical Turk through crowdsourcing is expensive and is also time-consuming to screen with a qualification task before the actual experiment. Thomson and Reiter (2020) proposed a gold standard methodology for evaluating similar Table-To-Text tasks. We adapted this gold standard evaluation technique for the ToTTo dataset. The annotation procedure is

discussed in section 4.2.1, and some examples are detailed in section 4.3 and Appendix C.

4.2.1 Error Categories for Annotation

Below are the eight categories of errors we used for annotating *Politics* domain outputs in ToTTo.

- **WORD**[W]: when incorrect words such as verbs, prepositions, adjectives and adverbs are found in the output.
- **NAME**[N]: when names of the Party, Leader, place (Electorate), Ambassador etc., are wrong (mostly nouns).
- **DATE_DIMENSION**[D]: when the Date and/or Month and/or Year are wrong.
- **NUMBER**[U]: when the number of seats and/or the number of votes and/or % of votes are incorrect.
- **OTHER**[O]: It includes mistakes in any of the below sub-categories.
  - GRAMMATICAL: when simple grammatical mistakes are identified in the output text. For example, missing articles such as 'a', 'the', and 'an', and the linking verb used for singular pronouns such as 'is' and 'was'. Any other verb mistakes belong to the WORD error. Other complex grammatical mistakes are not considered.
  - PUNCTUATION: when punctuation symbols are placed at inappropriate places, an apostrophe is missed for the Name of the Leader or Place.
  - GARBAGE: when the table data has the Politics party name in the abbreviation, it tries to produce garbage output.
  - Unclear: when the information is unclear.
- **CONTEXT**[C]: when the people will misunderstand a sentence i.e., the generated sentence is misleading, given the input data.
- **NOT-CHECKABLE**[X]: when the output has details that are not available in the Input Table data (i.e., relevant contents such as metadata, highlighted cells and their headers). The information may be right, but it requires checking other online resources to validate.
- **NON-ENGLISH**[NE]: when the Unicode characters in non-English names are either replaced with special characters or when these Unicode characters are omitted.

Our annotation scheme differs from Thomson and Reiter (2020) in terms of how the Date, Month and Year are handled. We introduced **DATE_DIMENSION**[D] category for ToTTo as the specific *Politics* domain had Date, Month and Year errors. There are also more **NON-ENGLISH**[NE] errors in the Unicode characters for the NAMES of a leader, place and/or party.

4.2.2 Other points for annotation

A single distinct token (i.e., word) is marked by highlighting that specific span of text for WORD, NAME, DATE_DIMENSION, NUMBER, OTHER (except GARBAGE sub-category) and NON-ENGLISH errors. For CONTEXT, OTHER-GARBAGE and NOT-CHECKABLE category of errors, it is difficult to reliably identify distinct tokens and therefore a group of tokens or relevant span of text can be marked as shown in the example annotations in Table 20, Table 21 and table 22 in Appendix C.

4.3 Results

Following the annotation guidelines defined in section 4.2.1, Table 2 and Table 3 provide the results of the manual error analysis made. Table 2 shows the overview of the error analysis identified in all four Table-to-Text models for a subset of 754 samples (Politics domain).

- **'NO ERROR'**: Around 46% of the samples out of this subset is error-free in all four models.
- **'OMISSIONS'**: Around 29% to 32% of the samples had omissions. However, these samples are not further analysed in the Table 3 (Individual Error Annotation count) due to the difficulty in objectively annotating omissions. We will independently study *Omissions* category of annotations in our future work. If a particular sample has both omission and error, the preference is given to the error alone, and its corresponding error count is included only in the *Errors Annotated* category in Table 2.
- **'META-DATA ISSUES'**: Around 6% of the samples required changes to the input records (Table metadata, cells and header) i.e., few

Category	M1: BLEU		GM2: t5-small		GM3: t5-base		GM4: t5-large	
	Count	%	Count	%	Count	%	Count	%
NO ERROR	346	46	355	47	371	49	387	51
OMISSIONS	244	32	218	29	240	32	232	31
META-DATA ISSUES*	46	6	43	6	40	5	37	5
ERRORS ANNOTATED	118	16	138	18	103	14	98	13
TOTAL COUNT	**754**		**754**		**754**		**754**	

Table 2: Sample/Sentence Count: Error Analysis of the model outputs for Politics domain of ToTTo. This table has the count of the samples with errors. Meta-data issues* are either i. when the right cells from table are not passed to the Input Data, or ii. when irrelevant cells (not highlighted in yellow) are passed as Input Data for few complex table structure.

Category	M1: BLEU	GM2: t5-small	GM3: t5-base	GM4: t5-large
WORD	63	74	49	47
NAME	14	23	24	10
DATE_DIMENSION	12	15	9	0
NUMBER	10	7	12	10
OTHER	10	11	6	6
CONTEXT	8	10	3	5
NOT-CHECKABLE	2	3	2	3
NON-ENGLISH	20	19	19	22
TOTAL ERROR COUNT	**139**	**162**	**124**	**103**

Table 3: Individual Error Annotation Count based on the *Errors Annotated* category taken from table 2. This table has the count of individual errors annotated from the samples. Hence, the total error count is higher in this table than table 2 (since each sample/sentence can contain multiple errors).

samples did not have the exact cell highlighted in the Table (compared to the Reference sentence), and few other samples had irrelevant cells passed in the Input. These samples are excluded from the Table 3 error annotations.

- **'ERRORS ANNOTATED'**: The error annotations in all four models ranged between 13% and 18%. This category is the main focus of this paper and is restricted to the eight categories of errors as presented in the Table 3.

Table 3 error analysis uncovered eight common classes of errors in all four models, which is elaborated further in each subsection along with examples [3].

4.3.1 WORD^{W} errors

This error is the dominant one committed by all four models. Our model had more WORD errors than two GeM Benchmark models (GM3: t5-base and GM4: t5-large). They belong to the below sub-groups.

- Most of them are VERB errors such as '*defeated*' versus '*succeeded*', '*won*' versus '*lost*', '*elected*' versus '*contested*', '*appointed*' versus '*nominated*' and so on.
- Some of them include errors in prepositions such as '*from*', '*with*', '*by*', '*to*', '*until*' and so on.
- Few errors are specific to the Politics related words. For example, '*swing*' that has a positive or negative percentage versus '*normal percentage of votes*'.

In Table 4, all four models made the WORD error. The word 'longest-lived' is the main error where the sentence semantic requires access to other data to compute the right word i.e., longest-lived or shortest-lived. The input header only has the term 'longevity' and could be the reason for all models to generalise it as longest-lived. 'Sanj Sanetomi' is a NON-ENGLISH error that is uniform in

[3] For better readability, the reference sentence and correct prediction sentence that have the right token without any errors are either marked with the superscript R (example, right-token^{R}) or in green colour (example, **right-token**) in the example annotations (i.e., in section 4.3 and Appendix C).

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: List of Japanese prime ministers by longevity
Section Title: Longevity of Japanese prime ministers

No	Prime Minister	Total time in Office	Date of Death	Lifespan
3	Sanetomi Sanjō	60 days	28 Feb 1891	53 years, 352 days

(b) **Output**: Prediction with Error Annotations and Reference

M1: BLEU prediction: Sanj Sanetomi[NE] was the longest-lived[W] prime minister in Japan at 53 years, 352 days.

Reference: The shortest-lived[R] Prime Minister was Sanetomi Sanjō, who died at the age of 53 years and 352 days.

Table 4: WORD error made by all models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: Members of the Australian House of Representatives
Section Title: None

Member	Party	Electorate	State	Term in Office
Hon Les Johnson	Labor	Hughes	NSW	1955-1966, 1969-1984
Robert Tickner	Labor	Hughes	NSW	1984-1996

(b) **Output**: Prediction with Error Annotations and Reference

GM3: t5-base prediction: Hughes Labor MHA[N] Les Johnson served until 1984 and Robert Tickner served from 1984 to 1996.

Reference: The Labor member for Hughes (NSW), Les Johnson in 1984 was replaced by Labor candidate Robert Tickner as **Member of the Australian House of Representatives**[R].

Table 5: NAME error made by our model and two benchmark models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title:1901 Bulgarian parliamentary election
Section Title: Results

Party	Votes	%	Seats	±
People's Party	66,910	22.8	29	+27
People's Liberal Party	65,680	22.4	24	+5
Progressive Liberal Party	58,860	20.1	40	+32
Total	344,787	100	164	-5

(b) **Output**: Prediction with Error Annotations and Reference

GM3: t5-base prediction: The result was a victory for the Progressive Liberal Party, which won **58,860**[U] of the 164 seats.

Reference: Despite receiving the third highest number of votes, the Progressive Liberal Party emerged as the largest party in the 1901 Bulgarian parliamentary election with **40**[R] of the 164 seats.

Table 6: NUMBER error made by all models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: 67th United States Congress
Section Title: Senate

Senator	Reason for Vacancy	Successor	Date of Successor's Installation
Philander C. Knox (R)	Died October 12, 1921. Successor was appointed.	William E. Crow (R)	October 24, 1921
William E. Crow (R)	Died August 2, 1922. Successor was appointed and subsequently elected.	David A. Reed (R)	August 8, 1922

(b) **Output**: Prediction with Error Annotations and Reference

M1: BLEU prediction: Philander C. Knox (R) served as senator until October 12, 1921 which was later served by William E. Crow, Jr. from October 24, **1922**[D] to August 2, **1923**[D] and later succeeded by David A. Reed, R, from August 8, **1932**[D].

Reference: Philander C. Knox (R) served until October 12, 1921 and William E. Crow (R) served from **October 24, 1921**[R] until August 2, **1922**[R], after whom David A. Reed (R) served from August 8, **1922**[R].

Table 7: DATE_DIMENSION error made by our model

all four models because of the tokenization used in t5-base transformers. One solution could be to use a multilingual tokenizer in our future work.

In Table 14, all models except GM4: t5-large could not predict the right WORD 'lost to' or the 'right leader name' for the verb *'defeated'*. One of the reasons the GeM t5-large model predicted the right WORD could be from the huge parameters it learned during pre-training. Being certain about the configuration details of our Table-to-Text model (M1: BLEU), it does not have any arithmetic reasoning abilities to predict the right verb *'defeated'* or *'lost to'* based on the number of votes.

In Table 13, the more obvious reason for WORD errors *'from'* and *'to'* is because the input table had both years (1987 and 1992) in a single field. One reason for this error is that the table does not comply with the First Normal Form (1NF) in relational databases (one of the common patterns identified in our analysis). Another reason is that the table does not have access to the additional insight semantic of 're-elected' as mentioned in the Reference.

4.3.2 NAME^{N} and NUMBER^{U} errors

These two errors are slightly higher in the Benchmark models (GM3 and GM4) than in the model (M1: BLEU).

NAME in the prediction got jumbled/swapped when two or more names are present in the Input Table data as shown in Table 15 in Appendix C. The Table 5 shows NAME hallucinations, where the GeM benchmark models (GM2: t5-small and GM3: t5-base) and our model (M1: BLEU) hallucinated the NAME '*MHA i.e., Member of the House of Assembly*' instead of the right NAME '*Member of the Australian House of Representatives*'.

NUMBER is a common error made by all models, where *'number (or %) of seats'* versus *'number (or %) of votes'* got swapped in the prediction as shown in Table 6. For other cases, the Table 16 and Table 17 show NUMBER hallucinations, where the two GeM benchmark models tried to hallucinate and compute the incorrect margin of votes even when the input table data did not explicitly pass this value.

4.3.3 DATE_DIMENSION^{D} errors

DATE_DIMENSION errors are more common in our model. As shown in Table 7, our model had the year hallucinated even when the right values (i.e., date dimension fields) are passed to the input data. GeM Benchmark models did not face this error except for few complex samples. Even when it had multiple date-dimension fields as shown in Table 18 in Appendix C, the GeM models predicted the year correctly but they committed a different error category (NAME error) in this example. The date-dimension errors will be the first class of errors we intend to address when we improve our model.

4.3.4 OTHER^{O} errors

This error is slightly higher in our model and GM2: t5small model. A few of the miscellaneous errors we encountered in all four models are missing apostrophe (***'s***), missing article ('the', 'a') and other spans of text that does not imply any meaning, for example, ***'GSSSDULSVDHSS'*** as shown in Table 20 in Appendix C. Table 21, Table 22 and Table 19 in Appendix C present the error annotations for the remaining three errors (CONTEXT, NOT-CHECKABLE and NON-ENGLISH).

4.4 Agreement between annotators

One of the authors (the first annotator) annotated 754 predicted outputs for four systems (i.e., 3,016 sentences). In addition, the second annotator annotated predicted outputs for a random 10% of the Politics domain of ToTTo by following the defined guidelines.

The second annotator was given a word document with the screenshot of the Input Table data with highlighted cells as shown in Table 1 excluding the Linearized representation of the input and Reference Text (to focus the attention of the Annotator 2 on the underlying table to annotate errors rather than being guided by the Reference Text). We provided four model outputs for each Input Table data. The second annotator annotated 80 predicted outputs (i.e., 320 sentences for four models), and it took approximately 5 hours to complete this experiment. The annotator marked each error with the corresponding category and provided remarks/corrections where ever possible.

The confusion matrix for the annotations made between the first annotator (A1) and second annotator (A2), along with Cohen's Kappa coefficient (K value), is presented in Table 8. The NUMBER, DATE_DIMENSION, CONTEXT, and NON-ENGLISH categories have a complete agreement. We have a high agreement for both the WORD (the most dominant error) and the NAME errors. NOT-CHECKABLE errors tend to be subjective and have a weak agreement. The average

Category	Both agree: error	Both agree: no error	A1-error	A2-error	K value
WORD	25	245	7	6	0.77
NAME	7	245	0	4	0.77
DATE_DIMENSION	4	245	0	0	1
NUMBER	8	245	0	0	1
OTHER	2	245	0	0	1
CONTEXT	3	245	0	1	0.80
NOT-CHECKABLE	2	245	0	4	0.49
NON-ENGLISH	7	245	0	0	1
TOTAL COUNT AND AVERAGE K VALUE	**58**	**245**	**7**	**15**	**0.79**

Table 8: Cohen's Kappa coefficient (K value): Confusion Matrix for the agreement between two annotators

Cohen's Kappa coefficient (K value) is 0.79, indicating a high agreement between two annotators.

5 Related Work

In the field of Machine Translation, error analysis has been carried out for a long time (Stymne and Ahrenberg, 2012). More recently, the Multidimensional Quality Metrics (MQM) framework based on a hierarchy of errors was applied to carry out error analysis of WMT data (Freitag et al., 2021). This analysis identified error types (error classes) responsible for the difference in output quality between human and machine-generated translations.

Within the NLG context, Cai et al. (2020) performed error analysis for the Topic-to-Essay NLG task and proposed a human annotation framework for evaluating sub-sentence grammar, sentence logic, repetition, semantic coherence and contextual consistency. Their experiment results show that the neural models produced relatively high semantic errors compared to the grammatical and repetition errors.

Within the Table-to-Text context, Thomson and Reiter (2020) designed a gold-standard error analysis methodology with a taxonomy of simple errors for the annotators to evaluate the factual accuracy of Table-to-Text NLG models. They apply this methodology to system-generated basketball summaries from the Rotowire dataset (Wiseman et al., 2017). Thomson and Reiter (2021) described a shared task where different evaluation techniques for basketball summaries from the Rotowire dataset were submitted and their results show that metric-based techniques struggled to detect factual errors.

van Miltenburg et al. (2021) suggested expanding Wiseman et al. (2017) taxonomy to include other taxonomies such as the SCARECROW annotation schema (Dou et al., 2021) and image captioning specific taxonomy by van Miltenburg and Elliott (2017), making the resulting expanded taxonomy aligned to the quality criteria recommended by Belz et al. (2020). van Miltenburg et al. (2021) emphasised avoiding complex terms such as 'hallucinations' and 'omissions' for error categories because these are process-level (system) rather than product-level (output) descriptions of the errors. Analysing the errors using process-level descriptions cannot be reliable. We, therefore, adhered to the simple category of errors based on product-level (output) descriptions in our error analysis.

6 Conclusion

We fine-tuned our neural Table-to-Text model (M1: BLEU) with the known configuration details and compared its outputs with the GeM benchmark model outputs. This analysis provided additional insights of error classes (such as incorrect VERB predictions for WORD errors, NAME and NUMBER swap when two or more of these details are in the Input Table, hallucinations for NUMBER and DATE_DIMENSION), which is not possible to determine from evaluation metric scores. Our analysis shows that these four Transformer based models can perform textual reasoning to some extent but lack a deeper level reasoning capabilities (for example, mathematical reasoning and for more complex table structure when multiple inputs are present). This level of insights from the manual error analysis will provide opportunities to overcome this reasoning capabilities in our model in the future work.

Limitations

This error analysis is focused only on the Politics domain of the ToTTo dataset. It needs to be expanded to other domains such as Sports, Arts, Entertainment and others from the ToTTo dataset. This is the first stage of our error analysis and is restricted to the eight common classes (categories) of errors. Omission-related errors need to be better developed with different severity levels, and meta-data issues have to be corrected.

Ethics Statement

This work seeks to perform error analysis for our model and three benchmark models from GeM, which are trained using the open-source ToTTo dataset. The ground-truth generation remains the same as the original dataset, and we did not introduce any further social bias to this dataset. The three benchmark model outputs are open-source and downloaded from GeM (`https://gem-benchmark.com/resources`). We did not modify any of these three model outputs while annotating errors in our experiment. We sought consent from our second annotator (Craig Thomson), who was provided with the necessary guidelines before performing the error annotations.

Acknowledgements

We thank Craig Thomson for helping with the annotations in this work. We also thank the three anonymous reviewers and the NLG (CLAN) reading group at the University of Aberdeen for their valuable feedback.

References

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.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Ping Cai, Xingyuan Chen, Hongjun Wang, and Peng Jin. 2020. The errors analysis of natural language generation — a case study of topic-to-essay generation. In *2020 16th International Conference on Computational Intelligence and Security (CIS)*, pages 86–89.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Kavita Ganesan. 2015. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021a. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2021b. A case for better evaluation standards in nlg.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of*

the 13th International Conference on Natural Language Generation, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

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.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2021. Controlling hallucinations at word level in data-to-text generation.

Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2019. A hierarchical model for data-to-text generation.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1785–1790, Istanbul, Turkey. European Language Resources Association (ELRA).

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson and Ehud Reiter. 2021. Generation challenges: Results of the accuracy evaluation shared task. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 240–248, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Emiel van Miltenburg and Desmond Elliott. 2017. Room for improvement in automatic image description: an error analysis.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Xunjian Yin and Xiaojun Wan. 2022. How do Seq2Seq models perform on end-to-end data-to-text generation? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7701–7710, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Appendices

A Standard Metrics Evaluation for ToTTo

The standard metrics computed for the ToTTo dataset are BLEU, PARENT and BLEURT. Table 9 is computed for the overall validation dataset and Table 10 is specifically computed for the Politics domain of ToTTo.

B Additional Metric Evaluation for ToTTo

We computed three additional metrics (BERTScore, METEOR and ROUGE2) in Table 11 and Table 12. BERTScore is taken from the official repository (Zhang* et al., 2020), METEOR and ROUGE2 metrics are taken from `https://huggingface.co/` datasets library. Best performing metric scores are in bold.

C Example Annotations

Example annotations for different types of error categories are annotated in this section, as per the guidelines defined in section 4.2.1 for the Politics domain of the ToTTo dataset.

Model	Overall			Overlap			Non-Overlap		
	BLEU	PARENT	BLEURT	BLEU	PARENT	BLEURT	BLEU	PARENT	BLEURT
M1: BLEU	45.9	**56.49**	0.1539	53.2	**61.04**	0.2705	38.4	52.10	0.0412
GM2: t5-small	43.7	54.46	0.1203	51.0	58.43	0.2376	36.6	50.63	0.0070
GM3: t5-base	**46.2**	56.20	**0.1651**	**54.0**	60.33	**0.2773**	**38.7**	**52.20**	**0.0566**
GM4: t5-large	44.7	55.28	0.1434	52.5	59.73	0.2591	37.2	50.97	0.0316

Table 9: Standard Metric Evaluation for the overall ToTTo Validation dataset

Model	Overall			Overlap			Non-Overlap		
	BLEU	PARENT	BLEURT	BLEU	PARENT	BLEURT	BLEU	PARENT	BLEURT
M1: BLEU	49.5	**59.67**	0.1370	55.0	**63.12**	0.2130	41.7	**55.08**	0.0362
GM2: t5-small	48.0	57.80	0.1530	53.0	60.77	0.2372	41.0	53.17	0.0413
GM3: t5-base	**49.9**	57.80	0.1518	**55.9**	61.16	0.2334	41.3	53.33	0.0435
GM4: t5-large	49.6	57.39	**0.1635**	54.6	60.94	**0.2416**	**42.7**	52.67	**0.0598**

Table 10: Standard Metric Evaluation for Politics Domain of ToTTo Validation dataset

Model	Overall			Overlap			Non-Overlap		
	BERT-Score	METEOR	ROUGE2	BERT-Score	METEOR	ROUGE2	BERT-Score	METEOR	ROUGE2
M1: BLEU	0.9330	0.6145	0.4713	**0.9418**	0.6697	0.5398	0.9246	0.5611	0.405
GM2: t5-small	0.9295	0.5972	0.4562	0.938	0.6488	0.5187	0.9212	0.5474	0.3956
GM3: t5-base	**0.9332**	**0.6189**	**0.4767**	0.9415	**0.6707**	**0.5433**	**0.9252**	**0.5688**	**0.4123**
GM4: t5-large	0.9318	0.6151	0.4674	0.9404	0.6689	0.5359	0.9234	0.5631	0.4012

Table 11: Additional Metric Evaluation for the overall ToTTo Validation dataset

Model	Overall			Overlap			Non-Overlap		
	BERT-Score	METEOR	ROUGE2	BERT-Score	METEOR	ROUGE2	BERT-Score	METEOR	ROUGE2
M1: BLEU	0.9364	0.6295	0.4841	0.9434	0.6672	0.5269	0.9280	0.5793	0.4272
GM2: t5-small	0.9361	0.6162	0.4669	0.9428	0.648	0.5055	0.9272	0.5739	0.4168
GM3: t5-base	**0.9372**	**0.6304**	**0.4858**	**0.9445**	**0.6704**	**0.5298**	0.9275	0.5774	**0.4276**
GM4: t5-large	0.9369	0.6293	0.4758	0.9435	0.6650	0.5135	**0.9283**	**0.5820**	0.4267

Table 12: Additional Metric Evaluation for Politics Domain of ToTTo Validation dataset

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: List of heads of state of Trinidad and Tobago

Section Title: President of Trinidad and Tobago

No	President (Birth - Death)	Took Office	Left Office	Elected
2	Noor Hassanali (1918 - 2006)	20 March 1987	17 March 1997	1987, 1992

(b) **Output**: Prediction with Error Annotations and Reference

GM3: t5-base prediction: Noor Hassanali (1918 – 2006) was the President of Trinidad and Tobago from^{W} 1987 to^{W} 1992.

Possible correction: from^{W} - in ; to^{W} - and in

Reference: Noor Hassanali was re-elected^{R} as the president in 1992.

Table 13: WORD error made by all four models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: Dan Maffei

Section Title: Electoral history

Party	Candidate	Votes	%	±
Republican	Ann Marie Buerkle	104,374	50.2	+ 8.0
Democratic	Dan Maffei	103,807	49.8	-4.7
Majority		567	0.4	-10.4
Turnout		208,181	100	+23.4

(b) **Output**: Predictions with Error Annotations and Reference

M1: BLEU prediction: Maffei defeated[W] Republican Ann Marie Buerkle by a margin of 567 votes or[W] 208,181 votes.

GM3: t5-base prediction: Maffei defeated[W] Republican Ann Marie Buerkle by a margin of 567 votes, a difference[W] of 208,181 votes.

GM4: t5-large prediction: Maffei lost to[R] Republican Ann Marie Buerkle by a 567-vote margin with 103,807 votes out of[R] 208,181 votes cast[R].

Reference: **Republican Ann Marie Buerkle defeated Maffei** with a 567-vote majority **of over 200,000 ballots cast**.

Table 14: WORD error made by all models except GM4: t5-large

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: Canterbury, New Zealand

Section Title: None

Members of Parliament for Canterbury

Electorate	Established	MP	Party
Rangitata	2008	Andrew Falloon	National
Selwyn	2008	Amy Adams	National
Waitaki	2008	Jacqui Dean	National

(b) **Output**: Predictions with Error Annotations and Reference

M1: BLEU prediction: The National MPs for Canterbury are Amy Adams for Selwyn and Jacqui Dean for Waitaki and Rangitata[N].

GM2: t5-small prediction: The National Party has represented Rangitata, Selwyn, Amy Adams[N], Waitaki and Jacqui Dean[N].

GM4: t5-large prediction: National's **Amy Adams of Selwyn**[R] and **Jacqui Dean of Waitaki**[R] were the MPs for Canterbury.

Reference: Rangitata and Selwyn are held by National MP's where **Selwyn is represented by Amy Adams**[R] while **Waitaki is represented by Jacqui Dean**[R].

Table 15: NAME error made by all models except GM4: t5-large

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: 2010 Wellington City mayoral election

Section Title: Result

2010 Wellington City mayoral election

Candidate	Count-1	Count-2	Count-3	Count-4	Count-5
Celia Wade-Brown	18,560	18,675	19,002	21,422	24,881
Kerry Prendergast (incumbent)	21,809	21,863	22,021	22,899	24,705
Jack Yan	5,817	5,924	6,144	7,426	0 (No count)

(b) **Output**: Predictions with Error Annotations and Reference

GM3: t5-base prediction: Wade-Brown won the 2010 Wellington City mayoral election by a margin of 1,354[U] votes.

GM4: t5-large prediction: Wade-Brown was elected mayor with 24,705[U] votes over a third of a second[O].

M1: BLEU prediction: Wade-Brown received 24,881 votes out of[W] 24,705 in the 2010 Wellington City mayoral election.

Reference: Celia Wade-Brown won the 2010 Wellington City mayoral election in the final round of the vote count by **24,881 to 24,705**[R].

Table 16: NUMBER error made by two benchmark models and WORD error made by other two models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: 2016 United States presidential election in Michigan
Section Title: Results
2016 United States presidential election in Michigan

Party	Candidate	Running mate	Votes	%	Electoral votes
Republican	Donald Trump	Mike Pence	2,279,543	47.50%	16
Democratic	Hillary Clinton	Tim Kaine	2,268,839	47.27%	0

(b) **Output**: Predictions with Error Annotations and Reference

GM4: t5-large prediction: Trump won the state of Michigan with 47.50% of the vote to Clinton's 47.27%, a margin of **17.13%**[U].

GM3: t5-base prediction: Donald Trump won the 2016 United States presidential election in Michigan with 47.50% of the vote to Hillary Clinton's 47.27%, a margin of **2.94%**[U].

M1: BLEU: In the 2016 United States presidential election in Michigan, Trump took 47.50 percent of the vote over Hillary Clinton's **47.27%**[R].

Reference: The 2016 United States presidential election in Michigan was won by Donald Trump with 47.50% of the total votes over Hillary Clinton's **47.27%**[R].

Table 17: NUMBER error made by two benchmark models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: Ashoknagar (Vidhan Sabha constituency)
Section Title: Members of Legislative Assembly

Election Year	Constituency	Name of M.L.A.	Party Affiliation
1967	Ashoknagar	Sadhan Kumar Sen	Communist Party of India
1969		Sadhan Kumar Sen	Communist Party of India
1971		Nani Kar	Communist Party of India (Marxist)
1972		Keshab Chandra Bhattacharya	Independent

(b) **Output**: Predictions with Error Annotations and Reference

M1:BLEU prediction: Sadhan Kumar Sen of CPI won the seat in 1967, 1969, **1971**[D], and Keshab Chandra Bhattacharya of Independent won in 1972.

GM4: t5-large and GM3: t5-base predictions: Ashoknagar was represented by Sadhan Kumar Sen of **CPI(M)**[N] in 1967 and 1969, Nani Kar of CPI(M) in 1971, and Keshab Chandra Bhattacharya of Independent in 1972.

Reference: Keshab Chandra Bhattacharya of Independent, won in 1972, **Nani Kar of CPI(M) won in 1971**[R], and Sadhan Kumar Sen of **CPI**[R] won in 1969 and 1967.

Table 18: DATE_DIMENSION error made by our model
NAME errors made by other two benchmark models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: List of political parties in Norway
Section Title: Parties currently in Parliament

Name	Ideology	Position	Leader	Affiliation International
Miljøpartiet de Grønne Green Party	Green politics	Centre-left	Rasmus Hansson, Une Aina Bastholm	Global Greens

(b) **Output**: Predictions with Error Annotations and Reference

M1: BLEU prediction: The **Miljpartiet de Grnne**[NE] Green Party is a political party in Norway.

GM3: t5-base and GM4: t5-large predictions: The **Milj ?? partiet de Gr ?? nne**[NE] Green Party is a political party in Norway.

Reference: The Green Party **Miljøpartiet De Grønne** is a green political party in Norway.

Table 19: NON-ENGLISH error made by all models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: Liberal Democratic Party (Serbia)
Section Title: Parliamentary elections

Year	Popular vote	% of popular vote	# of seats	Seat change	Coalitions
2007	214,262	5.31%	6 / 250	6	With GSS-SDU-LSV-DHSS

(b) **Output**: Predictions with Error Annotations and Reference

M1: BLEU prediction: In the 2007 parliamentary election, the Liberal Democratic Party (Serbia) with GSSS-DULSVDHSS[O] won[W] 5.31% of the vote.

GM3: t5-base and GM4: t5-large predictions:
In the 2007 parliamentary elections, the Liberal Democratic Party won[W] 5.31% of the vote.

Reference: The LDP's first electoral performance was during the 2007 Serbian parliamentary election, where the **LDP ran in a coalition together with the Civic Alliance of Serbia, the Social Democratic Union and the League of Social Democrats of Vojvodina**[R] which **collectively received**[R] 5.31% of the popular vote.

Table 20: OTHER and WORD errors made by our model; WORD error made by benchmark models

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: List of Governors of Guam
Section Title: Elected governors (1971 - present)
Democratic (4) Republican (5)

Governor	Served	Party	Elected	Previous Office	Lieutenant Governor
Joseph F. Ada (1943–)	January 3, 1987 – January 2, 1995	Republican	1986 1990	Lieutenant Governor of Guam	Frank F. Blas

(b) **Output**: Predictions with Error Annotations and Reference

GM3: t5-base and GM4: t5-large predictions:
Joseph F. Ada served as Governor of Guam from 1987 to 1995 and **as Lieutenant Governor from 1986 to 1990 under Governor Frank F. Blas**[C].

Reference: **Blas served as the Lieutenant Governor of Guam**[R] from 1987 until 1995 for two consecutive terms **under Governor Joseph Franklin Ada**[R].

Table 21: CONTEXT error made by all models

Complex sample: Joseph F. Ada got elected for two consecutive terms, in 1986 and 1990. Frank F. Blas was Lieutenant Governor during the same period. All the models struggled to predict the factual output. As the Input Table for this complex sample has multiple names, multiple date-dimension fields and does not comply with the First Normal Form (1NF), the predicted outputs are misleading.

(a) **Input**: Table with Title, Highlighted cells and their headers

Page Title: 1998 United States Senate elections
Section Title: Elections leading to the next Congress
Democratic (4) Republican (5)

State	Incumbent			
	Senator	Party	Results	Candidates
Georgia	Paul Coverdel	Republican	Incumbent re-elected.	Paul Coverdell (Republican) 52.3% Michael Coles (Democratic) 45.3% Bertil Armin Loftman (Libertarian) 2.5%

(b) **Output**: Predictions with Error Annotations and Reference

M1: BLEU prediction: Incumbent Republican Paul Coverdell won re-election to a second term[X] in Georgia.

GM3: t5-base prediction:
Incumbent Republican Paul Coverdell won re-election to a second term[X] over Democrat Michael Coles.

Reference: In Georgia, Incumbent Republican Senator Paul Coverdell **defeated Michael Coles in the 1998 United States Senate elections**[R].

Table 22: NOT-CHECKABLE error made by our model and GM3: t5-base model