Diversity Enhanced Table-to-Text Generation via Logic-Type Control

Anonymous ACL submission

Abstract

Generating natural language statements to convey logical inferences from tabular data (i.e., Logical NLG) is a process with one input and a variety of valid outputs. This characteristic underscores the need for a method to produce a diverse set of valid outputs, presenting different perspectives of the input data. We propose a simple yet effective diversity-enhancing scheme that builds upon an inherent property of the statements, their logic-types, by using a type-controlled table-to-text generation model. We demonstrate, through extensive automatic and human evaluations over the two publicly available Logical NLG datasets, that our proposed method both facilitates the ability to effectively control the generated statement type, and produces results superior to the strongest baselines in terms of quality and factuality-diversity trade-off.

1 Introduction

Table-to-text (T2T) generation is the task of generating natural language statements to convey information appearing in tabular data. This task is relevant in real-world scenarios including generation of weather forecasts (Goldberg et al., 1994), sports results (Wiseman et al., 2017), and more.

A statement generated from tabular data can be inferred based on different levels of information. These range from a value of a specific cell to the result of logical or numerical operations across multiple cells, such as the average value of a column, or a comparison between rows.

In NLG in general, and in T2T generation in particular, a *diverse set* of generated outputs given a single input is favorable, as it offers different perspectives on the data, provides the user with multiple options to choose from, and facilitates further improvement of output quality via post-generation re-ranking algorithms (Gimpel et al., 2013).

In this work, we propose a method for enriching the control and diversity of generated T2T outputs.

Worldwide cheese market cap

Year	Market cap
2022	81.2
2021	76.1
2020	63.8
...	...
1961	12.1
1960	14.1

(a) Diversity Enhancement via Type Control
The cheese market cap has risen by 17.4B USD between 2022 and 2020
The cheese market cap had passed a value of 60B USD in only 3 years
The average cheese market cap between 1980 to 2000 was 51.3B USD

(b) Diversity Enhancement via Decoding Techniques
2022 is the year with the **highest** cheese market cap with 81.2B USD
2022 is the year with the **largest** cheese market cap at 81.2B USD
In 2022, the **largest** cheese market cap was 81.2B USD

Figure 1: T2T generation of 3-statement sets for the table on the left; (a) LT controlled: each statement delivers a unique piece of information, yielded by the control employed: compare, count, and aggregation; (b) decoding-based diversity: all are focused on one fact, hence demonstrating a weak diversity.

To this end, we leverage a common semantic partition of T2T statements into *numeric-logic types* (LTs) (Chen et al., 2020a) representing different perspectives of the data (see Figure 1(a)).

Namely, we utilize these LTs to realize a *controlled* generation model, that allows guiding generated statements to a specific LT, out of the many different valid LTs corresponding to the input table. This controlled generation model enables our method, *Diversity enhancement via LT Control* (DEVTC) to produce a diverse set of statements representing multiple perspectives of the data, by conditioning upon several different LTs.

As previous T2T methods intrinsically can only produce a single output per input, they obtain output diversity through common decoding techniques that have been shown to suffer from a trade-off between diversity and quality measures such as fluency and adequacy (Ippolito et al., 2019). By this trade-off, high quality hinders diversity, as exemplified in Figure 1(b). In contrast, we show that DEVTC readily generates a diverse set of high quality statements, allows LT-control, all while surpassing the baseline models in terms of generation quality.

Through extensive experimentation, we show that by employing this simple LT-control scheme, DEVTC surpasses SOTA methods in the trade-off between diversity and quality, measured here in *fac-*

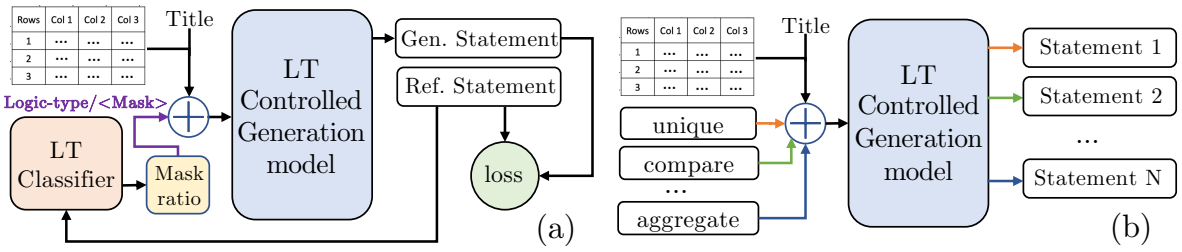

Figure 2: Framework; (a) **Train**: the LT-conditional model is trained to generate a reference statement given the statement LT as it is predicted by our LT classifier; (b) **Inference**: DEVTC is realized by inputting several different LTs along with a single table resulting in a diverse set of statements.

tuality which is a paramount concern in T2T. We also show that DEVTC generates statements adhering to the LT required by the user, and moreover, even in the absence of an input LT, outperforms the baselines on the two common benchmarks[1].

2 Related Work

The task of LOGICAL NLG, introduced by Chen et al. (2020a), involves the automatic application of complex numeric-logic operations on data tables along with the natural language expressions of them as statements. The task was accompanied by a dataset, LOGICNLG, that contains a set of (table, statement) pairs. In addition to the LOGICNLG dataset, Chen et al. (2020a) presented two methods based on GPT2 (Radford et al., 2019). Both methods receive the same input $\mathbf{T}$: a table in conjunction with a title, denoted as a natural language sequence, but differ in their generation scheme. GPT-TABGEN learns to generate a statement Y directly: $p_\theta(Y|\mathbf{T})$; while GPT-C2F generates a statement-template, $\tilde{Y}$, and conditions on it to create the final statement, effectively learning $p_\theta([\tilde{Y}; Y]|\mathbf{T})$. In a subsequent work, Chen et al. (2021) proposed DCVED, a scheme based on a conditional variational auto-encoder architecture. Their scheme can generate multiple statements for a single input, but these only undergo a re-ranking, and their diversity or quality aspects are not discussed. LOGIC2TEXT (Chen et al., 2020b) is a small dataset similar to LOGICNLG. In its associated task, a model receives an additional logical-form input, specifying its full logical description. Liu et al. (2021) aims to circumvent the problem of data scarcity of LOGIC2TEXT with an approach combining data-augmentation, data-weighting and semi-supervised learning using LT-controlled generation module. In contrast to their work, our trained model is robust to missing LTs, and, paired with a diversity enhancing scheme, is shown to improve both generation diversity and factuality. Recently, Zhao et al. (2023) successfully applied the proposed method to different tasks and models, extending it with additional LTs and a post-filtering module.

[1]Models and code will be made public upon acceptance.

3 Method

3.1 Statement-LT Classifier

To enable controlled generation learning, we augmented our training datasets with LT-control annotations. Specifically, we automatically annotated our training datasets with 7 LTs, namely, c = {count, comparative, superlative, unique, ordinal, aggregation, majority} by employing a BERT (Devlin et al., 2018) based classifier $p_\phi(c|Y)$ that was fine-tuned on 8.5K (statement, LT) pairs from the LOGIC2TEXT train set.

This classifier achieved 97% macro F1 on the corresponding test set. To measure the classifier's ability to transfer, we ran it on 200 randomly sampled statements from LOGICNLG annotated by experts achieving 90% macro F1.

3.2 LT-controlled T2T Generation Model

As depicted in Figure 2(a), we train an LT-controlled generation model, learning $p_\theta(Y|\mathbf{T}, c)$, obtaining LT annotations from the reference statement using the statement-LT classifier. The LTs are concatenated to the table and title to produce the input. The model is then trained to minimize the autoregressive cross-entropy loss between the generated and reference tokens.

During training, we apply a mask over the LT with probability $p_{mask} = 0.5$. Inducing an equal chance to receive a masked token as the LT in train-

ing, this ratio allows the model to learn how to condition on the LT, while also enabling robustness for scenarios where LT is unavailable for the model to condition on. The effects of other p_{mask} choices are discussed in Appendix A.3.

3.3 Diversity Enhancement via LT Control

Figure 2(b) presents our DEVTC inference-time flow. As the figure depicts, we utilize the above $p_\theta(Y|\mathbf{T}, c)$ model to generate multiple statements, each conditioned on a different LT sampled from a uniform LT distribution. By this process, DEVTC is able to produce statements with various types providing different perspectives on the data.

4 Experiments

4.1 Datasets

In our experiments, we use LOGICNLG (Chen et al., 2020a) and LOGIC2TEXT (Chen et al., 2020b). Each data-point in LOGICNLG consists of a parent-table crawled from Wikipedia from which 5 tables are derived, each containing a subset of the parent-table columns and an associated statement generated by crowd-workers. LOGIC2TEXT is similar but further provides statement logical-form (its full logical description) from which we extract the LT. In our experiments, we will use these LTs to train a statement-LT classifier (cf. Section 3.1) but will **not** use these extra annotations in training or evaluating the generation model. To the best of our knowledge, these two datasets are the only publicly available table-to-text datasets that include statement generation capturing complex logical and numerical operations from tables, making them the only datasets relevant for our scenario.

4.2 Metrics

Following previously proposed evaluation practices laid out by Chen et al. (2020a), we evaluate the quality of a generated text, with BLEU to measure consistency with the reference text; and the SP-ACC (SP-A) and NLI-ACC (NLI-A) metrics to estimate its factuality, using semantic parsing and a pretrained NLI model, respectively. Specifically, we focus on NLI-A that was found to better agree with human preference for factuality evaluation (Honovich et al., 2022). For measuring the diversity of the generated statements we use the three common n-gram based metrics Self-BLEUn (Zhu et al., 2018), Ent-n (Zhang et al., 2018) and Dist-n (Li et al., 2016).

LOGICNLG

Model	Size	BLEU 1/2/3 (↑)	SP-A (↑)	NLI-A (↑)
GPT-C2F	sm	46.6 / 26.8 / 13.3	42.7	72.2
GPT-TABGEN	sm	48.8 / 27.1 / 12.6	42.1	68.7
DEVTC$_{mask}$	sm	**50.0 / 28.6 / 14.4**	**43.0**	**73.4**
DCVED	med	49.3 / 28.3 / 14.2	44.3	73.9
GPT-C2F	med	49.0 / 28.3 / 14.6	45.3	76.4
GPT-TABGEN	med	49.6 / 28.2 / 14.2	44.7	74.6
DEVTC$_{mask}$	med	**50.8 / 29.2 / 15.2**	**45.6**	**77.0**

LOGIC2TEXT

Model	Size	BLEU 1/2/3 (↑)	SP-A (↑)	NLI-A (↑)
DCVED	med	46.4 / 31.2 / 20.1	**43.7**	71.9
GPT-C2F*	med	46.6 / 31.1 / 20.5	40.8	73.4
GPT-TABGEN*	med	46.1 / 32.4 / 21.0	41.0	70.3
DEVTC$_{mask}$	med	**47.8 / 32.6 / 22.2**	41.9	**74.4**

Table 1: Quality results on the test split of LOGICNLG and LOGIC2TEXT. Baseline models trained by us are marked with a *, all DEVTC and starred results are the average over 5 different seeds, **bold** marks statistically significant advantage. DEVTC is marked with *mask* to indicate the use a mask token as the type.

4.3 Hyper-parameters & Baseline Methods

We use the same hyper-parameters as in Chen et al. (2020a), apart from the learning rate (LR) for which we tried 6 values between $1e$-6 to $5e$-5 and chose the best LR per according to our model selection scheme, that uses the dev set BLEU3 score. As for models, we compare **DEVTC** based on GPT2-small/medium with **GPT-C2F** and **GPT-TABGEN**, and **DCVED**. DCVED is compared against the medium models only since it uses two GPT2-small and two fully-connected networks, adding up to a larger parameter count than GPT2-medium. Further details can be found in App. A.2.

5 Results

5.1 Quality Performance

Conventional T2T setup evaluation was done on the LOGICNLG and LOGIC2TEXT test-sets. As in (Chen et al., 2021), when evaluating on LOGIC2TEXT we follow the Logical NLG task formulation and do not use the logical-form annotations. Since LT annotations are unavailable in this scenario, our type-controlled models are conditioned on a mask token as control. As shown in Table 1, for both datasets, across metrics and model sizes DEVTC outperforms all baseline methods.

5.2 Factuality-Diversity Trade-off

To compare DEVTC and GPT-TABGEN (the strongest baseline) across the Factuality-Diversity plane, we generated a set of 5 statements per table for each method. Since, as opposed to DE-

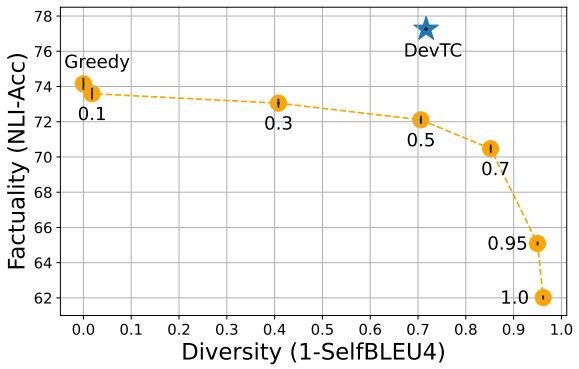

Figure 3: Factuality-diversity trade-off for LogicNLG: each dot in the orange line represents an average over 5 seeds (error bars are SEMs) of the baseline model (GPT-TabGen) with nucleus sampling parameters varied between 0 (greedy decoding) and 1. The blue star is our method (using greedy decoding) that surpasses the the baselines pareto frontier.

vTC, which natively enables the production of a diverse set of statements via LT-control, the baseline cannot produce a diverse set with greedy decoding, we utilized stochastic decoding, the most common practice to obtain a set of different outputs from a single model. Following Ippolito et al. (2019) we varied the top_p decoding parameter of the baseline to explore the factuality-diversity trade-off for the baseline. In contrast, DevTC allows us to achieve diversity without using stochastic decoding (which is known to reduce quality), by conditioning on different LTs. Fig. 3 shows results of DevTC obtained by conditioning on 5 LTs sampled uniformly from the 7 LTs, compared to the baseline paired with stochastic decoding as described above. To evaluate, we measured the diversity within each set, along with the average NLI-A. Figure 3 shows that DevTC is better positioned on the factuality-diversity plane, surpassing the baselines Pareto front. We attribute this gain in generation factuality to the use of more accurate supervision through the LTs, offloading the task of LT prediction from the model, and bypassing the quality degradation incurred by stochastic decoding. We found these results to be consistent across other diversity measures such as Ent-2/4 and Dist-2/4, decoding methods, and datasets (see Appendix A.4 for more results).

5.3 Human Evaluation

5.3.1 Factuality and Diversity

We complement the automatic evaluation results with human evaluation, to verify the success of our approach in gaining a given diversity with higher factually. We therefore choose the top_p decoding parameter of GPT-TabGen to be the one that produces the most similar output to DevTC in terms of diversity (i.e. 0.5). We sampled 100 tables from the set used in Section 5.2 and distribute them independently to 3 human experts. Each table was presented along with two 5-statement sets – one generated by DevTC, and the other by GPT-TabGen. The experts were asked which of the two sets is more factual, i.e., properly describes the data in the table (ties are also allowed), and which is more diverse – on Likert scale, from -2 (set-1 is much better) to $+2$ (set-2 is much better). Overall, the human evaluation findings are inline with the results appearing in Figure 3. In 55% of the samples presented to the annotators, DevTC was reported to be more factual compared to 21% for GPT-TabGen. The rest 24% were reported as a tie. DevTC advantage is statistically significant (P_{value}<0.05) using two-sided t-test. For diversity, there was a slight advantage to DevTC implying no significant difference inline with Fig 3. The annotators inter rater Cohen's Kappa is 0.553 indicating moderate agreement.

5.3.2 LT-Control

To verify our models proficiency in LT-control we asked the experts to classify the LTs of the above generated statements. The LT-consistency (i.e., the ratio of examples where control LT resulted in a generated statement classified to the same LT) on average over the 7 types is 79.8% (for comparison, a baseline model produced a ratio of 17%). The lowest consistency is for *ordinal*, which is characterized with relatively high lexical variance, and for which we had relatively scarce training data.

6 Conclusions

We presented DevTC, an innovative model for T2T generation that addresses and implements two prominent features of that task, overlooked by existing models: diversity and control. DevTC facilitates the generation of a statement of a desired LT, and the option to generate a diverse set of high quality statements, features that are unlocked by adding statement LT-control to the input. Results show the merit of our approach compared to existing baselines in generation quality as measured by common benchmarks, diversity-factuality trade-off in automatic and human evaluations.

7 Limitations

The main limitations of our work are automatic factuality evaluation and factual generation. In terms of automatic factuality evaluation, current SOTA table fact-checking metrics such as NLI-Acc and SP-Acc still present medium human agreement (See Figure 7). In terms of factual generation as determined by human evaluation, as all End-to-end T2T methods, GPT-TabGen, the method we use to show the ability of DevTC to improve diversity without sacrificing accuracy, suffers from weak human approval in terms of factuality. As we show in the main text, DevTC is able to improve the factuality over the baseline but still presents human approval factuality that is too low for business applications.

References

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

Wenqing Chen, Jidong Tian, Yitian Li, Hao He, and Yaohui Jin. 2021. De-confounded variational encoder-decoder for logical table-to-text generation. 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 5532–5542.

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020b. Logic2text: High-fidelity natural language generation from logical forms. arXiv preprint arXiv:2004.14579.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A systematic exploration of diversity in machine translation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1100–1111.

Eli Goldberg, Norbert Driedger, and Richard I Kittredge. 1994. Using natural-language processing to produce weather forecasts. IEEE Expert, 9(2):45–53.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. arXiv preprint arXiv:2204.04991.

Daphne Ippolito, Reno Kriz, Maria Kustikova, João Sedoc, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. arXiv preprint arXiv:1906.06362.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

Ao Liu, Congjian Luo, and Naoaki Okazaki. 2021. Improving logical-level natural language generation with topic-conditioned data augmentation and logical form generation. arXiv preprint arXiv:2112.06240.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. arXiv preprint arXiv:1707.08052.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. Advances in Neural Information Processing Systems, 31.

Yilun Zhao, Zhenting Qi, Linyong Nan, Lorenzo Jaime Yu Flores, and Dragomir Radev. 2023. Loft: Enhancing faithfulness and diversity for table-to-text generation via logic form control. arXiv preprint arXiv:2302.02962.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pages 1097–1100.

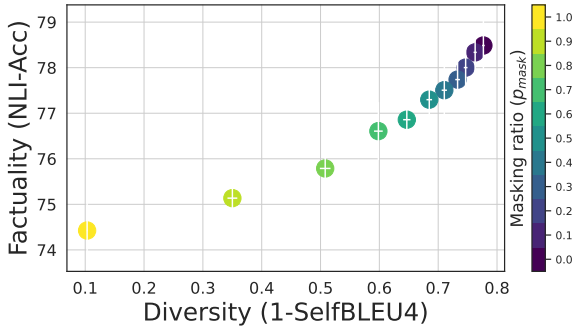

Figure 4: Factuality-diversity trade-off over the LogicNLG dataset for different p_{mask}, averaged over 5 seed (error-bars are SEMs).

A Appendix

A.1 Dataset Statistics

See Table 2.

Dataset	Parent tables	Statements	Train / Dev / Test
LOGICNLG	7,392	37,015	28,450 / 4,260 / 4,305
LOGIC2TEXT	5,554	10,753	8,566 / 1,095 / 1,092

Table 2: Datasets statistics.

A.2 Implementation Details

All models are trained with batch size of 32 on 1 NVIDIA A100 GPUs for 12 epochs. We use Adam optimizer (Kingma and Ba, 2014) and an autoregressive cross entropy loss to optimize the models. During test time, we use a greedy search to generate text and calculate the BLEU-1,2,3 scores with the 5 references from all 5 sub-tables as suggested by (Chen et al., 2020a). We base our implementation on Huggingface's Transformers (Wolf et al., 2019) version 4.16.2 in the (Paszke et al., 2019) flavour and use the pre-trained version of GPT-2 (Radford et al., 2019) small/medium with subword unit vocabulary of 30K. All models selection is based on the BLEU-3 score on dev set. All our models and models marked with a * were found to have the best performance with learning rate set to $1e$-5. Regarding DCVED, we note that, we report the original variant of DCVED without an additional generate-and-select scheme, since multiple generation and re-ranking is complementary and could potentially be applied to all compared methods.

A.3 Masking Ratio Effect

To analyze how the different LT masking ratios used in training impact model performance, we trained 11 LT-controlled models with p_{mask} varying from 0.0 (no masking) to 1.0 (always masked). In Figure 4 we compare these models using the same evaluation protocol as in Section 5.2. As expected, both factuality and diversity obtained by DEVTC gain significantly from strengthening the control. That is, as expected, a lower masking ratio means a more stable training process with better LT correspondence, which in turn results in higher diversity and better factuality on the test set.

Table 3 is complementary to the automatic evaluation and includes the standard error of the mean for our models.

LogicNLG

Model	Size	BLEU 1/2/3 (↑)	SP (↑)	NLI (↑)
DEVTC	sm	50.0±0.2 / 28.6±0.2 / 14.4±0.2	43.0±0.3	73.4±0.5
DEVTC (oracle)	sm	51.3±0.1 / 30.3±0.1 / 15.6±0.1	40.5±0.5	75.4±0.2
DEVTC	med	50.8±0.2 / 29.2±0.2 / 15.2±0.2	45.6±0.5	77.0±0.6
DEVTC (oracle)	med	52.3±0.2 / 31.1±0.2 / 16.7±0.2	42.7±0.5	78.2±0.2

LOGIC2TEXT

Model	Size	BLEU 1/2/3 (↑)	SP (↑)	NLI (↑)
DEVTC	med	47.8±0.2 / 32.6±0.1 /22.2±0.1	41.9±0.2	74.4±0.7
DEVTC (oracle)	med	48.4±0.2 / 33.6±0.2 / 23.2±0.1	42.6±0.7	76.1±0.5

Table 3: Quality results on the test split of LOGICNLG and LOGIC2TEXT, all DEVTC results are the average over 5 different seeds, the ±s represents the standard error of the mean.

A.4 Factuality-Diversity Trade-off: Other diversity measures

Figure 6 displays the factuality-diversity trade-off discussed in Section 5.2 for the other two diversity metrics, SelfBLEU4 and Dist2.

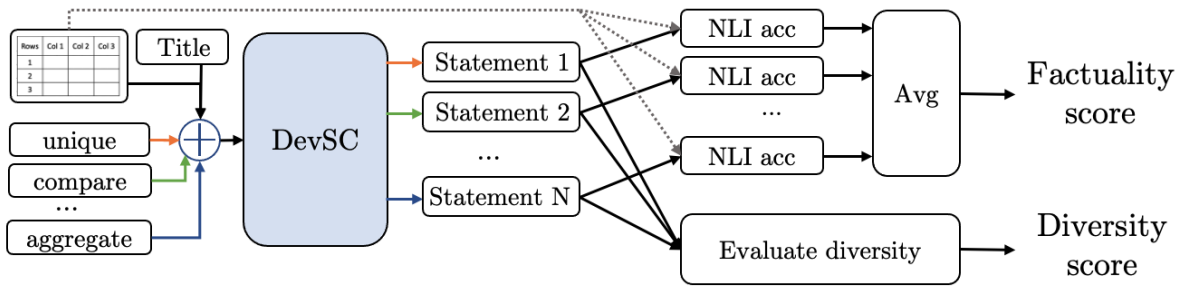

Figure 5: An illustration of the quality-diversity trade-off evaluation. NLI-Acc is a fact checking model proposed by Chen et al. (2020a) that labels the statement as true or false given the table.

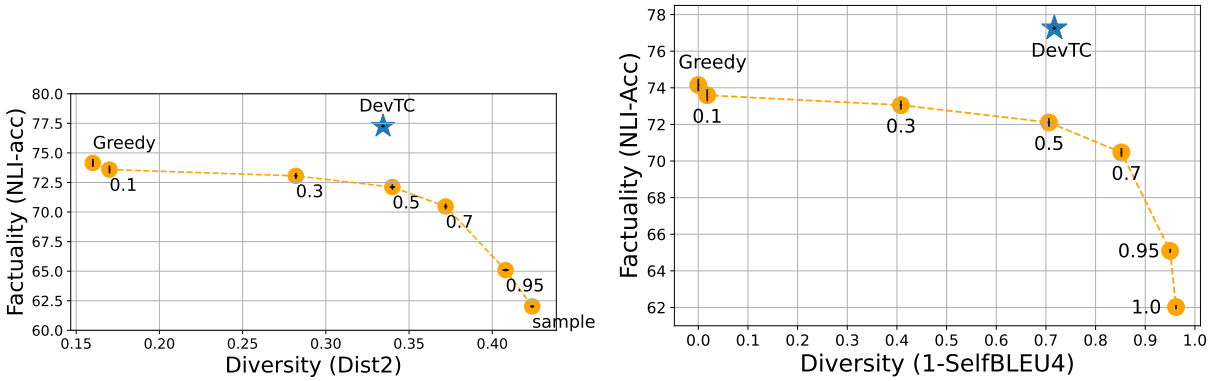

Figure 6: Factuality-Diversity trade-off for Dist-2 and Self-BLEU4: each dot in the orange line represents an average over 5 seeds (error bars are SEMs) of the baseline model (GPT-TABGEN*) with a different nucleus sampling decoding parameters (shown in the figure). The blue star is our method that surpasses the trade-off line created by the baseline and the decoding strategy.

(a)

Type	Generated statement
Superlative	The United State had the most Gold medal, with 4
Ordinal	The United State and Canada each received 4 medal
Comparative	The United State had 4 more medal than Latvia
Comperative	The United State had a higher Total than Latvia
Count	The United State and Canada had the same Total medal

(B)

Nation	Gold	Total
united states	4	5
Canada	1	4
latvia	1	1
germany	0	6
new zealand	0	1
united kingdom	0	1

Figure 7: 5 statements generated using DEVTC along with the table that was used for their generation, sentences marked in red display false type correspondence.