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

Anonymous ACL submission

Abstract

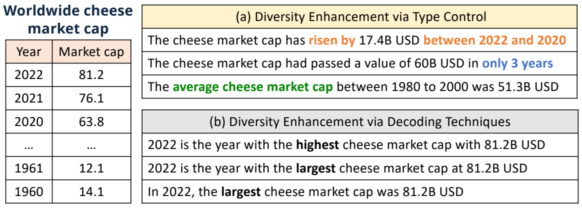
Generating natural language statements to convey logical inferences from tabular data (i.e., Logical NLG) is a process with one input and a variety of valid outputs. This characteristic underscores the ability for a method to produce a diverse set of valid outputs. We propose a simple yet effective diversity 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 method is able to surpass the strongest baselines along the quality-diversity plane, all while allowing users to effectively control the type of the generated statement.

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), sport 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. LogicNLG, introduced by Chen et al. (2020a), involves the automatic application of complex numeric-logic operations on the data and the natural language expressing of them and their results. The task was accompanied by a dataset, LogicNLG, that contains a set of (table, statement) pairs, and several baselines for statement generation.

Generally in NLG, a *diverse set* of generated hypotheses 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 all sorts of post-generation re-ranking algorithms (Gimpel et al., 2013).



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.

In T2T generation, and specifically in Logical NLG, diversity naturally emerges from the different *numeric-logic types* (LTs) used to infer the statements from the table (see Figure 1(a)). Here, we propose to use these LTs and realize a *controlled* generation model that enables our method, *Diversity enhancement via LT Control* (DevTC) to generate a diverse set of statements by conditioning each generation on a different type. In addition, a conditional generation model allows users to further guide generated statements to a specific LT, out of the many different valid statements corresponding to the input table. Existing T2T methods intrinsically can only produce a single output per input and 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, DevTC readily generates a diverse set of high quality statements, with a variety of LTs, without suffering any degradation in quality.

Through extensive experimentation, we show

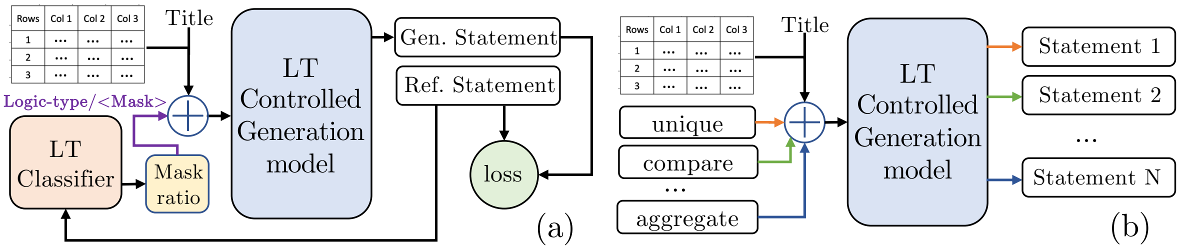


Figure 2: Framework; (a) **Train**: the LT-conditional model is trained to generate a reference statement given the statement LT as it is predicted by our LT classifier; (b) **Inference**: DEVTC is realized by inputting several different LTs along with a single table resulting in a diverse set of statements.

that by employing this simple LT-control scheme, DEVTC surpasses SOTA methods on the trade-off between diversity and quality, measured here in *factuality* which is a paramount quantity in T2T. We also show that DEVTC generates statements adhering to the LT required by the user and performs on par with current SOTA on the common benchmarks even in the absence of input LT on the two relevant datasets[1].

2 Related Work

Along with 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: $p_\theta(Y|\mathbf{T})$; whereas 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.

3 Method

3.1 Statement-LT Classifier

To enable controlled generation learning, we had to augment our training datasets with LT-control annotations. Specifically, we automatically annotated our training datasets with 7 LTs as proposed by Chen et al. (2020b), 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. The classifier further achieved 90% macro F1 on 200 randomly sampled statements from LOGICNLG annotated by experts.

3.2 LT-controlled T2T Generation Model

We propose to re-purpose GPT-TABGEN as a LT-controlled generation model, learning $p_\theta(Y|\mathbf{T}, c)$. At training, we predict the LT from the gold statement using the dedicated classifier and concatenate it to the table and title (see Figure 2(a)). 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 $p_{mask} = 0.5$, this ratio adds robustness for scenarios where LT is unavailable for the model to condition on, on the one hand and allows the model to learn how to condition on LT on the other. Effects of p_{mask} choices are discussed in Section 5.2.

3.3 Diversity Enhancement via LT Control

Figure 2(b) presents our inference-time flow in which we utilize the above $p_\theta(Y|\mathbf{T}, c)$ model

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

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 1: Datasets statistics.

for our Diversity Enhancement via LT Control (DevTC) scheme. Specifically, at inference time, given a table, we generate multiple statements, each conditioned on a different LT sampled from a uniform LT distribution, ending up with a diverse set of statements.

4 Experiments

4.1 Datasets

In our experiments, we use LogicNLG (Chen et al., 2020a) and Logic2Text (Chen et al., 2020b) (Table 1). 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 and NLI-Acc metrics to estimate its factuality, using semantic parsing and a pretrained NLI model, respectively. Specifically, we focus on NLI-Acc 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).

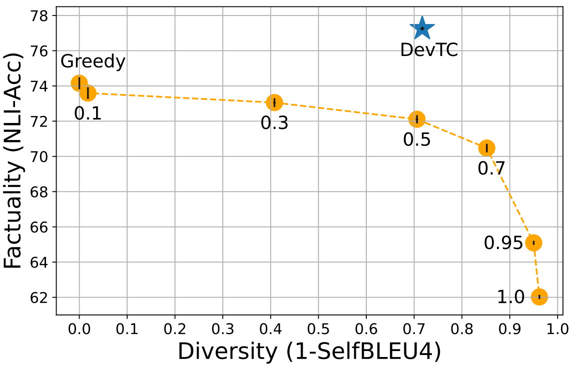


Figure 3: Factuality-diversity trade-off for LogicNLG: each dot in the orange line represents an average over 5 seeds (error bars are SEMs) of the baseline model (GPT-TabGen*) with nucleus sampling parameters varied between 0 (greedy decoding) and 1. The blue star is our method (using greedy decoding)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

4.3 Hyper-parameters & Compared Models

We use the same configurations and settings 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 method according to our model selection scheme, that uses the dev. set BLEU3 score. Of the baselines, only GPT-TabGen benefited from the sweep, and we marked the improved version as GPT-TabGen*. Further details can be found in Appendix A.1. As for models, we compare DevTC with the GPT-C2F and GPT-TabGen* across both small / medium GPT2 model versions. DCVED is considered medium since it uses two GPT2-small and two fully-connected networks, adding up to a larger parameter count than GPT2-medium.

5 Results

5.1 Factuality-Diversity Trade-off

To compare DevTC and the strongest baseline across the Factuality-Diversity plane, for each method we generated a set of 5 statements per table in the test-set. Since, as opposed to DevTC, 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. For DevTC, since we do not require a

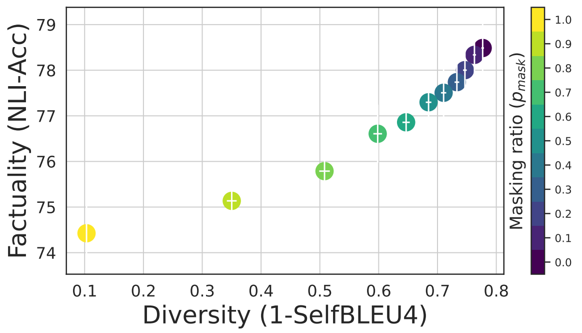


Figure 4: Factuality-diversity trade-off over the LogicNLG dataset for different p_{mask}, averaged over 5 seed (error-bars are SEMs).

stochastic decoding to obtain a diverse set of statements, we used greedy decoding and generated the set by conditioning on 5 LTs sampled uniformly from the 7 LTs. To evaluate, we measured the diversity within each set, along with the average NLI-Acc. In Figure 3 we show 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.3 for more results).

5.2 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 5.1. 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.

5.3 Robustness for Missing LTs

To demonstrate DevTC performance in a conventional setup when LT is unavailable as input at test time, we compared our LT-controlled model with a masked token as control, with SOTA methods, on both the LogicNLG and Logic2Text test-sets, using the standard evaluation protocol. For both datasets, across all metrics and model sizes, DevTC is leading the benchmark along with GPT-TabGen*. For detailed results, see Table 3 in Appendix A.2.

5.4 Human Evaluation

We complement the automatic evaluation results with human evaluation. We sampled 100 tables from the set used in Section 5.1 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 top_p decoding parameter set to 0.5. 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). Our results find that on 50% of the samples, DevTC reported to be more factual vs. 31% for GPT-TabGen*. 19% of the samples were reported as a tie. DevTC advantage is statistically significant (P_{value}<0.05) using two-sided t-test. For diversity, the average score was 0.14, implying no significant difference, in line with Figure 3. To verify our models proficiency in LT-control we additionally asked the experts to classify the LTs of the above generated statements. Table 2 shows the ratio of examples where control LT resulted in a generated statement classified to the same LT. It shows high LT-consistency for all LTs but *ordinal*, which is characterized with relatively high lexical variance, and for which we had relatively scarce training data.

Agg.	Comp.	Count	Maj.	Ord.	Super.	Unique
0.87	0.92	0.90	0.88	0.46	0.96	0.60

Table 2: Human evaluation of controlled-generation LT consistency.

6 Conclusions and Future Work

DevTC facilitates the generation of a statement of a desired LT, and the option to generate a diverse set of high quality statements. Both features are unlocked by adding statement LT control to the input. Results show the merit of our approach compared to existing baselines. In future work we plan to study how to further improve factuality, i.e., the faithfulness of the statements generated by our approach, to bring it to practical use.

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 NLI-SP 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.

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.

A Appendix

A.1 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.

A.2 Automatic Evaluations

Table 3 demonstrates DEVTCs performance in the conventional setup. In the table, our type-controlled models (marked as DEVTC) are using a mask token as control, the oracle version (marked as DEVTC (oracle)) are receiving the type as classified by the logic-type generator while the baselines are not receiving types. 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. We further note that, we report the original variant of DCVED without an additional generate-and-select scheme also reported by them, since multiple generation and re-ranking is complementary and could potentially be applied to all compared methods.

We see that for both datasets, across all metrics and model sizes, our model with p_{mask}=0.5 is leading the benchmark along with GPT-TABGEN*. Also, we note that the oracle methods enjoys the types perform the best by a great margin in four of the five metrics. We attribute the decline in SP-ACC to the different type distribution the model generates when in oracle mode that impacts the SP-ACC since different types are more likely to be labeled as accurate by SP-ACC.

Table 4 is complementary to the automatic evaluation and includes the standard error of the mean for our models.

A.3 Factuality-Diversity Trade-off: Other diversity measures

Figure 6 displays the factuality-diversity trade-off discussed in Section 5.1 for the other two diversity metrics, SelfBLEU4 and Dist2.

LOGICNLG

Model	Size	BLEU 1/2/3 (↑)	SP (↑)	NLI (↑)
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
GPT-TABGEN*	sm	49.6 / 28.5 / 14.2	44.8	73.2
DEVTC	sm	50.0 / 28.6 / 14.4	43.0	73.4
DEVTC (oracle)	sm	51.3 / 30.1 / 15.6	40.5	75.5
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
GPT-TABGEN*	med	50.8 / 29.4 / 15.2	46.1	76.1
DEVTC	med	50.8 / 29.2 / 15.2	45.6	77.0
DEVTC (oracle)	med	52.3 / 31.1 / 16.7	42.7	78.2

Logic2Text

Model	Size	BLEU 1/2/3 (↑)	SP (↑)	NLI (↑)
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	48.1 / 32.4 / 22.0	41.0	70.3
DEVTC	med	47.8 / 32.6 / 22.2	41.9	74.4
DEVTC (oracle)	med	48.4 / 33.6 / 23.2	42.6	76.1

Table 3: Quality results on the test split of LOGICNLG and Logic2Text. Baseline models trained with our learning rate are marked with a * , all DEVTC and starred results are the average over 5 different seeds, SEMs of our models are in Table 4. SP and NLI stands for SP-Acc and NLI-Acc from (Chen et al., 2020a)

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 4: 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.

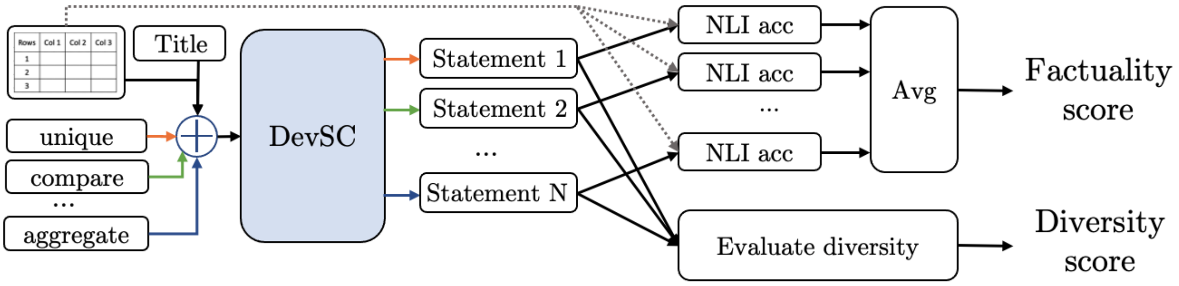


Figure 5: An illustration of the quality-diversity trade-off evaluation. NLI-Acc is a fact checking model proposed by Chen et al. (2020a) that labels the statement as true or false given the table.

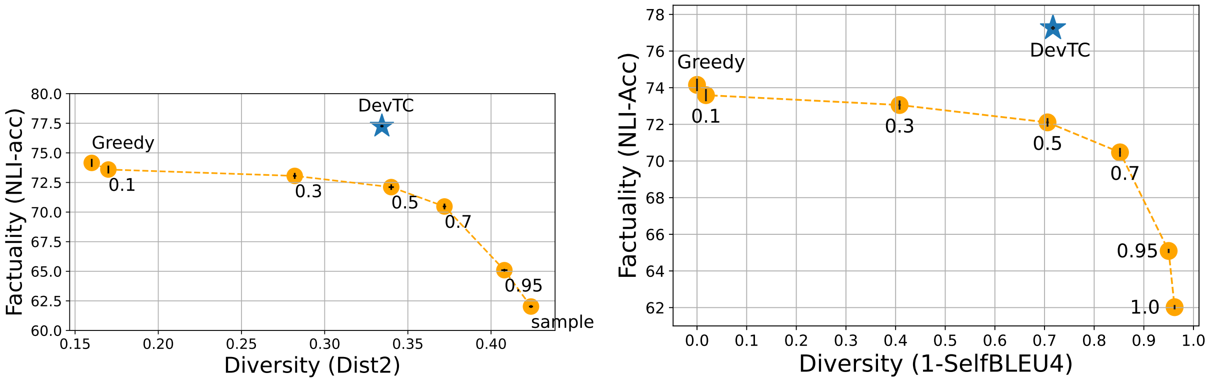
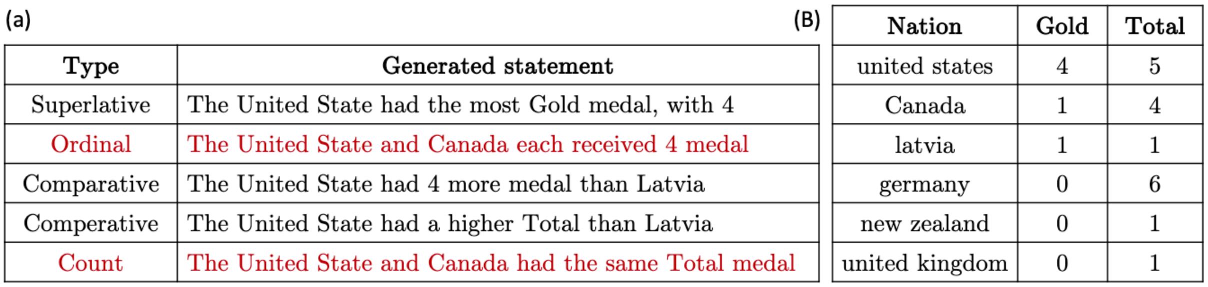


Figure 6: Factuality-Diversity trade-off for Dist-2 and Self-BLEU4: each dot in the orange line represents an average over 5 seeds (error bars are SEMs) of the baseline model (GPT-TabGen*) with a different nucleus sampling decoding parameters (shown in the figure). The blue star is our method that surpasses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