NADIA: Diverse Paraphrase Generation with Syntax Planning and Diversity-driven Sequence Calibration

Anonymous ACL submission

Abstract

Paraphrase generation is a fundamental task in natural language processing. In this work, we study diverse paraphrase generation, and propose a novel method to increase surface-form diversity while maintaining semantic similarity for the generated paraphrase. Our method disentangles the generation into syntax structure planning and semantic realization, which first produces a syntax tree as high-level guidance and then generates surface form of paraphrase conditioned on the syntax tree. We further introduce a diversity-driven calibration loss to rank the probability of model generated sequences and enhance the output diversity. We evaluate our method on both ParaNMT dataset and a newly proposed DiverseQuora dataset, and our model outperforms strong baselines with better quality and diversity on both datasets.

1 Introduction

Paraphrase generation is an important task in natural language processing, with the goal to transformer the source sentence into a different surface form while keeping the semantic meaning unchanged (Madnani and Dorr, 2010; Dou et al., 2022; Zhou and Bhat, 2021). It has various downstream applications such as question answering (Liu et al., 2020a), machine translation (Mallinson et al., 2017), and sentence simplification (Martin et al., 2022; Maddela et al., 2023).

While most studies in this domain focus on generating paraphrases with high semantic similarity, how to paraphrase with enhanced surface diversity is much less studied. Here, we define "*enhanced surface diversity*" as to generate sentences with largely different surface form compared with the original source input but still keep the semantic meaning unchanged. Surface-form diversity is an important feature for paraphrase generation because it helps to accommodate various audiences, contexts, and applications by generating multiple

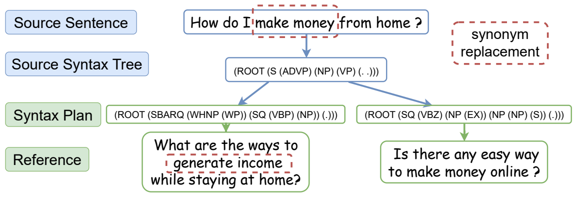

Figure 1: An example sentence and its paraphrases with different diversity. The syntax tree represents the surface-from organization of the target paraphrase.

ways to express the same idea. Diverse paraphrasing also ensures more robust and adaptable models, capable of understanding and producing a wider range of linguistic expressions. This can further benefit downstream applications by allowing more nuanced and varied outputs.

However, there remains challenges to generate diverse paraphrases with current token-level autoregressive language models. Achieving surface-form diversity while ensuring semantic fidelity to the input sentence is essential for effective paraphrase generation. Yet, the current training objective with maximum likelihood estimation (MLE) over each token in the target sequence cannot explicitly learn such disentanglement, and thus makes it hard to fulfill the aforementioned two objectives.

To overcome this, we propose a Syntax Driven Diverse Paraphrase framework (**NADIA**) to stably generate sentences with high diversity. First, syntactic structure is useful to represent the surface organization of a sentence, as shown in Figure 1. Thus, we explicitly incorporate syntactic structure as high-level guidance to control the surface-form generation and improve output diversity. Specifically, our model first produces a syntax tree as plan, and then conducts surface generation to produce paraphrases with synonym replacement conditioned on the syntax tree. By doing so, our model can effectively learn to disentangle paraphrase generation into syntax planning and semantic realization, thus generating more diverse outputs. Furthermore, to mitigate the issue of MLE training that lacks sequence-level objective, we introduce a diversity-driven calibration loss, which ranks model outputs and aligns sequence-level likelihood to both surface diversity and semantic similarity in the latent space. Therefore, our model learns to produce outputs with better diversity and quality.

To evaluate our model performance, we build up a new dataset, DiverseQuora, with more diverse targets compared to the existing paraphrase generation benchmarks. [1] Experiments on both ParaNMT and our newly proposed DiverseQuora dataset prove that NADIA with syntax planning and diversity-driven sequence calibration outperforms strong baselines with better quality and diversity.

2 Related Work

Paraphrase generation has received significant research attention. Li et al., 2016 studied using mutual information to generate more diverse responses. Prakash et al., 2016 first used deep neural networks to generate paraphrases. Wieting and Gimpel, 2018 (ParaNMT) and Kumar et al., 2020 (QQPos) built up widely used datasets for paraphrase. Compared to these datasets, DiverseQuora is more diverse and has better quality distilled from the Large Language Model. Most prior works in controllable paraphrase rely on reinforcement learning(Gupta et al., 2017, Li et al., 2018,Liu et al., 2020b), which is difficult to train and control the diversity level. Xu et al., 2018 studied using conditional embedding to control diverse generation, and Cao and Wan, 2020 studied extra loss in GAN to improve diversity. Similar to REAP(Goyal and Durrett, 2020) and BRIO (Liu et al., 2022), we use ordering to improve specific metrics of quality. Different from their work, we incorporate ordering into planning-based model to improve both surface diversity and semantic fidelity. AESOP(Sun et al., 2021) and GCPG(Yang et al., 2022) also use syntax information to control generation, but they rely on human-labeled exemplars. SGCP(Kumar et al., 2020) uses a syntax tree but within a fixed human labeled set. Both methods use contrastive loss to improve quality towards specific aspects.

3 Method

The overview of **NADIA** is shown in Figure 2. We first describe our planning based model architecture (§ 3.1), and then introduce the diversity-driven calibration loss (§ 3.1).

[1]Code and dataset will be released upon publication.

3.1 NADIA with Syntax-planning

Paraphrase generation is typically modeled as a sequence-to-sequence (Seq2seq) task with the conditional probability $P(y|x)$, where x denotes the input and y is the target. In this work, we explicitly incorporate target syntax feature z as high-level guidance into the generation process: Instead of directly generating surface target y, our model first computes $p(z|x)$ to generate a syntactic plan that represents the surface organization of the target, and then produces the final target conditioned on both input and plan with $p(y|z, x)$. In this way, our planning-based modeling disentangles the syntactic and semantic features and further improves diversity with the guidance of the syntax.

Concretely, as shown in Figure 2, the encoder takes the concatenation of the source sentence and source syntactic parse as input. For the decoder, instead of directly generating the target sentence, it first predicts the target syntactic parse, and then produce the target paraphrase according to it. As the generation of target depends on its syntactic plan, we can manipulate the target by sampling plans with desired attributes during inference, thereby enabling the model to enhance output diversity.

3.2 Diversity-driven Calibration Loss

The typical training with MLE lacks sequence-level objective and cannot directly optimize the model towards the desired goal (Zhao et al., 2023). We propose a diversity-driven calibration loss to provide sequence-level supervision and improve output diversity and semantic fidelity. Following Liu et al. 2022, we first train our model with the standard MLE objective. Then we sample multiple candidates from the fine-tuned model and design a multi-object calibration loss to align the model towards the desired goal.

As the goal is to improve surface diversity while maintaining semantic fidelity, we first design a multi-objective based scoring function to score each candidate:

$$S(\bar{y}) = \lambda_{sts} \cdot STS(\bar{y}) + \lambda_b \cdot BS(\bar{y}) + \lambda_s \cdot SD(y) \\ - \lambda_{r1} \cdot R_1(\bar{y}) - \lambda_{r2} \cdot R_2(\bar{y}) \tag{1}$$

where $\bar{y}$ is the candidate, $STS(*)$ represents sentence transformer similarity score (Reimers and Gurevych, 2019) calculated based on an off-the-shelf model [2], $BS(*)$ denotes BERT score, $SD(*)$

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

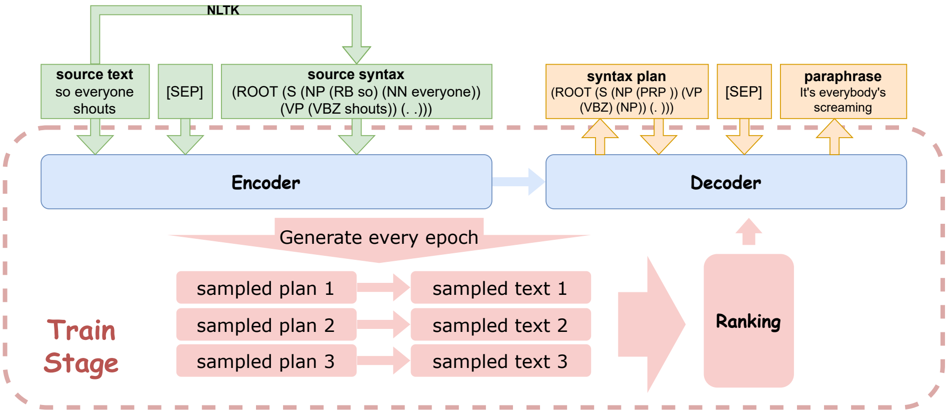

Figure 2: Model Architecture

is the syntax tree edit distance, and $R_1(*)$ and $R_2(*)$ stand for ROUGE-1/2. All scores are computed between $\bar{y}$ and the source input x, and we omit x for simplicity. λ_* are weights of each score, tuned as hyper-parameters.

To align the model outputs with the desired object, we propose a ranking-based calibration loss to optimize the model to assign higher probability to candidates with higher scores:

$$L_{cal} = \sum \max(logP(y_j|x) - logP(y_i|x) + |j-i|\lambda_{cal}, 0) \tag{2}$$

where y_i and y_j are two sampled candidate paraphrases with $S(y_i) > S(y_j), \forall i, j$. λ_{cal} is chosen empirically to control the margin.

The final loss is a combination of both token-level cross-entropy (L_{ce}) and sequence-level calibration (L_{cal}): $L = L_{ce} + \alpha L_{cal}$, where α is the weight. These two losses are complementary to each other where cross entropy ensures the model not deviate significantly from the reference while the calibration loss coordinates the model for better diversity.

4 DiverseQuora Dataset

Existing work on paraphrase generation mainly adopt NLI based dataset such as Quora and ParaNMT and convert the original paraphrase to a generation task (Yang et al., 2022; Li et al., 2018). However, the target paraphrases in these datasets usually have a high surface-form similarity as the source sentences, making them less applicable in our scenario. We introduce a new dataset, DiverseQuora, for diverse paraphrase generation.

Specifically, we sample about 10K source sentences from Quora (Kumar et al., 2020) dataset, and prompt ChatGPT to produce a diverse paraphrase candiate. We then filter the low quality paraphrases by verifying their semantic similarity re-prompting ChatGPT. The detailed prompts are in Appendix B.1. This yields 10,242 source-target pairs in total. To further validate the data quality, we randomly select 50 samples and manually check the sample quality, with the details in Appendix B.2. Finally, The data are split into train, validation and test sets, with the statistics reported in Table 1. As can be seen, the paraphrases in DiverseQuora have higher diversity compared to the existing datasets.

Dataset	Train	Val.	Test	Diversity
DiverseQuora	9,213	562	467	11.29
Quora	137,185	3,000	3,000	17.41
ParaNMT	493,081	500	800	18.53

Table 1: Statistics of DiverseQuora and existing paraphrase datasets. Diversity is measured by BLEU score between source and target, where lower score means better diversity.

5 Experiments

5.1 Datasets

Following previous works (Yang et al., 2022), we include ParaNMT-small (Chen et al., 2019) which is a subset of ParaNMT-50M dataset (Wieting and Gimpel, 2018). We also evaluated our model on the DiverseQuora dataset.

5.2 Baseline

Besides **Seq2Seq** where we directly finetune an encoder-decoder Transformer as a baseline, we further include the following comparison methods. The implementation details of both our models and baselines are in Appendix A.1.

Control Seq2Seq. This is a diversity-controlled model. We categorize the dataset into five subsets based on the edit distance to indicate the diversity, and then prepend the diversity as control codes. During inference, we use the highest diversity as the hint to generate diverse outputs.

ParaNMT						
	Diversity				Similarity	
Model	**R-2↓**	**BLEU↓**	**ED↑**	**SD↑**	**BERT↑**	**STS↑**
Reference	0.201	18.532	0.586	0.269	0.510	0.791
Seq2Seq	0.219	19.727	0.465	**0.246**	**0.490**	0.767
Control Seq2Seq	0.168	13.450	**0.607**	0.226	0.439	0.772
Seq2Seq Post Scoring	0.160	14.012	0.496	0.203	0.505	**0.774**
NADIA	**0.102**	**8.015**	0.558	0.224	0.436	0.741
w/o Calibration Loss	0.149	11.819	0.526	0.214	0.460	0.750
w/o Plan	0.131	12.025	0.533	0.218	0.455	0.750
DiverseQuora						
Seq2Seq	0.396	23.997	0.498	0.187	**0.643**	**0.886**
Control Seq2Seq	0.327	18.274	**0.605**	0.200	0.598	0.861
Seq2Seq Post Scoring	0.307	16.694	0.563	**0.224**	0.619	0.865
NADIA	**0.300**	**17.146**	0.604	0.209	0.594	0.854

Table 2: Experimental Results. ED stands for Edit Distance, SD stands for Syntax Tree Edit Distance,R-2 stands for Rouge F-1 Score, BERT stands for BERT Score, STS stands for Semantic Textual Similarity.

Input	Reference	NADIA
It is your first own studio.	It 's the first studio you have owned.	You've got your first studio!
It 's a big risk for him.	The risks for him are big.	He's taking a great risk.
Relax. Take it easy.	Just calm down.	Calm down, buddy.

Table 3: Three examples from **NADIA** output.

Seq2Seq Post Scoring. This is a post-scoring model, where we adopt Seq2Seq during training. In inference, we sample 8 outputs, and select the best one with the same scoring parameters as those used in the ordering loss, except that the syntax tree edit distance is replaced by edit distance.

5.3 Evaluation Metrics

We evaluate both diversity and semantic fidelity. For **surface diversity**, we adopt Rouge-2 and BLEU to measure the token overlap between source and generate paraphrase; we also include edit distance (ED) which calculates character level Hamming distance and syntax tree distance (SD) which computes tree edit distance(Zhang and Shasha, 1989) between the two syntax trees (keep only top 3 layers as in Figure 3) of source input and the generated paraphrase. For **semantic fidelity**, we leverage BERT score and Semantic Textual Similarity (all- MiniLM-L6-v2) to measure the similarity between source and paraphrase.

6 Results and Analysis

6.1 Automatic Results

Automatic metrics results are shown in Table 2. As can be seen, **NADIA** can generate outputs with both high surface-form diversity and semantic similarity, proving its effectiveness. Compared with Control Seq2Seq and Seq2Seq with Post Scoring, NADIA is able to achive a good balance of both two objectives, demonstrating its effectiveness of generating diverse paraphrase. Notably, compared with vanilla Seq2Seq that not pursuing diversity, our model has slightly lower BERT score and Semantic Textual Similarity. This is because changing the used words or word order inherently decreases these scores due to their order sensitivity. This is also evident in the scores between standard reference and input. Furthermore, after removing the planning or calibration loss, the results both drop, which show the effectiveness of the two components to jointly improve the model performance.

6.2 Case Study

We show sample outputs in Table 3 and Table 6. From the examples in Table 3, we can see our model output are more different to input sentence on syntax tree level(row 1 and row 2). Besides doing paraphrase, it is trying to leaking predicted information from pretrained model(row 3). This is because we use strategy to select less fine tuned checkpoint before combining with calibration loss.

7 Conclusion

We proposed a novel method to increase output diversity for the paraphrase task, which disentangles paraphrase generation into syntax planning and semantic realization. We further introduce a diversity-driven calibration loss to rank model generated outputs and enhance sequence-level diversity while maintaining semantic similarity. We propose a DiverseQuora dataset which is distilled from Large Language Model with diverse paraphrases. Experiments show that our model can generate both diverse and high-quality paraphrases compared to several strong baselines.

Limitations

Seeking diversity in paraphrase will intrinsically decrease some similarity scores like BERT score and Semantic Textual Similarity. Our model has slightly lower similarity metric compared to base seq2seq model. In the future, we will investigate how to find better metrics which can evict this issue. The ordering loss is hard to train on small dataset. In the future, we seek to make it easier to control.

References

Yue Cao and Xiaojun Wan. 2020. DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421, Online. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Yao Dou, Chao Jiang, and Wei Xu. 2022. Improving large-scale paraphrase acquisition and generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9301–9323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 238–252, Online. Association for Computational Linguistics.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2017. A deep generative framework for paraphrase generation.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:329–345.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020a. Revision in continuous space: Unsupervised text style transfer without adversarial learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8376–8383.

Mingtong Liu, Erguang Yang, Deyi Xiong, Yujie Zhang, Yao Meng, Changjian Hu, Jinan Xu, and Yufeng Chen. 2020b. A learning-exploring method to generate diverse paraphrases with multi-objective deep reinforcement learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2310–2321, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*, 36(3):341–387.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2923–2934, Osaka, Japan. The COLING 2016 Organizing Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Qiongkai Xu, Juyan Zhang, Lizhen Qu, Lexing Xie, and Richard Nock. 2018. D-page: Diverse paraphrase generation.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Haibo Zhang, Xue Zhao, Wenqing Yao, and Boxing Chen. 2022. GCPG: A general framework for controllable paraphrase generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4035–4047, Dublin, Ireland. Association for Computational Linguistics.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18:1245–1262.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Experiment Details

A.1 Implementation Details

All models are instantiated by BART using base size. During inference, beam size is set to 5, and length penalty is set to 1.0. In training, all 8 samples are sampled with temperature=1.2, $\lambda_{sts}, \lambda_b, \lambda_{r1}, \lambda_{r2}, \lambda_s$, are set to $1.0, 0.333, 1.0, 0.0, 0.2$ based on validation. α for L_{cal} is set to 1000. BERT score are calculated based on RoBERTa model, and Semantic Textual Similarity are calculated with “all-MiniLM-L6-v2”.

A.2 Explanation of Syntax Tree

We use NLTK [3] to compute the syntax tree of both source and target sentences. For source syntax tree, we do not trim the tree, and concatenate the source sentence and the source syntax tree as input. For target, we trim the target syntax tree to $height = 3$ in our implementation. An concrete example is shown in Figure 3.

[3] https://www.nltk.org/

Dataset	Avg. Quality	Std. Quality
DiverseQuora	4.00	1.01
Quora	2.92	1.07

Table 4: Human evaluation of the paraphrase.

B More detailed of DiverseQuora Dataset

B.1 Prompt for DiverseQuora Construction

We leverage ChatGPT to produce diverse paraphrase given an input sentence. Concretely, we first prompt ChatGPT to produce a paraphrase candidate with the prompt:

"Given a sentence: _input_. Please rewrite the sentence. You need to keep the semantic meaning unchanged, while making the surface form different compared to the original sentence. You can use synonyms or/and change the sentence structure to make them different towards surface form."

To ensure the semantic similarity of the generated paraphrase, we verify the quality by prompting the ChatGPT again with the following prompt:

"sentence 1: [_sent1_]; sentence 2:[_sent2_]; Do sentence 1 and sentence 2 have the same semantic meaning? Answer "yes" or "no":"

If the candidate does not satisfy the above condition, we will repeat the process.

B.2 DiverseQuora Quality Evaluation

We further evaluated the quality of DiverseQuora and original Quora using the evaluation criteria described in Table 5. For each dataset, we randomly selected 50 examples from the training set, hide source information, merge and random shuffle them, and then evaluate them using the evaluation criteria described in Table 5. The results are shown in Table 4. Our dataset is of higher quality than the original Quora dataset (Kumar et al., 2020). We also manually selected some examples to show the difference between DiverseQuora and Quora in Table 7. Because the Quora dataset is generated by filtering negative examples from the original Quora Question Pairs dataset, some pairs are not good paraphrases but rather similar questions (such as row 9).

C More Examples Generated by NADIA

Here are some examples generated by **NADIA**. Through leveraging the power of BART model and Calibration Loss, we generate some examples with diversity and good quality.

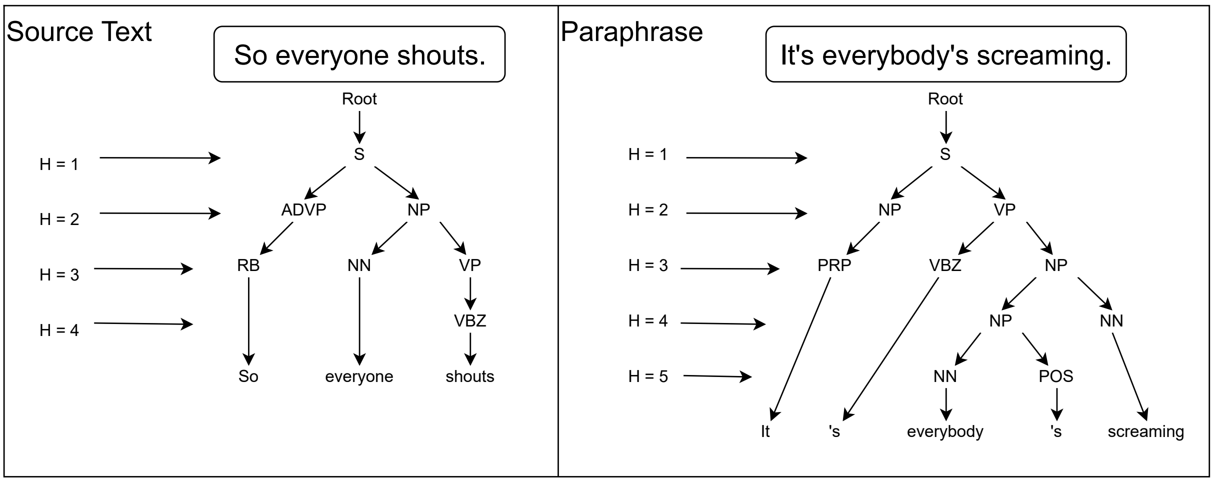

Figure 3: Showing the syntax tree examples. We select only top 3 layer in predicted syntax plan.

Paraphrase Quality Scale	Definition
5: Excellent	The paraphrase expresses the original meaning in a new and creative way, while still being accurate and fluent. It is clear that the paraphraser put a lot of thought into their work.
4: Good Expresse	Expresses the original meaning accurately and fluently. It is similar to what a human would generate within a few seconds.
3: Fair Generally	conveys the original meaning, but may be less fluent or original. It may contain some minor differences, such as removing unimportant information or adding well-known basic information.
2: Poor	Does not accurately convey the original meaning. It may introduce new information, lack important information, use too many of the same words as the original sentence, or is poorly written.
1: Very poor	Very similar to the original sentence, or expresses a very different meaning, or the paraphrase is difficult to understand.

Table 5: The human evaluation template.

Input	Reference	NADIA
It is your first own studio .	It's the first studio you have owned.	You've got your first studio!
The police think the bombing and today's gunfight in the apartments may be related.	The police suspect that the bombing may be tied to the apartment gunfight from earlier today.	The police think there was a connection between yesterday's explosion and today's shooting.
All this gold's gonna make things different.	Everything will be different with this gold.	The gold will change everything.
We will perform the opposite reaction to the plan he drew up .	He drew up the plan, so we 're going to perform the opposite reaction.	We're going to do a different reaction to this plan.
Let me show you to your seats.	I'll show you where you sit.	I'll show you the seat.
It's a big risk for him.	The risks for him are big.	He's taking a great risk.
Can't pass documents to this workspace.	Documents can not be forwarded to this workspace .	The document can not be transferred to the workspace.
What we want is to talk to your daddy.	We just want to talk to your dad.	We'd like to talk to your father.
Relax. Take it easy.	Just calm down.	Calm down, buddy.
Things have been getting a little weird around here.	There's some weird stuff going on.	It's getting kind of weird.
I have English , science , and chemistry books.	There are English books and science books and chemistry books.	I've got books in English, science and chemistry.

Table 6: More examples generated by **NADIA**.

Source	**Quora**	**DiverseQuora**
What are the benefits of a billing software?	What is the benefit of billing software?	What advantages does a billing software offer?
How do i get my likes and followers up on instagram?	How do i increase likes on instagram?	What strategies can I use to increase my likes and followers on Instagram?
How do i travel around the world without any money?	How can i travel without an id or money?	How can I journey around the globe without any funds?
Which laptop is best under 25000 inr?	Which is best laptop under 25000 with all features like vga and hdmi port?	What laptop is the optimal choice for under 25000 Indian rupees?
How do you take a screenshot on a mac laptop?	How do you take a screenshot on a mac?	What is the procedure to capture a screenshot on a mac laptop?
What happens if you actually drink bleach?	What happens when you swallow bleach?	What would occur if you ingested bleach?
How can i create a magnetic field?	How is a magnetic field created?	What steps do I need to take in order to generate a magnetic field?
Why was Hindi news channel NDTV India banned for one day?	What are your views on the government's decision of banning NDTV India for a day?	What was the reason for temporarily suspending the Hindi news channel NDTV India for one day?
Is the Aam Aadmi party's agenda economically comprehensive?	What is the agenda of Aam Aadmi party?	Does the Aam Aadmi Party's program possess an all-encompassing economic aspect?
How do i find ask someone to become a mentor?	How do i ask someone to be my mentor?	What would be the best way for me to request someone to be my mentor?

Table 7: Examples from DiverseQuora.