Multi-Task Instruction Training of Diffusion-Based Text Generative Models

Anonymous ACL submission

Abstract

Recent advancements in language models (LMs) have demonstrated remarkable adaptability across diverse tasks, excelling in both discriminative and generative domains with impressive multitasking capabilities. Recent attention of LMs has shifted towards non-autoregressive diffusion models, leveraging denoising generation for sequence-to-sequence modeling. However, the extent to which current diffusion-based LMs can handle multitasking remains unclear. In this study, we introduce a novel framework tailored to designing a diffusion model for multi-task language modeling. Inspired by latent image diffusion models, our approach involves a general transformer-based diffusion model leveraging pretrained encoders, facilitating multi-task learning with adaptable input embedding encoders. We define a diffusion loss within the trainable decoder's latent space, which interacts with any encoder via a cross-attention mechanism. This framework establishes a flexible non-autoregressive LM capable of handling potentially noisy data by leveraging robust instruction embeddings from encoders, enabling instruction tuning. We demonstrate the efficacy of our model across various setups, including single-task and multi-task scenarios, showing its ability to produce high-quality outputs by effectively utilizing and merging training task information in the continuous latent space.

1 Introduction

With the recent success of diffusion models in vision (Rombach et al., 2022a), there has been growing interest in adapting them to text, which have shown superior diversity in language generation while maintaining competitive quality compared to auto-regressive counterparts such as GPT (Radford et al., 2019) and T5 (Raffel et al., 2020). Additionally, diffusion models offer controllability either on semantic concepts (Li et al., 2022) or concatenated input sentences(Gong et al., 2022).

However, it is *unclear* whether diffusion models excel as multi-task learners in language modeling tasks, given their ability for diverse generation and controllability in the latent space. Furthermore, existing continuous text diffusion models, like DiffuSeq (Gong et al., 2022), constrain the input transformer's capacity by including input along with the latent space to be diffused, resulting in token wastage. SeqDiffuSeq (Yuan et al., 2022) fixes the encoder-decoder to a BART model and lacks a general cross-attention conditioning mechanism that can leverage any pre-trained encoder. In this work, we aim to address the following questions:

1. How can we enable effective multi-task learning in diffusion-based language models, given their promising capability for diverse generalization (Gong et al., 2022)?
2. How can we leverage state-of-the-art embedding modules/encoders in a plug-and-play fashion for multi-task learning with diffusion models, akin to image-counterparts such as Stable Diffusion (Rombach et al., 2022a)?

To address these challenges, we introduce a novel framework for multi-task learning with text diffusion models, leveraging any input representations through a general cross-attention mechanism. Our exploration encompasses architectural design and multi-task training techniques aimed at stabilizing and enhancing performance. We anticipate that our analysis will serve as an initial and valuable endeavor towards better understanding of multi-task language training with diffusion models.

2 Methodology

2.1 Latent Text Diffusion Model

A diffusion model (Ho et al., 2020) operates through a Markov chain of steps where the forward process gradually introduces noise to the original sample, and the model learns to *reverse* (*i.e.*, denoise) the noisy sample to reconstruct the

original. Initially, a sample $z_0 \sim p(z)$ is drawn from the distribution aimed to model. Then, a sequence of transformations is applied over T time steps to eventually produce $z_T \sim \mathcal{N}(0, 1)$. This transformation is defined as $q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I})$, where β_t governs the noise schedule, determining the rate at which noise is introduced to z_0 to transform it into z_T. During the reverse diffusion process, the model learns a denoising function $p_\theta(z_{t-1}|z_t) \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$, where μ and Σ are typically estimated using a model such as a U-Net (Ronneberger et al., 2015) or a transformer (Vaswani et al., 2017). In our implementation, we keep the learned variance fixed to the original variance schedule and predict the mean itself.

In our framework, we focus on sequence-to-sequence text generation tasks with *continous* text diffusion models, where given $(\mathbf{x}, \mathbf{y})$, we aim to generate the output sequence $\mathbf{y}$ conditioned on the input sequence $\mathbf{x}$ (NULL for unconditional generation). Additionally, we incorporate a task description $\mathbf{c}$ to facilitate multi-task learning. The discrete token inputs necessitate additional steps of embedding (conversion from text to continuous space in the forward process) and rounding (conversion from continuous space to text during the reverse process), which we adopt directly from (Li et al., 2022). Unlike standard diffusion, where $\mathbf{z}_0$ is sampled from a task-and-input conditioned distribution denoted as $\mathbf{z}_0 \sim q_0(\mathbf{z}_0|\mathbf{x}, \mathbf{c})$, in our approach, during the reverse process, we parameterize the reverse model (a Transformer) as $p_\theta(z_{t-1}|z_t; \mathbf{x}, \mathbf{c}) \sim \mathcal{N}(\mu_\theta(x_t, t, \mathbf{c}), \Sigma_\theta(x_t, t, \mathbf{c}))$. Following the standard DDPM, our objective is to align the joint distributions of the forward and reverse processes.

We follow DDPM and adopt the canonical *simple loss* for the learning of generating $\mathbf{y}$ via the continuous latent variable $\mathbf{z}$:

$$\mathcal{L}^{\text{e2e}}_{\text{simple}}(\mathbf{y}) = \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}_{0:T}|\mathbf{w})} \Big[\mathcal{L}_{\text{simple}}(\mathbf{z}_0) + \|\text{EMB}(\mathbf{w}) - \mu_\theta(\mathbf{z}_1, 1)\|^2 - \log p_\theta(\mathbf{y} \mid \mathbf{z}_0)\Big], \text{ with}$$

$$\mathcal{L}_{\text{simple}}(\mathbf{z}_0) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(\mathbf{z}_t|\mathbf{z}_0)} \|\mu_\theta(\mathbf{z}_t, t) - \hat{\mu}(\mathbf{z}_t, \mathbf{z}_0)\|^2$$

The second term of $\mathcal{L}^{\text{e2e}}_{\text{simple}}$ accounts for the embedding loss, while the third term ensures that the latent representations map back to words in the discrete space. It is important to highlight that diffusion modeling with text employs a technique for *non-autoregressive* generation: all tokens are generated simultaneously, leveraging the inherent stochasticity in the Gaussian sampling process to produce diverse samples.

2.2 The Proposed Architecture

To facilitate multi-task training, we introduce a novel architecture capable of conditioning the diffusion denoising process on input representations obtained from a pre-trained encoder. This is accomplished through a general cross-attention mechanism. Figure 1 illustrates our proposed architecture. Specifically, we leverage the BERT-base architecture and utilize a pre-trained BERT model to encode task information in the form of task instructions. The encoded task information is then combined with the source text information using the cross-attention mechanism within a new BERT architecture (termed *cross-attention BERT*). Finally, another BERT model is employed to decode the embeddings generated by the cross-attention BERT, producing coherent target text (termed *self-attention BERT*).

Cross Attention Mechanism One key innovation of our architecture, distinct from existing text diffusion models, is the incorporation of cross-attention to seamlessly integrate task information from a pre-trained BERT into the input source text. This approach offers flexibility as the cross-attention mechanism can accommodate task descriptions of varying lengths.

Training and Inference Details We utilize the BERT-base transformer, comprising 12 layers, with an embedding dimension set to 128. During training, we employ $T = 2000$ diffusion steps with a square-root noise schedule, and a learning rate of 1e-5. Additionally, we implement importance sampling, as proposed by Nichol and Dhariwal (2021), to train our diffusion model. This technique optimizes the loss by assigning higher weight to timesteps with greater loss, following the approach originally proposed by Nichol and Dhariwal (2021). To facilitate the mapping of discrete text into the latent space and back, we adhere to a similar process of embedding and rounding as proposed in the work by (Li et al., 2022). During embedding, discrete tokens are mapped into a continuous space as the initial stage of the diffusion process. Similarly, during the denoising phase, the continuous latent space is mapped back to the discrete vocabulary.

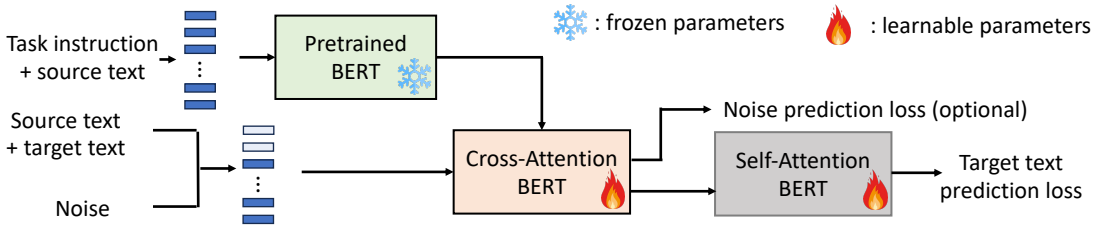

Figure 1: The proposed model for multi-task instruction training of diffusion models based on the BERT architecture.

Consistent with the observations in (Li et al., 2022), we directly model the mean μ instead of the noise.

Sampling Details Following Li et al. (2022), we sample for $T = 2000$ steps and leverage the clamping trick during sampling to ensure that each representation commits to specific words. We also follow the Minimum Bayes Risk (MBR) decoding (Koehn, 2004) which chooses the output sentence minimizing a loss such as the BLEU score.

3 Experiment Results

3.1 Datasets and Tasks

To implement our multi-task training paradigm, we first pretrain our models on the *Super Natural Instructions v2 (NIV2)* dataset (Wang et al., 2022). NIV2 is a diverse dataset comprising approximately 1600 tasks across 76 task-types, designed to evaluate generalization to unseen tasks. As NIV2 includes multiple languages but the tasks specifically require English, we filter out non-English data. We utilize the instruction templates provided by the dataset creators.

After pretraining, we further fine-tune the model on four standard datasets for single-task evaluation, namely **Text Paraphrasing, Question Generation, Text Simplification,** and **Open Domain Dialogue**. These datasets are commonly used for evaluating single-task diffusion models such as Diffuseq (Gong et al., 2022) and SeqDiffuSeq (Yuan et al., 2022).

3.2 Evaluation Setting

Baselines We consider several groups of baselines for comparison:

- Auto-regressive language models, including Transformer (Vaswani et al., 2017) (size 110M, base version) and pre-trained GPT2 (Radford et al., 2019).
- Non-autoregressive models: LevT (Gu et al., 2019) (an iterative NAR model), and recently proposed seq2seq text diffusion models such as encoder-only DiffuSeq (Gong et al., 2022) and encoder-decoder based SeqDiffuSeq (Yuan et al., 2022).

Note that for this study, the encoders are set to a pretrained BERT (Devlin et al., 2018) (BERT Base)[1].

Evaluation Metrics To measure the generation quality, we report the BLEU (Papineni et al., 2002), Rouge-L (Lin, 2004) and BERT Score (Zhang et al., 2019). Following Yuan et al. (2022), we also report the diversity of generation by reporting the intra-sentence dist-1 (distinct unigram).

3.3 Main results

The results are summarized in Table 1. Our method outperforms all baselines in most metrics. Compared with the auto-regressive based methods, our method achieves better (if not comparable) performance while using much smaller models. Compared to current state-of-the-art diffusion-based models, our model is also able to achieve better scores in most cases, partly due to the ability of learning from multiple tasks.

3.4 Ablation Studies

Single task vs. multi-task training We compare our model trained with and without multi-task pretraining. The results are shown in Table 2, where we observe that adopting multi-task pretraining significantly improves the performance[2].

Impact of MBR size We investigate the impact of the MBR decoding sample sizes. The results are shown in Figures 2 - 5 in the appendix. Our model consistently outperforms Diffuseq in most cases, as well as demonstrating a less sensitivity to the MBR sample sizes.

[1]Note that Gong et al. (2022) evaluate the scenario when embeddings are fixed with a BERT-Tiny model, but this can lead to incompatible representations in the same diffusion space due to which it is important to consider a separate encoder and decoder.

[2]We find that without multi-task pretraining, the model is usually not stable and can encounter numerical issues, in which case we evaluate the model using the checkpoint before encountering the problem.

Method	Paraphrase Task				Question Generation Task			
	BLEU↑	ROUGE-L↑	BERTScore↑	dist-1↑	BLEU↑	ROUGE-L↑	BERTScore↑	dist-1↑
Transformer-base	0.2722	0.5748	0.8381	0.9748	0.1663	0.3441	0.6307	0.9309
GPT2-base FT	0.1980	0.5212	0.8246	0.9798	0.0741	0.2714	0.6052	0.9602
GPT2-large FT	0.2059	0.5415	0.8363	0.9819	0.1110	0.3215	0.6346	0.9670
LevT	0.2268	0.5794	0.8344	0.9790	0.0930	0.2893	0.5491	0.8914
DiffuSeq	0.1805	0.55274	0.7910	0.9734	0.1511	0.3473	0.5882	0.9158
DiffuSeq (w/MBR=10)	0.2413	0.5880	0.8365	0.9807	0.1654	0.3677	0.6035	0.9106
SeqDiffuSeq	0.2328	–	0.8291	0.9806	0.1720	–	0.6135	0.9270
SeqDiffuSeq (w/ MBR=10)	0.2434	–	0.8400	0.9807	0.1746	–	0.6174	0.9248
Ours	0.2206	0.5711	0.8234	0.9713	0.1690	0.3646	0.6041	0.9147
Ours (w/MBR=10)	0.2563	0.6010	0.8489	0.9786	0.1757	0.3723	0.6126	0.9069

Method	Text Simplification Task				Open Domain Dialogue Task			
	BLEU↑	ROUGE-L↑	BERTScore↑	dist-1↑	BLEU↑	ROUGE-L↑	BERTScore↑	dist-1↑
Transformer-base	0.2693	0.4907	0.7381	0.8886	0.0189	0.1039	0.4781	0.7493
GPT2-base FT	0.3083	0.5461	0.8021	0.9439	0.0108	0.1508	0.5279	0.9194
GPT2-large FT	0.2693	0.5111	0.7882	0.9464	0.0125	0.1002	0.5293	0.9244
LevT	0.2052	0.4402	0.7254	0.9715	0.0158	0.0550	0.4760	0.9726
DiffuSeq	0.2873	0.5289	0.7771	0.9270	0.0088	0.0916	0.5106	0.9539
DiffuSeq (w/MBR=10)	0.3641	0.5869	0.8137	0.9264	0.0117	0.1103	0.5210	0.9425
SeqDiffuSeq	0.3709	–	0.8211	0.9081	0.0084	–	0.4382	0.9650
SeqDiffuSeq (w/MBR=10)	0.3712	–	0.8214	0.9077	0.0112	–	0.4425	0.9608
Ours	0.3225	0.5560	0.7928	0.9231	0.0120	0.1049	0.4953	0.9024
Ours (w/MBR=10)	0.3752	0.5961	0.8201	0.9196	0.0150	0.1168	0.5023	0.8762

Table 1: Comparisons with various baselines, including traditional autoregressive models and representative non-autoregressive diffusion models on four downstream tasks.

Method	Paraphrase Task				Question Generation Task			
	BLEU	ROUGE-L	BERTScore	dist-1	BLEU	ROUGE-L	BERTScore	dist-1
Single-Task (no MBR)	0.1847	0.5231	0.7978	0.9611	0.1770	0.3566	0.6215	0.9201
Multi-Task (no MBR)	0.2206	0.5711	0.8234	0.9713	0.1690	0.3646	0.6041	0.9147
Single-Task (w/MBR=10)	0.1970	0.5329	0.8101	0.9665	0.0459	0.1525	0.5066	0.9119
Multi-Task (w/MBR=10)	0.2563	0.6010	0.8489	0.9786	0.1757	0.3723	0.6126	0.9069

Method	Text Simplification Task				Open Domain Dialogue Task			
	BLEU	ROUGE-L	BERTScore	dist-1	BLEU	ROUGE-L	BERTScore	dist-1
Single-Task (no MBR)	0.2406	0.4811	0.7531	0.9536	0.0112	0.1003	0.5036	0.9105
Multi-Task (no MBR)	0.3225	0.5560	0.7928	0.9231	0.0120	0.1049	0.4953	0.9024
Single-Task (w/MBR=10)	0.2976	0.5329	0.7845	0.9512	0.0138	0.1130	0.5080	0.8820
Multi-Task (w/MBR=10)	0.3752	0.5961	0.8201	0.9196	0.0150	0.1168	0.5023	0.8762

Table 2: Single task versus multi-task training. Multi-task training significantly boosts the performance.

4 Conclusion

We propose the first framework to enable multi-task learning with text-diffusion models. We introduce a novel architecture to leverage any pre-trained representations to enable multi-task learning combined with the diverse generation capability of text diffusion models. Our experiments demonstrate competitive performance w.r.t AR and NAR baselines, demonstrating the efficacy of our approach.

Limitation Due to resource restrictions, we were only able to build and train our model on a relatively small scale. Compared to autoregressive LLMs with billions of parameters, our diffusion-based model is too small to observe emergent abilities, and the evaluation is limited to small-scale datasets for select tasks. We have tried some initial effort to scale up our model; however, we encountered multiple challenges such as training instability and slow convergence compared to autoregressive LLMs. Overcoming these challenges to scale up text diffusion models is an interesting research direction worthy of further investigation.

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

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*.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Jiatao Gu, Changhan Wang, and Jake Zhao. 2019. Levenshtein transformer. *CoRR*, abs/1905.11006.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. 2021. Step-unrolled denoising autoencoders for text generation. *arXiv preprint arXiv:2112.06749*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2022. Seqdiffuseq: Text diffusion with encoder-decoder transformers. *arXiv preprint arXiv:2212.10325*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Related Work

Discrete Text Diffusion Models Diffusion models have exhibited remarkable success within the realm of continuous data, particularly in the context of image generation (Rombach et al., 2022b), which has contributed to the generation of high-quality images. More recently, diffusion models have garnered attention for text generation as well. Taking inspiration from the diffusion model, Sohl-Dickstein et al. (2015) initially introduced a diffusion process involving binary random variables. Building upon this, Hoogeboom et al. (2021) expanded the diffusion model to encompass categorical data by incorporating surjective flow layers. Austin et al. (2021) further generalized the multinomial diffusion model proposed by Hoogeboom et al. (2021) by introducing corruption through transition matrices instead of uniform transition probabilities.

The application of densing diffusion techniques in non-autoregressive text generative models led to remarkable outcomes in machine translation tasks, as demonstrated by Savinov et al. (2021), achieving state-of-the-art results. In contrast to modeling the discrete state space in textual data, Li et al. (2022); Gong et al. (2022); Yuan et al. (2022) adopted diffusion models within the embedded latent space.

Continuous Text Diffusion Models Recent work has focused on diffusion modelling in the continuous space. Li et al. (2022) proposed a continuous diffusion model for improved controllable generation by leveraging a pre-trained classifier, further introducing the notions of 'embedding' and 'rounding' to ensure compatibility of the discrete token space with the continuous latent space. Following this, Gong et al. (2022) proposed DiffuSeq, a continuous diffusion model aimed at conducting sequence to sequence generation by concatenating the input and output sequences and only partially noising/denoising the output during the diffusion process. It should be noted that they leverage only a single transformer for the generation and only add noise to the output, which essentially wastes the entire sequence length which the model is capable of generating. Further, Yuan et al. (2022) propose SeqDiffuSeq, an encoder-decoder based model along with a different noise schedule and token based noising for conducting the diffusion process. They train both components simultaneously, and fix the architecture to a BART (Lewis et al., 2019) model, not demonstrating how one can leverage pre-trained representations.

B Extra Experimental Results

Figures 2, 3, 4 and 5 show the comparisons on the four evaluations metrics on the four datasets with increasing MBR sample sizes.

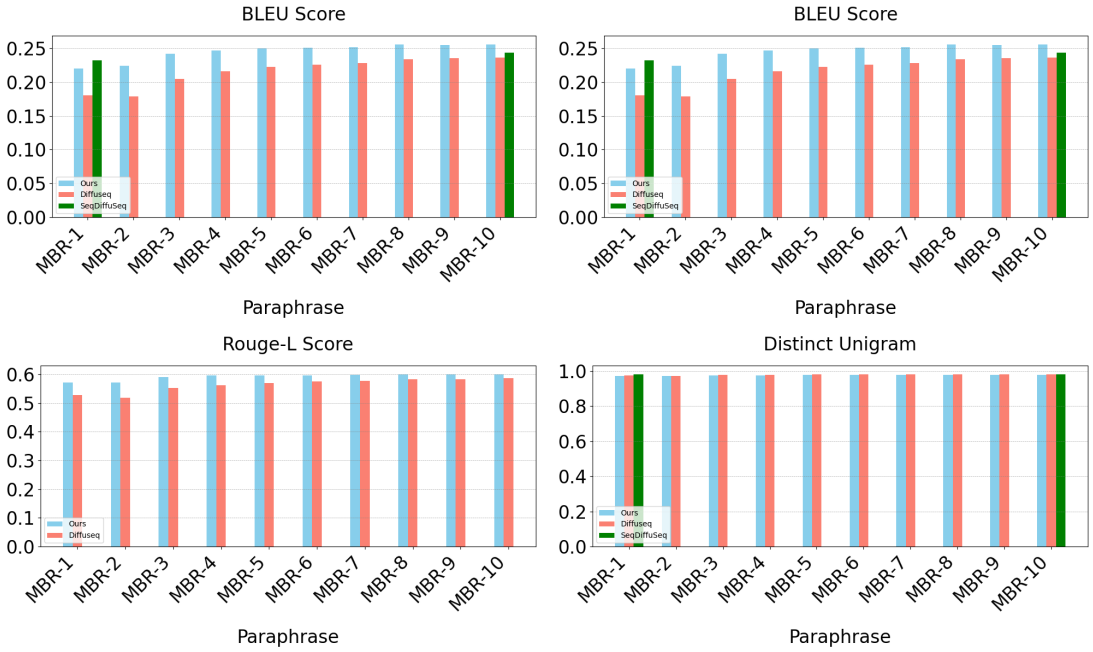

Figure 2: Comparisons of BERT, BLEU, Rouge-L and dist-1 scores with increasing MBR sample sizes on Paraphrase task.

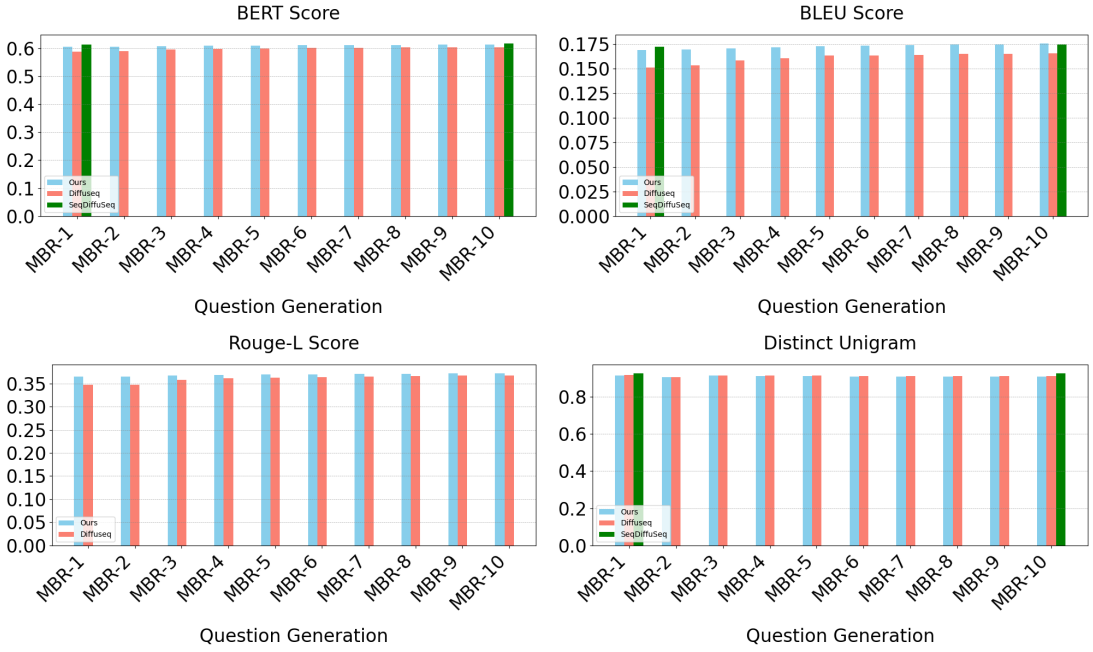

Figure 3: Comparisons of BERT, BLEU, Rouge-L and dist-1 scores with increasing MBR sample sizes on Question Generation task.

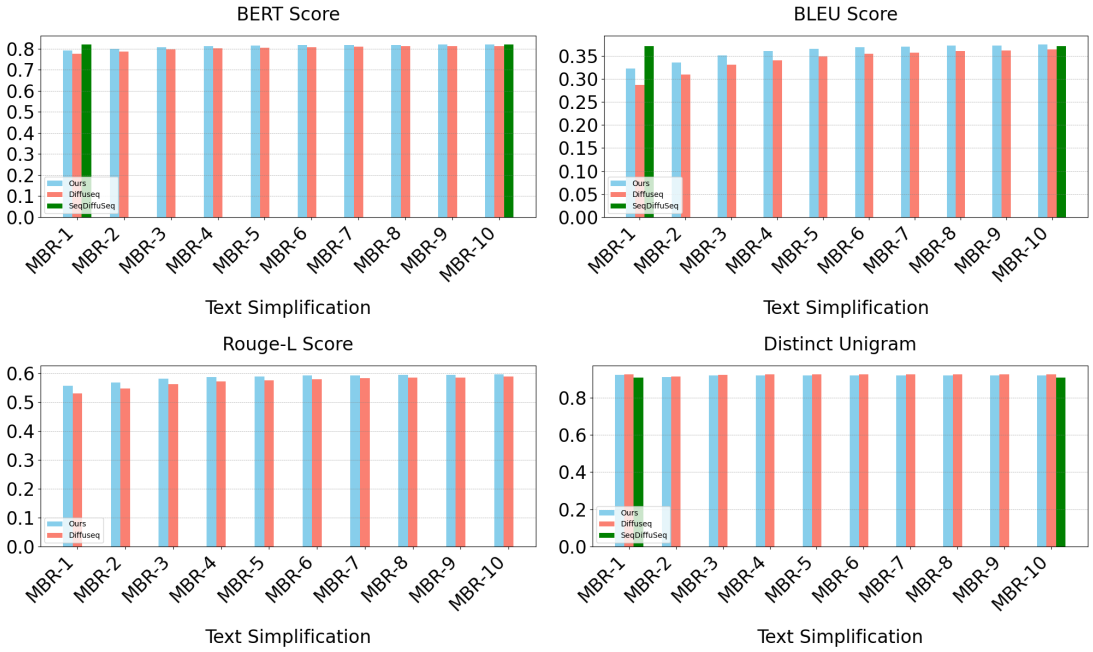

Figure 4: Comparisons of BERT, BLEU, Rouge-L and dist-1 scores with increasing MBR sample sizes on Text Simplification task.

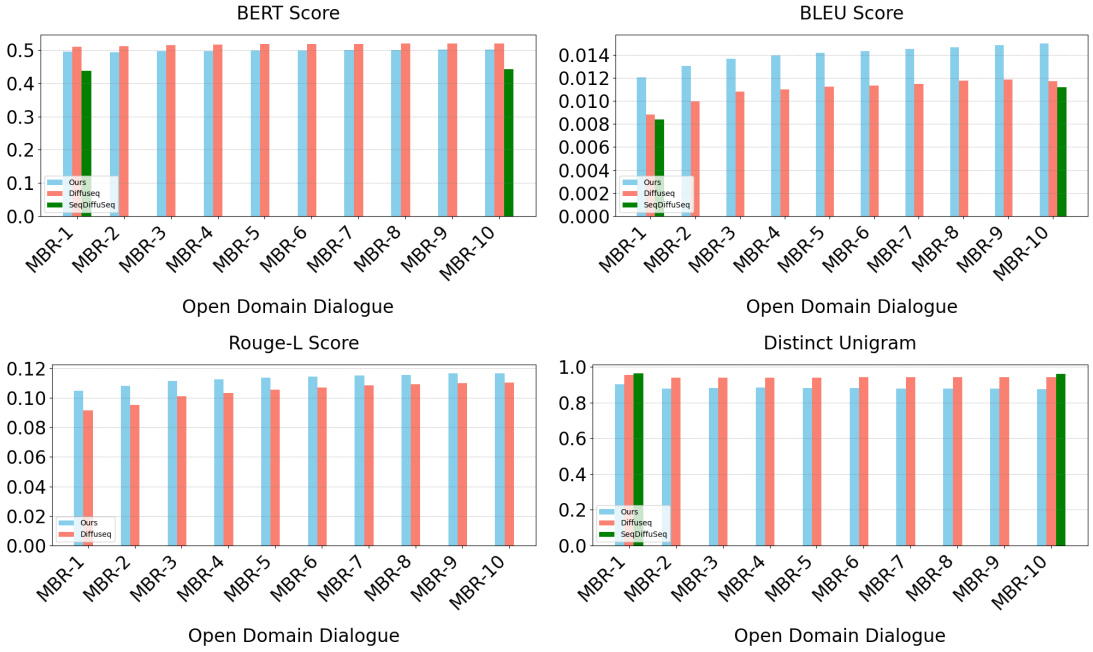

Figure 5: Comparisons of BERT, BLEU, Rouge-L and dist-1 scores with increasing MBR sample sizes on Open Domain Dialogue task.