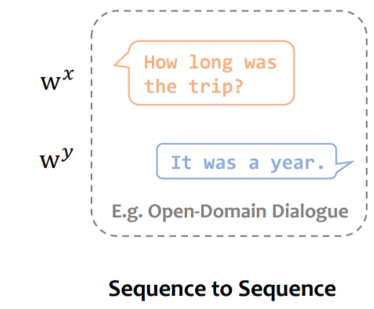
DiffuSeq: Sequence to Sequence Text Generation With Diffusion Models

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, Lingpeng Kong

Shark-NLP
Shanghai, China
hisansas@gmail.com



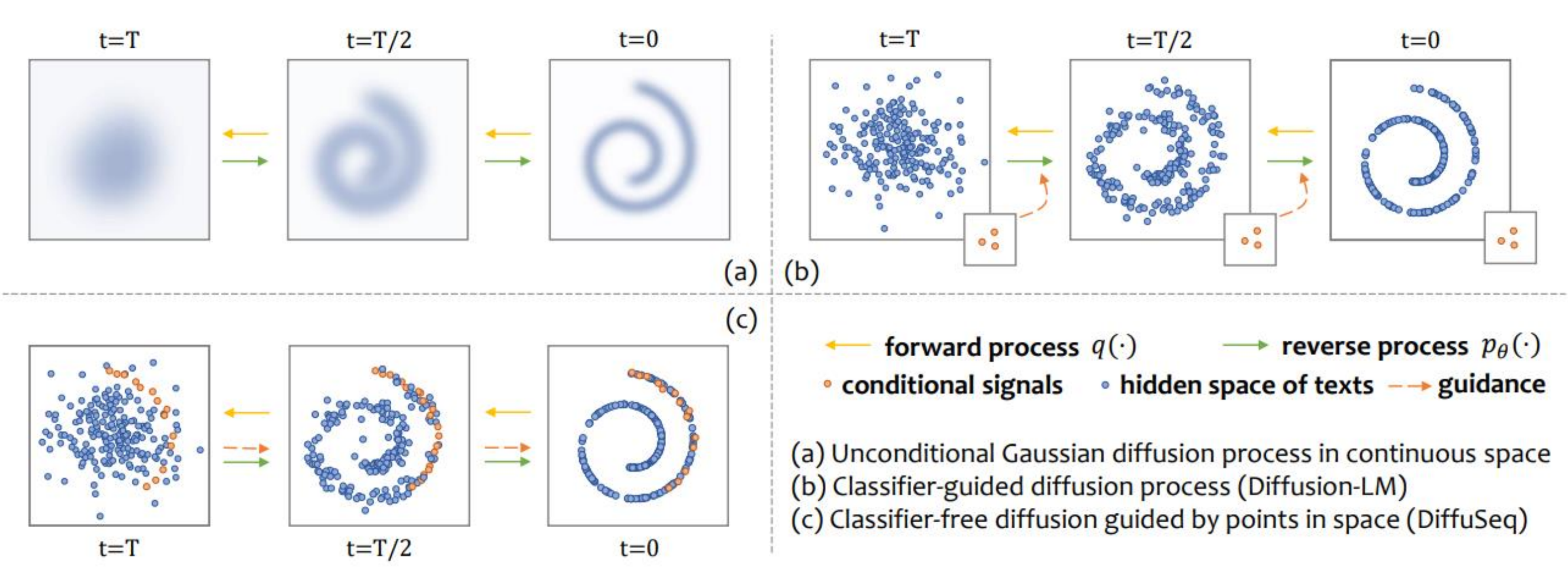
Background and contribution highlights

Recently, diffusion models have emerged as a new paradigm for generative models. Despite the success in domains using continuous signals such as vision and audio, adapting diffusion models to natural language is under-explored due to the discrete nature of texts, especially for the conditional generation. We tackle this challenge by proposing DiffuSeq: a diffusion model designed for sequence-to-sequence (Seq2Seq) text generation tasks.

- Our proposed DiffuSeq as a conditional language model is trained end-to-end in a classifier-free manner.
- We establish a theoretical connection among autoregressive (AR), non-autoregressive (NAR) and DiffuSeq models.
- DiffuSeq is a powerful model for text generation, matching or even surpassing competitive AR, iterative NAR, and large-PLMs on quality and diversity.

Connections to previous work

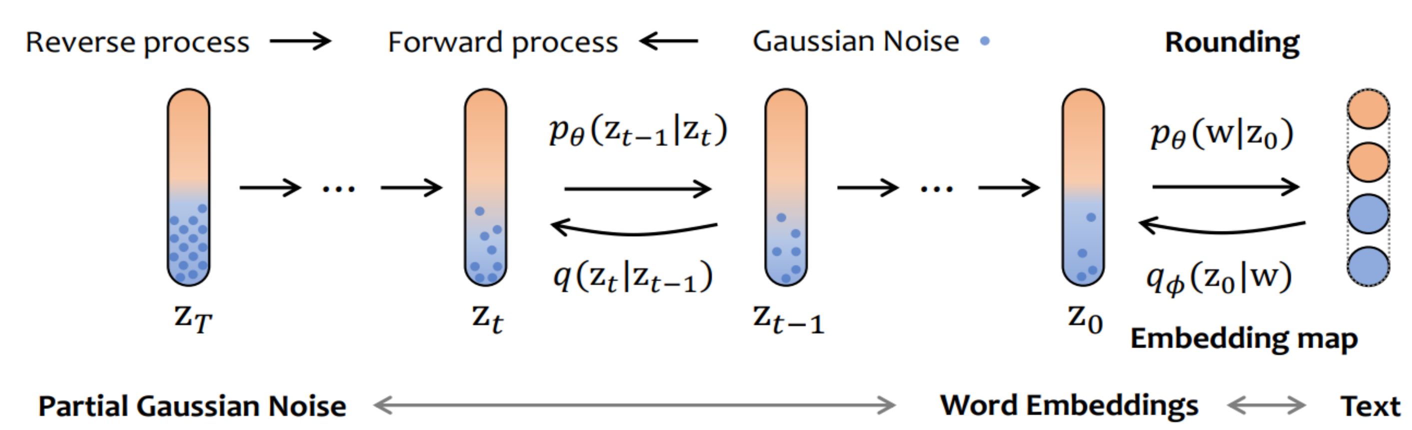
For conditional text generation, Diffusion-LM uses an extra-trained classifier to provide guidance but only adds fine-grained constraints on the generated outputs. In the more general seq2seq setting, applying it can be challenging. To address this, we propose a model DiffuSeq. It captures input guidance using a single model and doesn't rely on a separate classifier.



The demonstration of unconditional, classifier-guided, and classifier-free diffusion models

Methods for DiffuSeq

DiffuSeq learns a unified feature space by concatenating the space of x and y into z, and the embedding space is jointly trained. In the forward process, for z_t, we only impose partial noise on the space of y. The x signal stays in an un-noised state. In the reverse process, the neural network is optimized with the help of conditional signals x as guidance.



During training, we employ importance sampling to ensure sufficient training for more difficult data.
During inference, in addition to the rounding operations, we use an additional anchoring operation that replaces the recovered x part with the original x_0 to ensure that the x part remains un-noised.

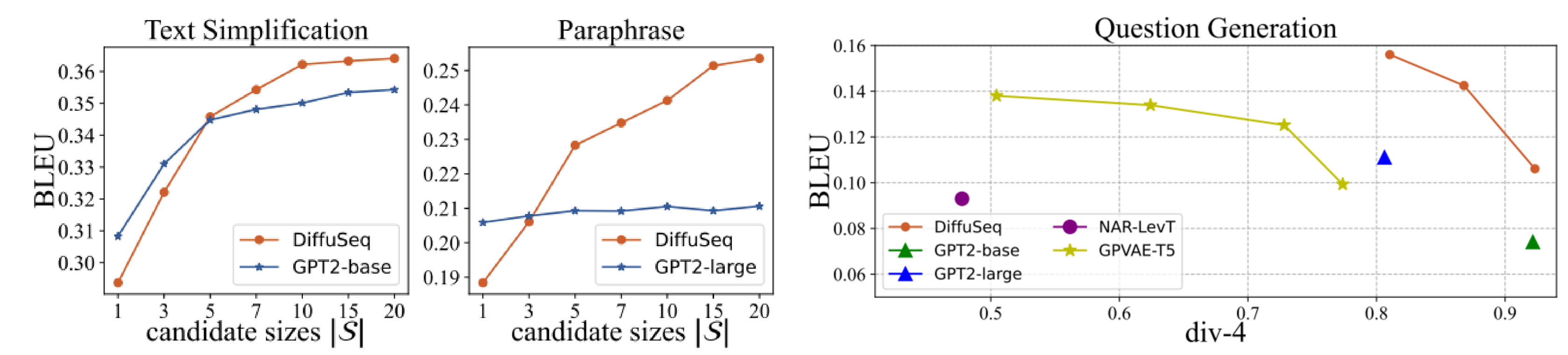
Selected Results for Different Seq2Seq Tasks

Our results demonstrate that DiffuSeq achieves comparable or even higher generation quality. Besides, it consistently shows its superiority in generating diverse outputs given the same input sequence. This is a desirable property in many NLG applications.

Tasks	Methods	BLEU↑	R-L↑	Score↑	dist-1↑	selfB↓ / div-4↑	Len
Open Domain Dialogue	GRU-attention ◇	0.0068	0.1054	0.4128	0.8998	0.8008/0.1824	4.46
	Transformer-base ◇	**0.0189**	0.1039	0.4781	0.7493	0.3698/0.6472	19.5
	GPT2-base FT •	0.0108	**0.1508**	0.5279	0.9194	0.0182/0.9919	16.8
	GPT2-large FT •	0.0125	0.1002	**0.5293**	0.9244	0.0213/0.9938	16.8
	GPVAE-T5 •	0.0110	0.1009	0.4317	0.5625	0.3560/0.5551	20.1
	NAR-LevT ‡	0.0158	0.0550	0.4760	**0.9726**	0.7103/0.1416	4.11
	DiffuSeq (Ours) ‡	0.0139	0.1056	0.5131	0.9467	**0.0144/0.9971**	13.6
Question Generation	GRU-attention ◇	0.0651	0.2617	0.5222	0.7930	0.9999/0.3178	10.1
	Transformer-base ◇	0.1663	0.3441	0.6307	0.9309	0.3265/0.7720	10.3
	GPT2-base FT •	0.0741	0.2714	0.6052	0.9602	**0.1403/0.9216**	10.0
	GPT2-large FT •	0.1110	0.3215	**0.6346**	**0.9670**	0.2910/0.8062	9.96
	GPVAE-T5 •	0.1251	0.3390	0.6308	0.9381	0.3567/0.7282	11.4
	NAR-LevT ‡	0.0930	0.2893	0.5491	0.8914	0.9830/0.4776	6.93
	DiffuSeq (Ours) ‡	**0.1731**	**0.3665**	0.6123	0.9056	0.2789/0.8103	11.5
Text Simplification	GRU-attention ◇	0.3256	0.5602	0.7871	0.8883	0.9998/0.3313	18.9
	Transformer-base ◇	0.2693	0.4907	0.7381	0.8886	0.6924/0.5095	18.5
	GPT2-base FT •	0.3083	0.5461	0.8021	0.9439	0.5444/0.6047	16.1
	GPT2-large FT •	0.2693	0.5111	0.7882	0.9464	0.6042/0.5876	15.4
	GPVAE-T5 •	0.3392	0.5828	**0.8166**	0.9308	0.8147/0.4355	18.5
	NAR-LevT ‡	0.2052	0.4402	0.7254	**0.9715**	0.9907/0.3271	8.31
	DiffuSeq (Ours) ‡	**0.3622**	**0.5849**	0.8126	0.9264	**0.4642/0.6604**	17.7

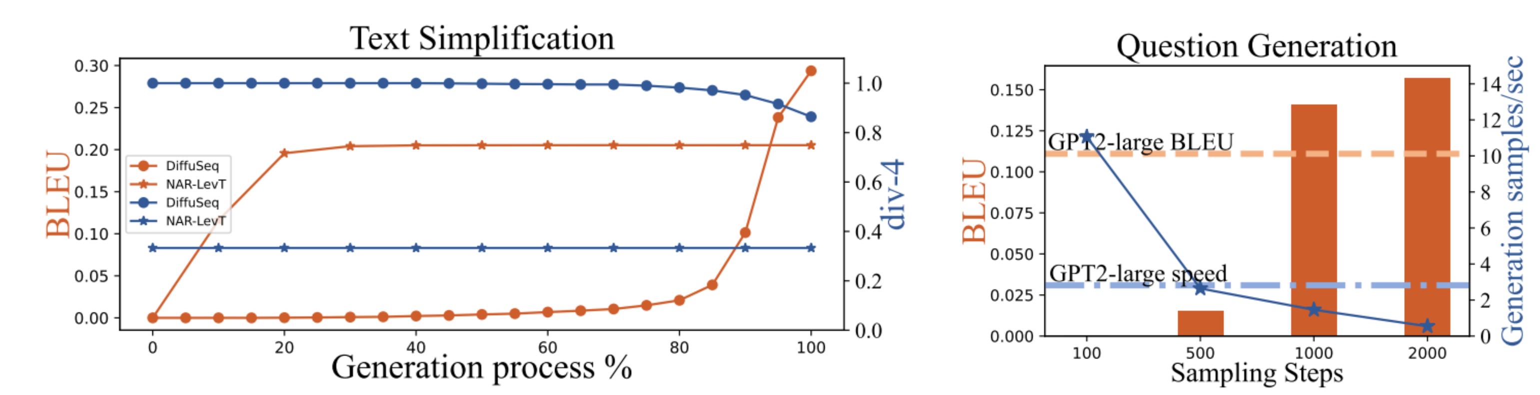
Diversity Analyses

We conducted Minimum Bayes Risk (MBR) on candidate sets with different sizes to select the best result. We observed that with an increase in size, the quality score increases as well. This rising trend is more evident for DiffuSeq than GPT2. This is because GPT2 tends to generate highly similar candidates, which impedes the effectiveness of MBR.
The right most figure validates this by showing that DiffuSeq enjoys better quality-diversity trade-offs than other strong baselines.



The increase of BLEU score with different candidate sizes, and the trade-off between quality and diversity

We investigate LevT and DiffuSeq's step-wise quality and diversity curves. It appears that DiffuSeq tends to explore more possible results in the first half of the generation process and converges to several potential candidates when it is close to the end of the steps, so the quality score rises in the end and the diversity score is always high. This is likely due to the noise injected in the generation process.



The curve of BLEU/div-4 score along with generation process (percentage of steps). Inference speed

As for the inference speed. By reducing the number of diffusion steps to 1,000, DiffuSeq can provide a good balance between quality and speed in practical scenarios.

References

Ho, J., Jain, A., & Abbeel, P. (2020). "Denoising diffusion probabilistic models.". *Advances in Neural Information Processing Systems, 33*, 6840-6851.

Song, J., Meng, C., & Ermon, S. (2020). "Denoising Diffusion Implicit Models.". In *International Conference on Learning Representations.*

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., & Hashimoto, T. B. (2022). "Diffusion-LM improves controllable text generation.". *Advances in Neural Information Processing Systems, 35*, 4328-4343.

Ho, J., & Salimans, T. (2021) "Classifier-Free Diffusion Guidance.". In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.*

DiffuSeq Github Repo