DiffuSeq:
Sequence to Sequence Text Generation With Diffusion Models

Shark-NLP
Shanghai, China
hisansas@gmail.com

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

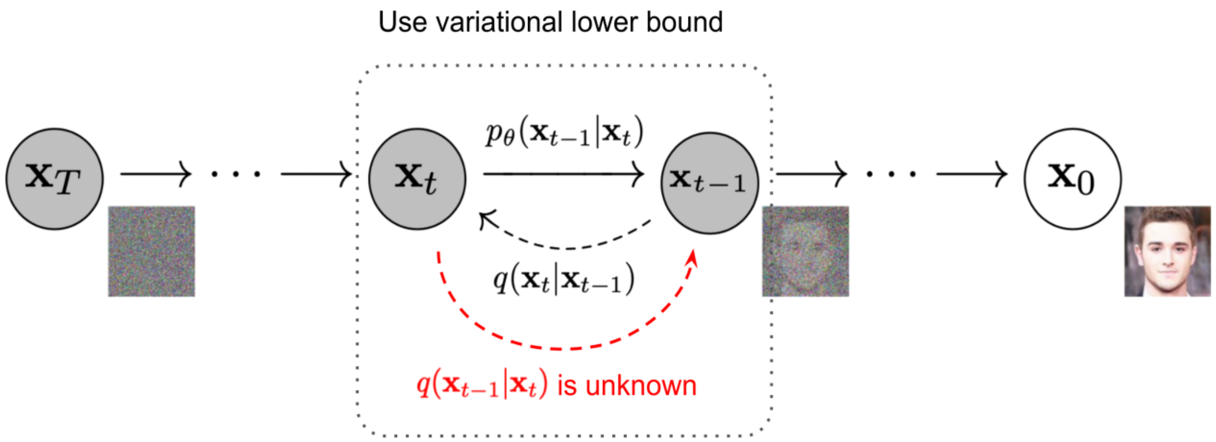
https://github.com/Shark-NLP/DiffuSeq

Contents

1.1 Preliminary

Diffusion process in continuous space:

(applied in vision, audio, time series and etc.…)



1. Noise-conditioned score network (NCSN; Yang & Ermon, 2019)
2. Denoising diffusion probabilistic models (DDPM; Ho et al. 2020)

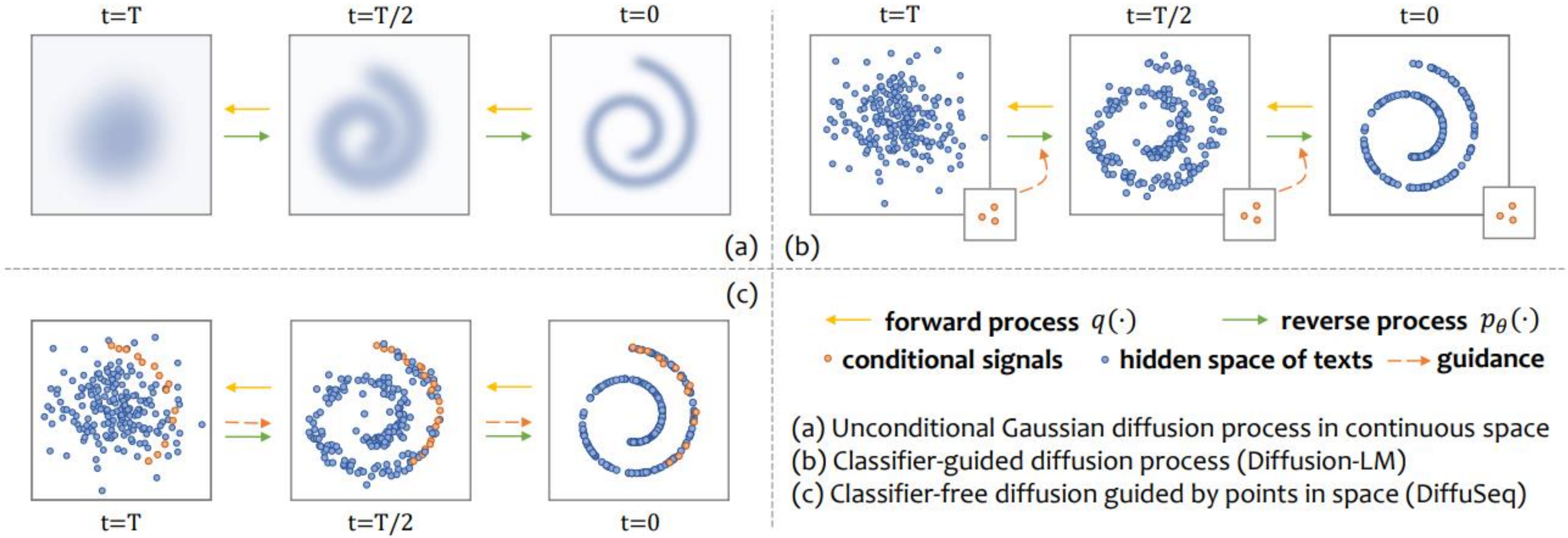
- Forward process:
  - $\mathbf{x}_0 \sim q(\mathbf{x}) \rightarrow \mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
  - $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t \mathbf{I})$
- Reverse process:
  - $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t))$
  - $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}$
- Training loss:
  - $L_t = D_{KL}(q||p_\theta)$
  - Parameterization of $L_t =$

$$\mathbb{E}_{\mathbf{x}_0}(||(\mathbf{x}_0 - f_\theta(\mathbf{x}_t, t))||^2)$$

1.2 Motivation

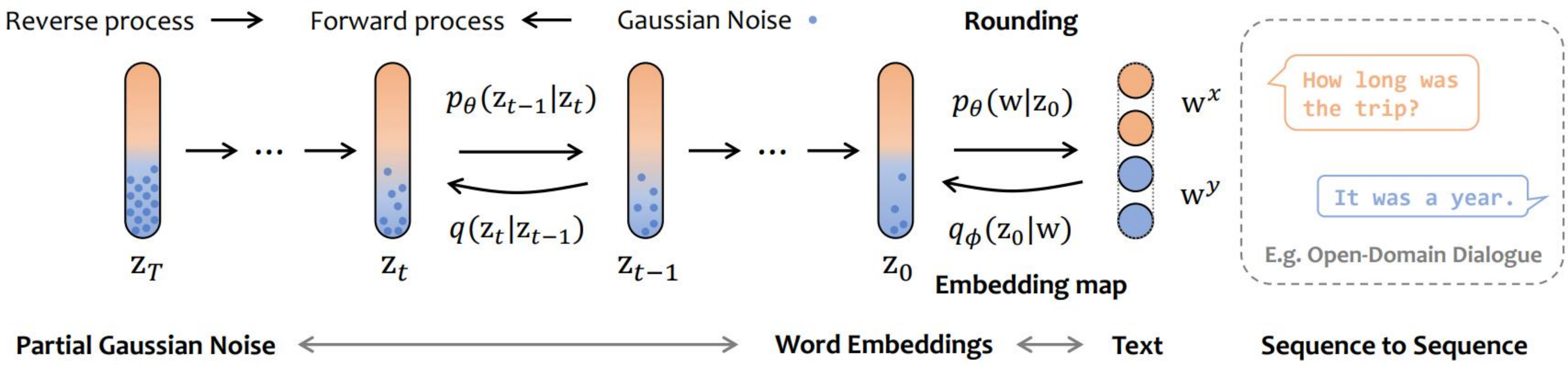
From unconditional models to conditional models:

Diffusion-LM (classifier-guided) v.s. DiffuSeq (classifier-free)



Seq2Seq tasks: $\mathbf{x} \rightarrow \mathbf{y}$

2.1 DiffuSeq



- Forward Process with Partial Noising:
  - $q(\mathbf{z}_0|\mathbf{w}^{x\oplus y}) = \mathcal{N}\left(\mathrm{EMB}(\mathbf{w}^{x\oplus y}), \beta_0\mathbf{I}\right); \mathbf{z}_t = \mathbf{x}_t \oplus \mathbf{y}_t$
- Reverse Process with Conditional Denoising:
  - $L_t = \mathbb{E}_{\mathbf{x}_0, \mathbf{y}_0}(||(\mathbf{y}_0 - f_{\tilde{\theta}}(\mathbf{z}_t, t))||^2)$

- Training:
  - importance sampling
- Inference:
  - Rounding to embeddings
  - Anchoring input signals

2.2 Connections of different models

AR/iter-NAR/DiffuSeq: Generation process is along with different dimensions:

$$p_{\text{AR}}(\mathbf{w}^y_{1:n}|\mathbf{w}^x) = \underbrace{p(w^y_1|\mathbf{w}^x)}_{\text{initial prediction}} \underbrace{\prod_{i=1,\ldots,n-1} p(w^y_{i+1}|\mathbf{w}^y_{1:i}, \mathbf{w}^x)}_{\text{progressive left-context prediction}},$$

$$p_{\text{iter-NAR}}(\mathbf{w}^y_{1:n}|\mathbf{w}^x) = \sum_{\mathbf{w}^y_1,\ldots,\mathbf{w}^y_{K-1}} \underbrace{\prod_{i=1\ldots n} p(w^y_{1,i}|\mathbf{w}^x)}_{\text{initial prediction}} \underbrace{\prod_{k=1..K-1} \prod_{i=1\ldots n} p(w^y_{k+1,i}|\mathbf{w}^y_{k,1:n}, \mathbf{w}^x)}_{\text{progressive full-context prediction}}.$$

$$p_{\text{DIFFUSEQ}}(\mathbf{w}^y|\mathbf{w}^x) = \sum_{\substack{\mathbf{w}^y_T,\ldots,\mathbf{w}^y_1 \\ \mathbf{y}_T,\ldots,\mathbf{y}_0}} p(\mathbf{w}^y|\mathbf{y}_0, \mathbf{w}^x) \prod_{t=T,\ldots,1} p(\mathbf{w}^y_t|\mathbf{y}_t, \mathbf{w}^x) p(\mathbf{y}_{t-1}|\mathbf{w}^y_t)$$

3.1 Experiments

Four tasks: Dialogue, QG, Text Simplification, Paraphrase

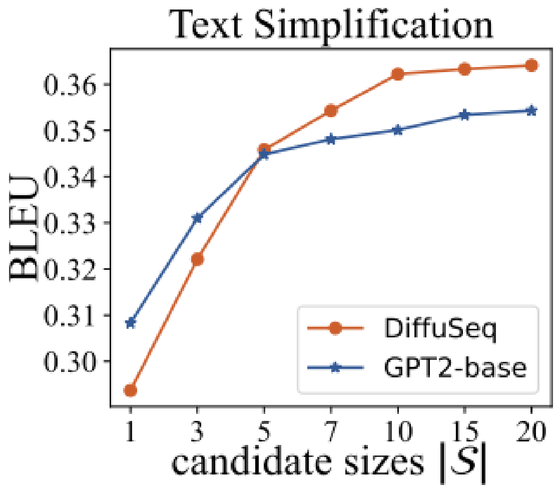
Three groups of baselines: Plain encoder-decoder, PLMs, NAR

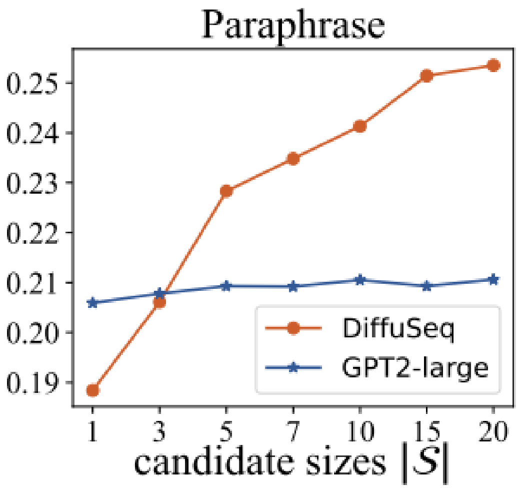
Tasks	Methods	BLEU↑	R-L↑	Score↑	dist-1↑	selfB↓ / div-4↑	Len
Paraphrase	GRU-attention ◇	0.1894	0.5129	0.7763	0.9423	0.9958/0.3287	8.30
	Transformer-base ◇	**0.2722**	0.5748	0.8381	0.9748	0.4483/0.7345	11.2
	GPT2-base FT •	0.1980	0.5212	0.8246	0.9798	0.5480/0.6245	9.67
	GPT2-large FT •	0.2059	0.5415	0.8363	**0.9819**	0.7325/0.5020	9.53
	GPVAE-T5 •	0.2409	**0.5886**	**0.8466**	0.9688	0.5604/0.6169	9.60
	NAR-LevT ‡	0.2268	0.5795	0.8344	0.9790	0.9995/0.3329	8.85
	DIFFUSEQ (Ours) ‡	0.2413	0.5880	0.8365	0.9807	**0.2732**/**0.8641**	11.2

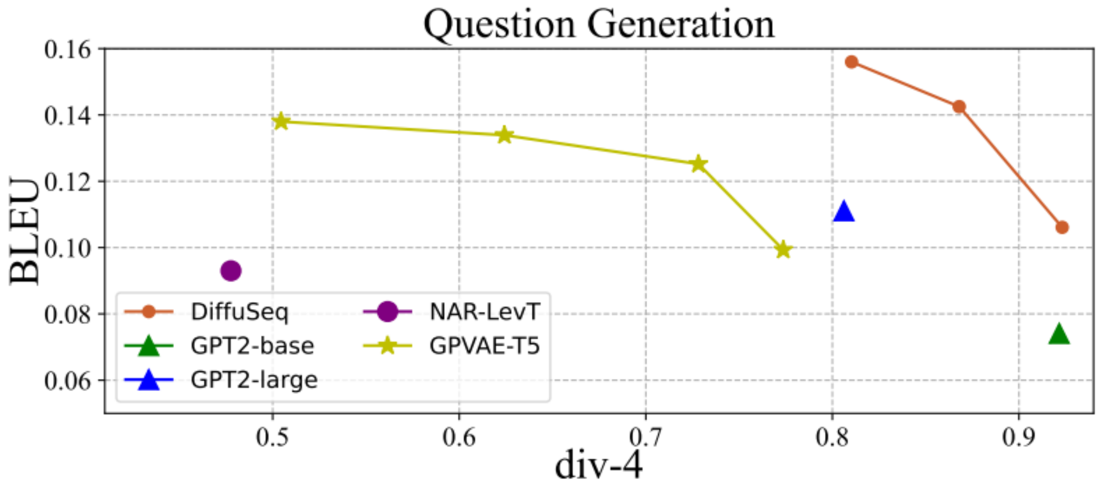
Comparable quality, better diversity

3.2 Analysis

Diversity Ensures Quality

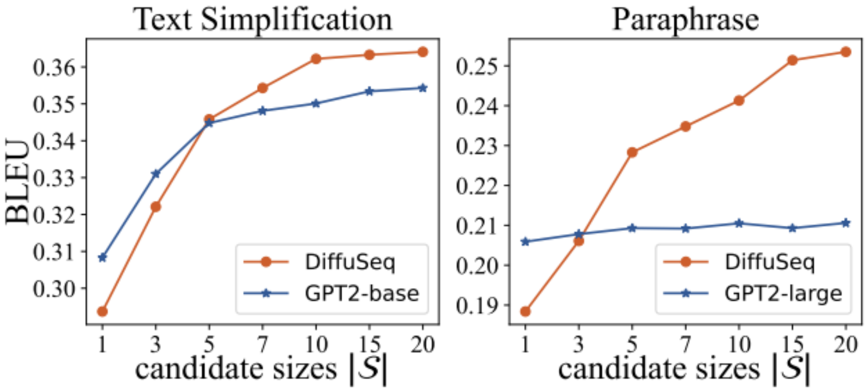


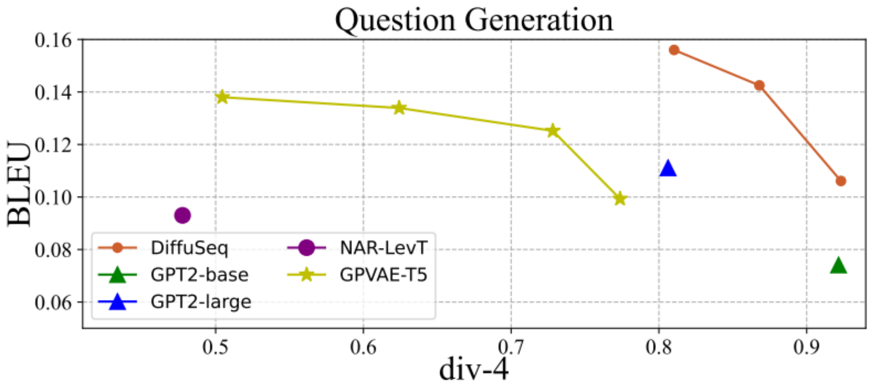




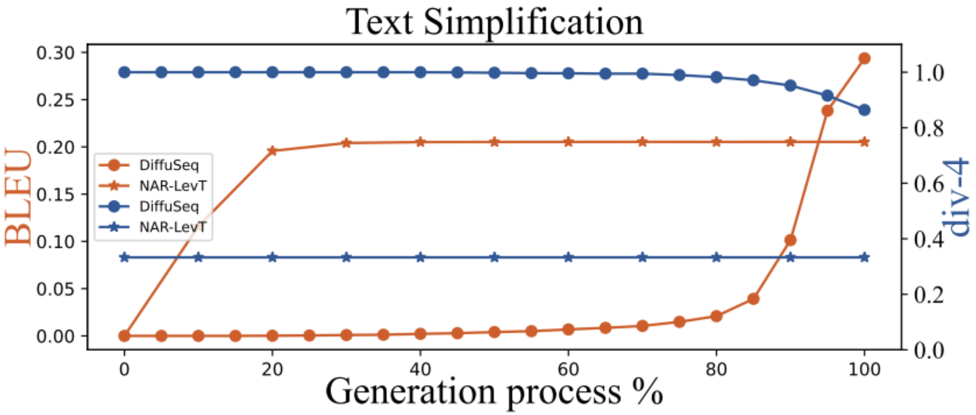
3.2 Analysis

Diversity Ensures Quality



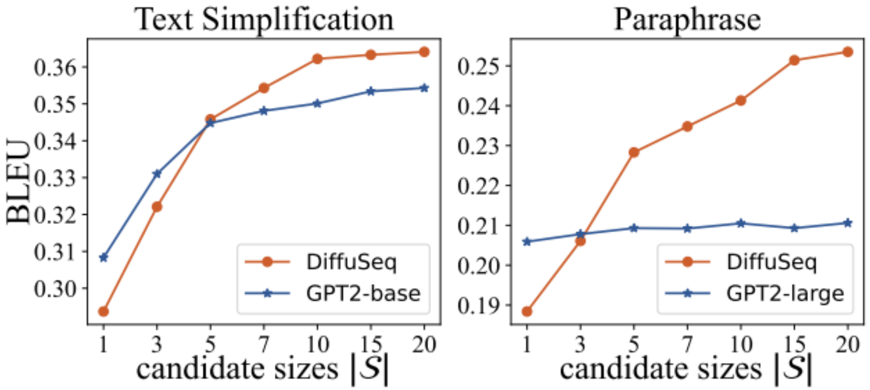


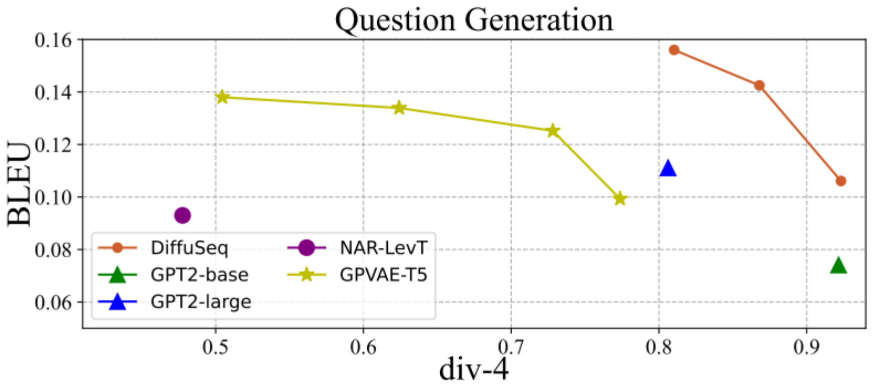
Step-wise Analysis against Iterative NAR



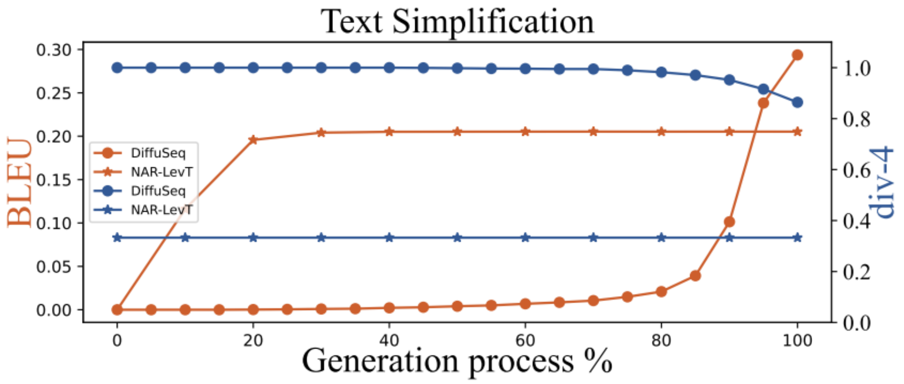
3.2 Analysis

Diversity Ensures Quality

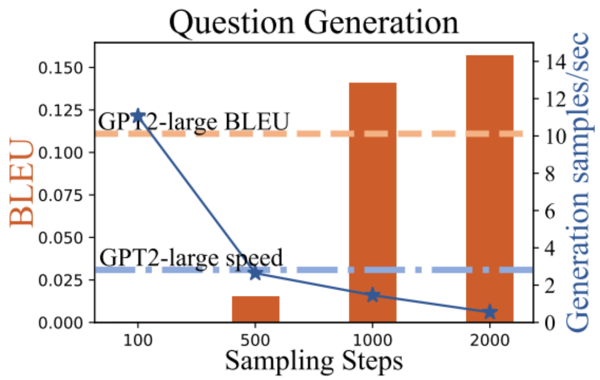




Step-wise Analysis against Iterative NAR



Inference Speed



4 Conclusion and future work

- DiffuSeq: as a new generation paradigm
  - Potential: competitive results on Seq2Seq tasks
  - Analysis: diversity DiffuSeq v.s. iter-NAR
- Future work: inference speed and sentence fluency

Thank you for watching!

Shark-NLP
Shanghai, China
hisansas@gmail.com

https://github.com/Shark-NLP/DiffuSeq