北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

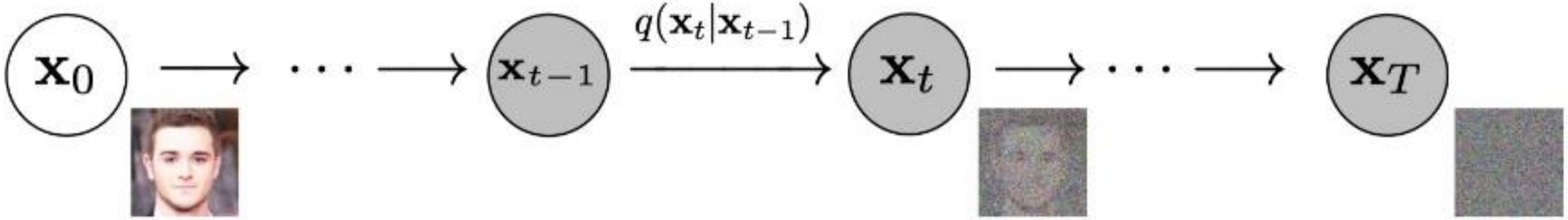
LREC-COLING 2024

Effective Integration of Text Diffusion and Pre-Trained Language Models with Linguistic Easy-First Schedule

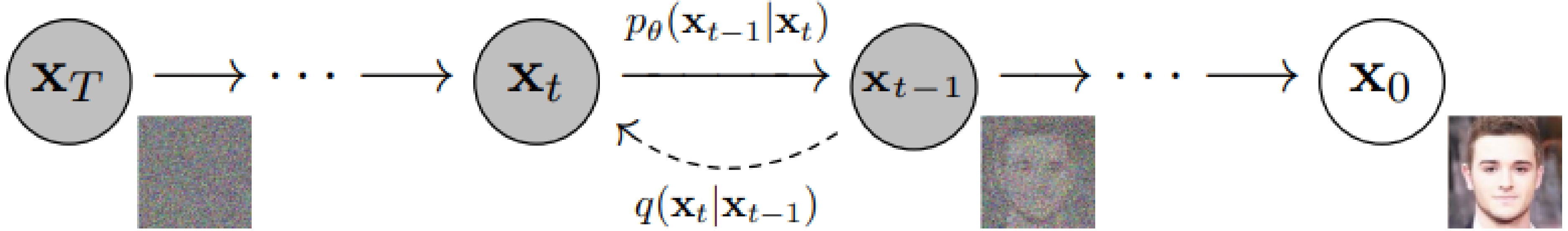
Yimin Ou, Ping Jian

Diffusion Models

➢ Forward Process



➢ Reverse Process



Diffusion Models for Text Generation

- **Discrete diffusion** (Hoogeboom et al., 2021; Austin et al., 2021)

Hoogeboom et al.(2021) respectively explores the diffusion process for discrete states with categorical transition kernels, uniform transition kernels, and absorbing kernels. However, replacing continuous diffusion with a discrete corruption process affords some flexibility (Dieleman et al., 2022).

- **Continuous diffusion** (Li et al.,2022; Gong et al., 2023)

DiffusionLM (Li et al., 2022) applies standard diffusion operations on the word embedding space and uses the rounding technique to map continuous space to discrete space during the reverse process.

Motivation

Limitations of existing text diffusion methods:

- The discrete nature of text data results in compatibility issues between continuous diffusion models (CDMs) and pre-trained language models (PLMs). That is, the performance of diffusion models even degrades when combined with PLMs.

- Existing noise schedules in text diffusion models do not consider the linguistic differences among tokens in a sequence, which violates the easy-first policy for text generation, causing the inaccurate generation of keywords and rare words.

Motivation

Our solutions:

- Utilize a pre-trained decoder to convert the denoised embedding vectors into natural language instead of using the widely used rounding operation. In this way, CDMs can be more effectively combined with PLMs.

- Linguistic easy-first schedule that that defines the importance of words in a sentence in terms of word relevance and information contents, conforming to easy-first-generation linguistic features and bringing about improved generation quality.

Method

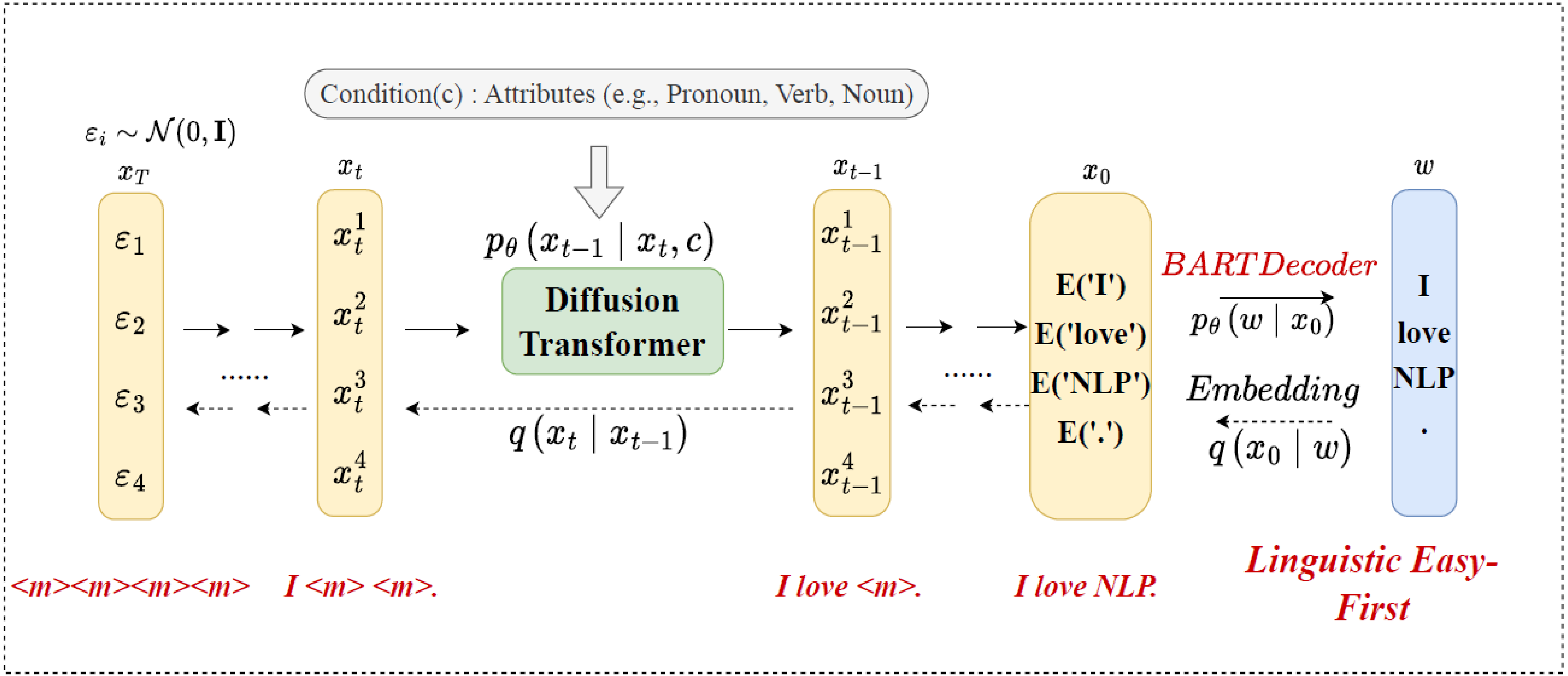


Figure 1: The overview of Diffusion-LEF.

Method

Linguistic Easy-First Schedule

- Word Relevancy:
  TextRank (Mihalcea and Tarau, 2004)

$$weight\left(v_i, v_j\right) = \frac{1}{\left|Out_{(v_j)}\right|}, \qquad (8)$$

where $Out_{(v_j)}$ represents the set of outdegree of node v_j. The score of node v_i is defined as follows:

$$score\left(v_i\right) = (1 - d) + d \sum_{v_j \in In(v_i)} \frac{weight(v_j, v_i)}{\sum_{v_k \in Out(v_j)} weight(v_j, v_k)} score\left(v_j\right), \qquad (9)$$

where $In_{(v_i)}$ represents the set of indegree of node v_i, d represents the damping coefficient, typically set to 0.85.

- The amount of Information:
  Entropy (Bentz and Alikaniotis, 2016; He et al., 2022)

$$H(w) = -p(w)\log(p(w)), \qquad (10)$$

$$p(w) = \frac{f_w}{\sum_{j=1}^{V} f_j}, \qquad (11)$$

where $p(w)$ represents the probability of the word w and f is the word frequency in the corpus.

Method

Linguistic Easy-First Schedule

➢ **The importance of word**

$$I(w) = \frac{score(w)}{\sum_{w' \in d} score(w')} + \frac{H(w)}{\sum_{w' \in d} H(w')}, \quad (12)$$

Experiment / Analysis

- E2E datasets
- Four classifier guided control tasks:

  Semantic Content, Parts of-Speech, Syntax Tree, Syntax Spans
- One classifier-free control task: Length

Methods	Semantic Content		Part-of-Speech		Syntax Tree		Syntax Spans		Length	
	Accuracy↑	Fluency↓	Accuracy↑	Fluency↓	Accuracy↑	Fluency↓	Accuracy↑	Fluency↓	Accuracy↑	Fluency↓
PPLM	9.9	5.32	-	-	-	-	-	-	-	-
FUDGE	69.9	2.83	27.0	7.96	17.9	3.39	54.2	4.03	46.9	3.11
DiffusionLM	81.2	2.55	90.0	5.16	86.0	3.71	93.8	2.53	99.9	2.16
DiffusionLM+BERT	77.4	2.68	86.2	5.43	82.3	3.92	89.3	3.13	99.9	2.68
Diffusion-LEF	81.7	2.46	91.2	5.09	86.3	3.68	94.4	2.48	99.9	2.14
Diffusion-LEF+BERT	**82.4**	**2.32**	**92.4**	**4.82**	**89.4**	**3.48**	**95.5**	**2.36**	**100**	**2.10**

Table 1: Main results on five controllable generation tasks.

Experiment / Analysis

➢ Main Result

Methods	Semantic Content		Part-of-Speech		Syntax Tree		Syntax Spans		Length	
	Accuracy↑	Fluency↓	Accuracy↑	Fluency↓	Accuracy↑	Fluency↓	Accuracy↑	Fluency↓	Accuracy↑	Fluency↓
PPLM	9.9	5.32	-	-	-	-	-	-	-	-
FUDGE	69.9	2.83	27.0	7.96	17.9	3.39	54.2	4.03	46.9	3.11
DiffusionLM	81.2	2.55	90.0	5.16	86.0	3.71	93.8	2.53	99.9	2.16
DiffusionLM+BERT	77.4	2.68	86.2	5.43	82.3	3.92	89.3	3.13	99.9	2.68
Diffusion-LEF	81.7	2.46	91.2	5.09	86.3	3.68	94.4	2.48	99.9	2.14
Diffusion-LEF+BERT	**82.4**	**2.32**	**92.4**	**4.82**	**89.4**	**3.48**	**95.5**	**2.36**	**100**	**2.10**

Table 1: Main results on five controllable generation tasks.

- Diffusion-LEF achieves the highest Fluency and Accuracy scores
- Excellent text generation quality and fine-grained control ability
- Efficiently integrate with the PLMs and thus make full use of the merits of both.

Experiment / Analysis

➢ Human Evaluation

Methods	Semantic Content	Part-of-speech	Syntax Tree	Syntax Spans	Length
DiffusionLM	3.56	3.63	3.61	3.42	3.81
DiffusionLM+BERT	2.81	3.10	2.96	3.04	3.20
Diffusion-LEF	3.89	4.05	4.12	3.72	3.98
Diffusion-LEF+BERT	**4.32**	**4.54**	**4.61**	**4.21**	**4.14**

Table 2: Human evaluation scores of different methods on five controllable generation tasks.

Conclusion

We propose a text Diffusion model Diffusion-LEF

- effectively combined with pre-trained language model BERT, thus taking advantage of diffusion model and pre-trained language model.

- the linguistic easy-first schedule that incorporates the measure of word importance, conforming to easyfirst generation linguistic features and bringing about improved generation quality.

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

LREC-COLING 2024

Thank you for your listening!