TSU HITS's Submissions to the WMT 2024 General Machine Translation Shared Task

Vladimir Mynka♠, Nikolay Mikhaylovskiy♠◊
♠Higher IT School, Tomsk State University, Tomsk, Russia
◊NTR Labs, Moscow, Russia
vladimirmynka34821@gmail.com, nickm@ntrlab.com

Abstract

This paper describes the TSU HITS team's submission system for the WMT'24 general translation task. We focused on exploring the capabilities of discrete diffusion models for the English-to-{Russian, German, Czech, Spanish} translation tasks in the constrained track. Our submission system consists of a set of discrete diffusion models for each language pair. The main advance is using a separate length regression model to determine the length of the output sequence more precisely.

1 Introduction

This report gives an overview of TSU HITS submissions in the WMT 2024 general machine translation tasks. We focused on exploring the capabilities of discrete diffusion models for the English-to-{Russian, German, Czech, Spanish} translation tasks in the constrained track. Our main contributions are

1. the use of regression-based output length prediction model
2. the use of the input length as a key feature for the output length prediction

The report is organized as follows. In the Section 2, we provide a general description of the discrete diffusion approach to machine translation, as it is not yet very widespread. In the Section 3, we describe the experimental setting and training processes. Section 4 discusses the results.

2 Discrete Diffusion Approach to Machine Translation

2.1 Diffusion: Preliminaries

Diffusion approaches (Sohl-Dickstein et al., 2015 , Ho et al, 2020) to generating objects (for example images) include forward (data to noise) and reverse (noise to data) diffusion processes. In the forward process, a small amount of noise is gradually added to the data. In the classical direct diffusion process, the original object x_0 is repeatedly and additively perturbed by a small Gaussian random noise, and in a fixed number of steps T goes into state x_T with a normal distribution (and thus is converted to noise):

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t;\ x_{t-1}\sqrt{1-\beta_t}, \beta_t\right), \quad (1)$$

where $\forall\, t = \overline{1..T}\ \beta_t \in (0;1]$ are the hyperparameters that regulate the diffusion rate.

During the reverse diffusion process, the machine learning model step by step reconstructs the object's states from x_T to x_0, and this denoising restores an object from the original distribution:

$$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}\left(x_{t-1};\ \mu_\theta(x_t,t), \sigma_\theta(x_t,t)\right), \quad (2)$$

where θ are the model's trainable weights.

Texts in typical representations do not have the property of continuity and are a sequence of tokens with discrete values that do not have an order relation and correspond to the categorical data type. Thus, we follow the path of adapting the diffusion processes to categorical data - such approaches are called discrete diffusion.

2.2 Discrete Diffusion for Text Generation

Diffusion models with discrete state spaces were first introduced by Sohl-Dickstein et al. (2015), who considered a diffusion process over binary

Proceedings of the Ninth Conference on Machine Translation, pages 205–209
November 15-16, 2024 ©2024 Association for Computational Linguistics

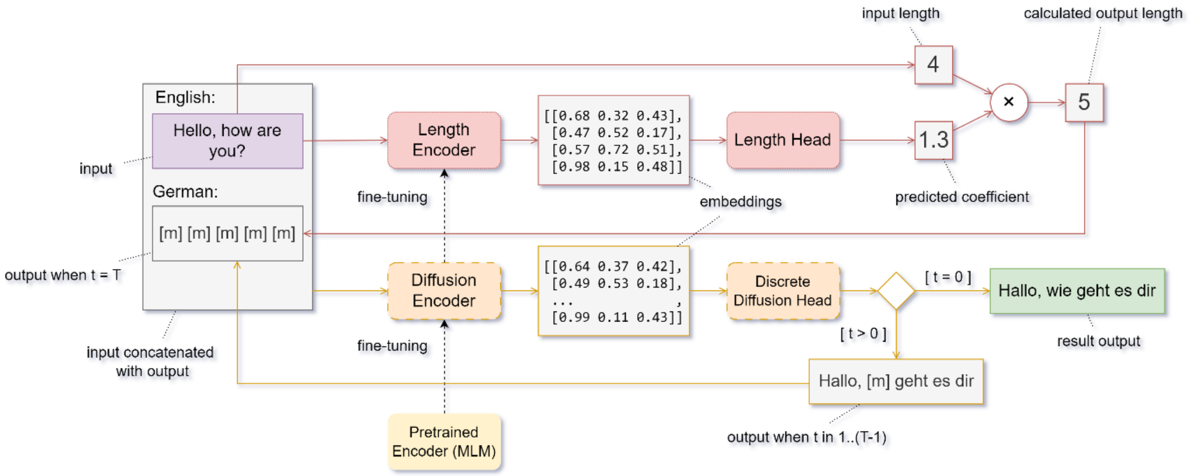

Figure 1: Overview of the system

random variables. Hoogeboom et al. (2021) extended the model class to categorical random variables with transition matrices characterized by uniform transition probabilities. We follow Austin et al. (2021) to define a discrete diffusion model for texts.

Namely, we consider each text token x_t to be a discrete random variable with K categories. For text data, $K = |V|$ is the size of the vocabulary. (He et al., 2023). The forward transition probabilities can be represented by matrices: $[Q_t]_{ij} = q(x_t = j|x_{t-1} = i)$. The process of adding noise can then be written as

$$q(x_t|x_{t-1}) = Cat(x_t;\ p = x_{t-1}Q_t) \qquad (3)$$

where $Cat(\cdot)$ is a category distribution (Austin et al., 2021).

2.3 Masked Language Models and Discrete Diffusion

He et al. (2023) noted the relationship between the discrete diffusion process and the task of pretraining of masked language modeling (MLM) encoder models. Namely, they suggested incorporating an absorbing state, e.g., [MASK] for BERT, into the Markov process of diffusion. In particular, each token in the sequence either stays the same or transitions to [MASK] with some probability. Formally, each entry of the transition matrix at step t is as follows,

$$[Q_t]_{ij} = \begin{cases} 1 & if\ i = j = [\mathrm{M}] \\ \beta_t & if\ j = [\mathrm{M}], i \neq [\mathrm{M}] \\ 1 - \beta_t & \text{if i } = \text{ j } \neq [\mathrm{M}] \end{cases} \qquad (4)$$

where [M] is short for [MASK].

Such a Markov process converges to a stationary distribution $q(x_T)$ that places all the probability mass on the sequence with all [MASK] tokens.

The most common transformer (Vasvani et al., 2017) models pre-trained for the MLM task are models from the BERT family (Devlin et al, 2019). He et al. (2023) suggested DiffusionBERT that uses a pretrained BERT model as an encoder due to the similarity of the tasks. The length of the output sequence of the DiffusionBERT model is fixed and is set to different values depending on the problem solved.

2.4 Discrete Diffusion for Translation

Reid et al. (2023) suggested a diffusion model using Levenstein operations for machine translation. They have tested the model on WMT14 EN-DE dataset. It is unclear from the paper how do the authors determine the target length of the output sequence.

Zheng et al. (2023) suggest a reparameterized discrete diffusion (RDM) approach to text generation, and report results for the machine translation task on the IWSLT14 DE-EN, WMT14 EN-DE and WMT16 EN-RO datasets. To determine the translation length, the authors of RDM trained a separate model similar to the one of Ghazvininejad et al. (2020). They pose the problem of determining the length of the output sequence as a classification problem, selecting k best options out of N possible, where N is the maximum text length that the model used can process. Similarly to Gao et al. (2024), several options are selected and the best one is chosen based on the metrics of the overall text quality.

Ye et al. (2023) explore the possibilities of increasing applicability domain of discrete diffusion approaches, while considering an approach similar to DiffusionBERT, except that instead of the BERT encoder, the authors use the RoBERTa model (Liu et al., 2019). The quality of machine translation is assessed on the IWSLT14 DE-EN and WMT14 EN-DE data sets, using the same quality metrics and the same idea for determining the length as in the RDM approach.

3 System Overview

3.1 General Translation Process

The general translation process is presented on Figure 1. Our system consists of a discrete diffusion model and an output length prediction model.

On each diffusion step, a concatenation of the source text and output is used as the input to the generative model, but the absorbing tokens are distributed only within the output part. We do not use any special separation tokens, but just use the prompt "{Source Language}: {Source Text} \n {Target Language}: {x_t}".

Since we use XLM-RoBERTa's (Conneau et al, 2020) positional embedding model as an encoder and are forced to fit the input sequence of the model into 512 tokens, we apply punctuation splitting of the source texts, limiting the maximum size of the source text to 200 tokens, and then glue the results back. We also do not use the extended context to improve translation; this is left for the future work.

We take a fixed number of the diffusion steps T equal to 50. Tokens that were unmasked in the previous steps are likely to be replaced with subsequent ones, just like in DiffusionBERT (He et al., 2023). The standard argmax approach is used as a sampling method. We do not use temperature and do not limit the number of tokens to choose from.

3.2 Generative Model

We largely follow Ye et al. (2023) and use XLM-RoBERTa (Conneau et al, 2020) family pre-trained model that includes a multilayer transformer encoder and a single-layer MLM head.

We fine-tune both the encoder and the head for discrete diffusion text generation that differs from MLM mainly by the percentage of the masked tokens. We use the cross-entropy weighted relative to the diffusion step t loss proposed by Zheng et al. (2023):

$$L_t = -\lambda_{t-1} \sum_{i}^{N} y_i \log p_i \tag{5}$$

where y_i is the true probability (0 or 1) of token with index i in model dictionary, p_i is the predicted probability, N is the size of the dictionary, λ_{t-1} is the parameter that depends on the percentage of the masked tokens at the steps t and $t-1$.

Generative Model	
Architecture	XLM-RoBERTa-Large
Optimizer	AdamW($\beta_1 = 0.9, \beta_2 = 0.98$)
Weights decay	0.01
Learning Rate Schedule	Cosine
Max learning rate	5E-05
Batch size	16
Accumulation step	8
Steps	30000
Warmup ratio	0.01
Loss	(Section 3.2)
Number format	FP16
Length Model	
Hidden size	1024
Optimizer	AdamW($\beta_1 = 0.9, \beta_2 = 0.999$)
Learning Rate Schedule	OneCycleLR (Smith et al, 2017), two phases
Max learning rate	7E-07
Batch size	8 / 16
Steps	30000
Embedding calculation	Mean pooling
Activation	ELU
Loss function	MSE
Number format	FP16

Table 1: Hyperparameters of the models

Following Chang et al. (2022), we use the cosine noise schedule:

$$\beta_t = \cos(\frac{\pi t}{2T}) \tag{6}$$

3.3 Length Predictor

Our length predictor also consists of an encoder and a task-specific head. Although our length prediction model is based on the same XLM-RoBERTa, physically these two models are completely separate. We tried not to fine-tune the encoder for the length problem and to use the standard XLM-RoBERTa, but we got worse metrics on the test data.

We use a regression predictor of the output length, unlike other works that use classifiers with the number of categories equal to the length of the context, for example, 512 tokens. Our regression head is a two-layer perceptron with ELU-activation. Standard MSE loss is used when the length predictor is trained.

	#tokens	#model parameters
EN-DE	68,333	444,158,849
EN-RU	91,932	492,511,151
EN-ES	31,380	368,444,201
EN-CS	65,514	438,382,718

Table 2: Numbers of tokens and model parameters after pruning the tokenizer

The main improvement in length prediction is because of the use of the input length. There is a fairly strong relationship between the length of the text in the source language and the length of its translation, which, in general, is almost linear. We suggest taking this into account when the target is defined. Our model predicts the ratio of the input and output lengths, normalized by the average ratio for the training set. We employ standard mean pooling to convert the matrix of token embeddings obtained from the encoder into a common embedding of text, which will be used as features for the length head.

3.4 Training Data

The WikiMatrix dataset (Schwenk et al., 2021) was used as a train dataset for EN-DE, EN-RU, EN-CS language pairs; Neulab-TedTalks (Tiedemann, 2012) was used for EN-ES. The training sets were trimmed to 480 thousand examples when training the generation model and to 240 thousand when training a length prediction model.

3.5 Pruning the tokenizer

Due to the computational limitations we reduce the token set of our models for each pair of languages to the minimum required (all the other tokens are replaced with [UNK]). The effect of reduction on the number of model parameters is demonstrated in Table 2. According to our observations, it increases the quality of models when tested on validation datasets for the selected language pair, but may degrade the quality of general translation when tested on complex examples.

Pruning the tokenizer was made before trimming the training sets to keep as much tokens as possible.

4 Results

The official automatic scores of our system on the test data are presented in the Table 3. The gap between our results and the leading system is significant.

	AutoRank	MetricX	CometKiwi
EN-DE	13.3	5.6	0.395
EN-RU	10.8	9.8	0.421
EN-ES	16.3	14.2	0.289
EN-CS	19.5	16.6	0.235

Table 3: System official scores

4.1 Model size

We used XLM-Roberta-Large with 561 million parameters as the main model for generating translation, while other systems participating in the competition this and last years had tens of billions of parameters. This makes our model largely uncompetitive. Unfortunately, today there are no pretrained open-weight encoder models comparable to leading open-weight decoder models in terms of parameters number and pretrain token count.

4.2 Quantity and quality of training data

Due to technical limitations, we used only a small part of the translation datasets provided, no more than 480 thousand examples for each language pair. Increasing the training set and better cleaning should significantly improve the quality, especially when using a larger pretrained model.

Acknowledgments

We are grateful to Prof. O. Zmeyev for the support of this research and Dr. V. Goiko for the computing resources provided.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual,* pages 17981–17993.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. 2024. Empowering Diffusion Models on the Embedding Space for Text Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4664–4683, Mexico City, Mexico. Association for Computational Linguistics.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2023. DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4521–4534, Toronto, Canada. Association for Computational Linguistics.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS 2021, December 6-14, 2021, virtual,* pages 12454-12465.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint 1907.11692

Machel Reid, Vincent J. Hellendoorn, and Graham Neubig. 2023. Diffuser: Discrete diffusion via edit-based reconstruction. In *Proceedings of The Eleventh International Conference on Learning Representations,* https://openreview.net/forum?id=nG9RF9z1yy3

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume,* pages 1351–1361, Online. Association for Computational Linguistics.

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

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems, NeurIPS 2017,* pages 5998–6008

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu and William T. Freeman. “MaskGIT: Masked Generative Image Transformer.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022): 11305-11315.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Quanquan Gu. 2023. Diffusion language models can perform many tasks with scaling and instruction-finetuning. arXiv preprint arXiv:2308.12219.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. 2023. A reparameterized discrete diffusion model for text generation. arXiv preprint arXiv:2302.05737.

Leslie N. Smith and Nicholay Topin. 2017. Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates. arXiv preprint arXiv:1708.07120.