A Study on English Translation Syntactic Reconstruction and Diversity Generation Driven by Diffusion Models

Lili Wang[1,*] and Na Zhao[2]
[1]School of Foreign Languages, Liaodong University, Dandong, 118001, Liaoning, China
[2]School of Chemical Engineering and Machinery, Liaodong University, Dandong, 118001, Liaoning, China
Corresponding author: Lili Wang (e-mail: lovewanglili713@163.com).

ABSTRACT The lack of syntactic diversity and expressive variety in English translation hinders the improvement of translation quality. In response to this, a syntactic reconstruction and diversity generation method of English translation based on diffusion models is proposed in this paper. It consists of two parts: Firstly, it builds a syntax structure encoder from the source language sentences to embed them in a latent semantic space; Secondly, it designs a forward diffusion process to add Gaussian noise to gradually destroy the syntactic features of the source language; and then, it constructs a reverse denoising network in UNet architecture and self-attention mechanism and gradually recovers the target language syntax structure from the pure noise, and finally, it introduces conditional guidance mechanism into the reverse denoising network, which integrates semantic constraints and syntactic templates to control the style and structure of the generated sentences. Experiments on the WMT translation dataset show that the syntactic diversity index of the generated translation reaches 0.87, an improvement of 31.8 % compared to the Transformer baseline model; the BLEU score is 41.3, and the semantic fidelity score is 4.2. This method effectively solves the problem of sentence rigidity in translation systems and provides a new technical path for neural machine translation.

INDEX TERMS KEYWODRS:LightGBM; Economic efficiency; Parameter optimization; SHAP value; Ensemble learning

I. INTRODUCTION

The ability of neural machine translation systems for cross-linguistic information transmission is quite high in terms of semantic understanding and, in terms of syntactic structure, very homogeneous translation characteristics are generated. In the encoder-decoder model with auto-regressive mechanisms, the model is trained to choose high-frequency sentence patterns, leading to very similar syntactic outputs for different translations of the same sentence. This uniformity also reduces the expressiveness and readability of the translation besides narrows the application value of the translation systems to situations that are clearly required of stylistic differences, such as literary works and creative texts [1-2].

The traditional methods add up the variability through decoding strategy manipulations or random sampling, but the variability mostly occurs at the lexical selection level and it is hard to achieve deep syntactic reorganization. There are many syntactic differences between source language and the target language, which make it difficult to make direct structural mapping, and there are no existing models that have a systematic exploration mechanism for the syntactic space.

The powerful and diverse generation ability of diffusion models of image generation is leading to a solution to the rigid translation syntax in which the model slowly infers noise to the datagrams and then deverts the structured output out of random regions into the opposite direction by reversing denoising. This generation allows for the adoption of high dimensional space, of which there are many possibilities. When syntactic structure is used for translation, we can look at syntactic form as a distribution in continuous latent space and deconstruct and recreate the sentence by adding and removing noise to syntactic content [3-4]. With conditional guidance we could apply semantic constraints to generation and make translation syntactically new and meaningful for the original text. This is a new solution to one-to-one constraint of sequence generation since it provides translation systems with “degrees of freedom” in the syntactic dimension that allow to produce a different high quality translation in different application conditions or style.

In this paper, a complete process of English translation syntactic reconstruction framework based on diffusion model is constructed, which realizes the process of syntactic structure encoding and conditional generation.

This designed syntactic encoder and the graph convolutional network transform the syntactic tree structure into continuous vector representation by extracting dependency relation features. The idea of a hierarchical noise strategy is to apply differential perturbations to the syntactic and semantic information, so that during diffusion the structural information is degraded gradually. The noise residual can be accurately predicted at each time-step by the improved UNet denoising network, which fuses multiple scales features and introduces a time-aware mechanism. The proposed method is a classifier-guided one which combines syntactic template constraints and semantic fidelity requirements for the sampling process to obtain controllable diversity generation. Experiments confirm that this approach is able to get a good balance between syntactic diversity and translation quality, paving a new road for enhancing the expressive ability of neural machine translation [5-6].

II. RELATED WORD

The era of statistics has been over and NMT has moved into the deep learning age. Modern translation systems have been based on sequence-to-sequence models, with the help of attention mechanisms. The source language features are extracted by the encoders, while target language sequences are generated autoregressively by the decoders. The new released Pre-trained Language Models has further improved the quality of translation. Large-scale parallel corpora can be used to learn cross-lingual representations with multilingual models, and they have demonstrated a good generalization capability on low resource language pairs [7-8]. These methods, however, basically use maximum likelihood estimation to seek to minimize the objective, and the training tends to encode the "peak region" of the data-distribution, leading to a clustering of the generated results at high frequency sentence patterns.

Researchers have tried several different strategies for decoding such as sampling temperature adjustment, kernel sampling, and variants of the bundle search, but all the above methods change only the choice of words in the decoding process, without changing the syntactic structure. In order to alleviate the pattern collapse issue to some degree, reinforcement learning based approaches can be applied to guide the model along non-optimal paths by utilizing reward functions; however, the requirement to design the reward function and the training instability of the RL-based approach limit their applicability [9-10].

In syntax-aware translation models, structural alignment is provided by explicitly modelling the syntactic structure in the source language and the target language (either by transforming the syntactic tree of the source language into the syntactic tree of the target language, or by generating the syntactic sequence of the target language from the syntactic tree of the source language). The methods are dependent on the correctness of the syntactic parser and have difficulties coping with cross-linguistic differences and syntactic ambiguity.

Many successful computer vision generative models motivated a study on their potential use for natural language processing. For the text generation, variational autoencoders (VAEs) learn the distribution of the latent space and posterior collapse problem makes the decoder forget about the latent variables and rely on the input. In text, generative adversarial networks (GANs) suffer from the problem of passing discrete symbolic gradients over the network and requires either policy gradient or reinforcement learning training. The iterative denoising task in diffusion models leads to more stable and diverse samples than in GAN for image synthesis. The model simulates a process of distorting data from noise to data and forward will gradually destroy the data structure and backward will learn to recover the data. The content produced using conditional generation is controlled either by classifer guided or classifier free and the generation quality and diversity trade-off is lower than GANs. Diffusion models can be used for natural languages generation but so far they lack discrete symbolic spaces due to continuous embedding spaces in previous work. For a translation problem, the task of exploring the syntactic space while maintaining semantic constraints can be treated as a syntactic diffusion problem that can be solved by a suitable technical tool: the progressive generation of diffusion models.

III. METHODS

A. SYNTACTIC STRUCTURE ENCODER DESIGN

The syntactic structure encoder employs a two-layer encoding architecture to process source language sentences. The bottom-layer encoder extracts lexical-level semantic features based on a Transformer structure, mapping the input sequence $X = \{x_1, x_2, \ldots, x_n\}$ to a hidden state sequence $H = \{h_1, h_2, \ldots, h_n\}$. The top-layer encoder incorporates the results of dependency syntax tree parsing, constructing a syntactic adjacency matrix $A \in \mathbb{R}^{n\times n}$, and aggregating syntactic relations between nodes through a graph convolutional network. The encoding process is represented as follows:

$$H_{\text{syn}} = \text{GCN}(H, A) = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}HW\right) \quad (1)$$

where $\tilde{A} = A + I$ is the adjacency matrix with self-loops added, $\tilde{D}$ is the degree matrix, and W is the learnable parameter. Syntactic and semantic features are fused through a gating mechanism to generate a unified sentence representation vector z:

$$z = \sigma(W_g \cdot [H_{\text{syn}}; H]) \odot H_{\text{syn}} + (1 - \sigma(W_g \cdot [H_{\text{syn}}; H])) \odot H \quad (2)$$

This representation vector contains both lexical semantic information and syntactic structural constraints, serving as the initial input to the diffusion process. The 256-dimensional vector space output by the encoder can distinguish different syntactic patterns, providing structured guiding signals for subsequent conditional generation.

B. FORWARD DIFFUSION PROCESS AND NOISE ADDITION MECHANISM

The forward diffusion process corrupts the syntactic encoding vector to create a typical Gaussian noise. The process runs over 1000 time steps and noise components are added at each time step, based on a pre-defined table of noise scheduling. Cosine Annealing Schedule on the intensities of the noises; low intensities at the beginning to preserve syntactic structure information, high intensities in the intermediate phases to diminish the semantic features and saturation level intensities in the final phases. In particular, the noise scaling factor is linearly ramped from 0.0001 to 0.02 for the time step t, so that the syntactic features are gradually lost.

Diffusion uses a non-uniform noise at the different dimensions of the syntactic representation vector. The 64 dimensions representing syntactic dependencies are weaker during diffusion so that they are retained for longer. Standard-intensity noise is fed to the 192 dimensions related to semantic content to speed up semantic decoupling. A hierarchical noise model allows the model to recover the syntax of the code and then the meaning in the backwards generation. The noise residuals are sampled at each diffusion step and all diffusion trajectory data are taken as training data for the reverse denoising network. The KL divergence of the pure noise vector at the diffusion endpoint and the standard normal distribution are not smaller than 0.03, which is sufficient for the samples to be random.

C. INVERSE DENOISING NETWORK ARCHITECTURE

The inverse denoising network is an improved UNet to reconstruct the noise residual at current time step. It is a 4 layer downsampling network followed by 4 layer upsampling network with skip connections in between. The downsampling part has convolutional layers to shrink the input (256 dimensional) into a 64 dimensional bottleneck representation with global dependency on the syntactic structure. The upsampling part slowly increase the number of features while transpose the convolutions and have the syntactic information in each level of abstraction combined with skip connections.

The positional embedding of the temporal step is sinusoidal one fed to each layer of the network. The weights and bias parameters of the convolutional layers are modulated by an affine transformation of the temporal step index and the temporal step index is mapped to a 128 dimensional vector. To have the network adapt to the diffusion process dynamically, this time-aware mechanism is added. The initial time steps are used to reconstruct the syntactic structure and the rest to refine the word selection.

The attention module is inserted after every upsampling layer, and after the bottleneck layer to correlate the dimensionality of the syntactic representation vector to strengthen the inner relation of the syntactic components of subject, verb and object. Multi-head architecture consists of eight heads with each head having dimensions of 32, and can extract multiple syntactic features simultaneously, such as verb tense, noun number and case, clause structure etc. Multi-head attention is a special case of multi-head architecture. The denoised prediction output from the network at the current time step is used directly to correct noise, helping to prevent the syntactic collapse due to the accumulation of errors.

D. CONDITION-GUIDED SYNTACTIC GENERATION STRATEGIES

The conditional guidance method is one of several classifiers-directed approaches to guiding the direction of syntax generation. An independent syntax classifier is trained to identify target syntax sample and non-target syntax sample. Denoising net output a representation of text and then input to classifier. The classifier is a 3-layer fully connected network that outputs a probability distribution over 12 types of syntax. In the backsampling, we add the classifier gradient and the denoising prediction gradient and then fuse the two signals to guide the direction of the generation trajectory to the indicated syntax structure. The corrected denoising gradient is computed as follows:

$$\epsilon_\theta^{\text{guided}}(z_t, t) = \epsilon_\theta(z_t, t) - s \cdot \nabla_{z_t} \log p_\phi(c \mid z_t) \quad (3)$$

The strength parameter s for the guiding is chosen as 7.5 to strike a balance between the semantic accuracy and syntactic variety. We employ the classifier loss function with a cross-entropy loss and a contrastive loss, so that boundaries between different syntactic patterns are sharp in the feature space.

CLIP implements semantic constraints on modalities. Each encoder encodes a sentence of the source language into a 512 dimensional semantic vector and at each denoising step cosine similarity between current state h produced by the decoder and semantic vector is calculated. If similarity < 0.75, semantic correction is applied (denoising direction is reverted back to the semantics of the source content). Amplitude of the correction decays at every step of time according to the following formula:

$$\Delta z_t = \lambda \cdot \left(1 - \frac{t}{T}\right) \cdot \frac{z_{\text{src}} - z_t}{\|z_{\text{src}} - z_t\|} \quad (4)$$

The attenuation coefficient lambda is set to 0.3 to avoid excessive constraints that could impair syntactic innovation.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETUP AND DATASET DESCRIPTION

The WMT2019 English-Chinese parallel sentence pairs are used to build such a data set for this experiment, with a total of 1.8 million parallel sentences and the length of sentences ranging from 5 words to 50 words. It was chosen because it contains various kinds of texts (news, literature and spoken language) and because its texts exhibit syntactic variation (as it is appropriate to test the ability of the model to create a variety of texts). Data was then partitioned in

a 8:1:1 partitioning with training data, validation data and test data, respectively. The model was executed on PyTorch framework and trained with eight NVIDIA A100 GPUs and a batch size of 128. Cosine annealing was used to decrease the learning rate from 5e-4 to 1e-6. Weights decay rate was set to 0.01 and optimizer used was AdamW. The diffusion steps were set to 1000 and the DDIM acceleration algorithm was used for sampling to decrease the number of diffusion steps to 50. It is worth of note that the important hyper parameter settings for each module of the model are shown in Table 1.

TABLE 1. Butler parameter configuration

Hyperparameter	Encoder	Denoising Network	Classifier	Diffusion Process
Hidden Dimension	256	512	256	256
Attention Heads	8	8	4	-
Network Layers	6	8	3	-
Dropout Rate	0.1	0.15	0.2	-
Guidance Strength	-	-	7.5	-
Total Timesteps	-	-	-	1000
Noise Schedule Range	-	-	-	0.0001-0.02

B. SYNTACTIC DIVERSITY AND SEMANTIC FIDELITY ANALYSIS

To test, 5000 sentences randomly selected from the test set were translated into 10 different versions. The syntactic diversity was calculated by the mean edit distance between the syntactic trees of the generated translations, the larger the distance, the more diversity in the syntax. The degree of semantic fidelity was assessed by three professional translators who assessed the translated texts with a blind test based on 5-point scale to decide the extent of transfer of meaning from the source language. For each model, the average inference time and lexical richness of the generated text were also measured. The same beam search parameters were used in all models, beam width was 5 and beam length penalty coefficient was 0.6. The results are shown in Table 2.

TABLE 2. Results of syntactic diversity and semantic fidelity analysis

Model	Syntactic Diversity	Semantic Fidelity	BLEU Score	Inference Time (s)	Lexical Richness
Transformer	0.66	3.8	38.7	0.12	2847
mBART	0.71	4.0	40.1	0.28	3126
Ours (Small)	0.82	3.9	39.5	0.35	3542
Ours (Standard)	0.87	4.2	41.3	0.52	3891
Ours (Large)	0.89	4.3	42.1	0.78	4103

Syntactic diversity score of standard model is 0.87, 0.87 is 31.8% better than Transformer, and 22.5% better than mBART. Not only does the standard model explore different syntactic spaces, the diffusion model is also able to explore different syntactic spaces. Semantic fidelity score 4.2 is higher than the two baselines, that is, the conditional guidance mechanism limits the generation process and avoids semantic drift. BLEU score is slightly better than mBart, that means the translation quality is not limited by diversity. Inference time 0.52s is slightly higher than TransTransformer and mBT, but still acceptable, and DDIM sampling algorithm is efficient to reduce inference cost for the diffusion models. Very high lexical richness (3891) compared to the baselines (the lexical expression spaces triggered by diffusion process) more than baseline.

C. COMPARISON OF ABLATION EXPERIMENT AND MODEL PERFORMANCE

Five variants were designed for the experiment based on the above three changes: removing the graph convolutional layer from the syntactic encoder and only keeping the Transformer encoder, removing the conditional guidance mechanism and adopting unconditional diffusion generation, removing the self-attention module and replacing it with only convolutional layers to build the denoising network, and replacing the cosine noise scheduling with linear noise scheduling. All hyperparameter values were the same for all variants and the number of epochs was 20 throughout each of them. During the test phase, 3000 translated units were tested for each model and the following data were recorded: syntactic diversity index, semantic fidelity, BLEU, failure rate of models. The definition of generation failure was that a significant number of the output texts had many serious syntactic errors or were semantically completely different from the source text, according to manual annotation. Also, the effect of different values of guidance strength parameters (1.0 to 15.0 by increments of 2.5) was evaluated and the performances of the model were compared based on model controllability and creativity. The results are shown in Table 3.

TABLE 3. Ablation experiment results

Model Variant	Syntactic Diversity	Semantic Fidelity	BLEU Score	Generation Failure Rate (%)
Full Model	0.87	4.2	41.3	2.3
w/o Graph Convolution	0.79	4.1	40.5	3.8
w/o Conditional Guidance	0.92	3.4	35.2	12.6
w/o Stratified Noise	0.81	3.9	39.7	5.1
w/o Self-Attention	0.74	3.7	38.1	6.9
Linear Noise Schedule	0.83	4.0	40.2	4.2

The results of ablation experiment are shown in Table 3. removing the conditional guidance raise the syntactic diversity index to 0.92, meanwhile decrease the semantic fidelity to 3.4, and increase generation failure rate to 12.6% showing that the conditional guidance plays an important role in controlling the quality of generation. removing the graph convolution layer, the syntactic diversity drop to 0.79 which shows that the ability of encoding syntactic structure is vital to explore various sentence pattern. The lack of hierarchical noise strategy cause the drop in diversity to 0.81, which shows that differential noise addition more adequately keep the syntactic framework information. Removing the self-attention module leads to a drop in diversity index to 0.74, and an increase in failure rate to 6.9%, which shows that the model long distance dependency modeling for syntax integrity is very important. All the metrics were better under cosine noise scheduling than linear scheduling, and the decline curve of the former is more nonlinear which is better adapted to hierarchical syntactic features structure. The full model achieves the best performance on all the metrics, with the BLEU score of 41.3 and 2.3% failures showing stable generation.

V. CONCLUSION

The model maps syntactic structures to a continuous latent space, thereby bypassing the limitations of traditional methods of sequence generation, and thus provides a gradual recovery process from noisy to structured language. The graph convolution mechanism proposed by the syntactic encoder is able to well reflect dependency relations and structure, and effectively input rich structured prior knowledge for the subsequent generation. The hierarchical noise strategy maintains the syntactic framework information when dealing with diffusion process, without uncontrolled generation when the information is completely destroyed. The conditional guidance mechanism is used to control the generation direction using the gradients of classifiers, which releases syntactic innovation space, ensures semantic fidelity and solves the contradiction between diversity and controllability.

The current method still has inference speed problems, although the number of steps is reduced to 50 in DDIM sampling, the performance is still not as good as the traditional autoregressive model. There is limited capacity for the model to process more complex sentences; when the source sentence is more than 60 words, syntactic accuracy drops drastically. There is no full validation of cross-language transfer capability; it is not sure whether syntactic knowledge acquired during diffusion can be transferred to low-resource language couples. The future work can be done in three ways: optimizing the denoising network architecture and reducing the computational complexity, introducing incremental sampling strategies to enhance the efficiency of inference; retrieval enhancement mechanism and use of syntactic templates in parallel corpora to guide the generation; and multimodal conditional guidance to enrich the expression forms of translation, adding prosody, emotion and other paralinguistic information.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This work was supported by Joint Program Project of Science and Technology Plan from Scientific Research Project of Science and Technology Department of Liaoning Province in 2023. The number is 2023JH2/101700002.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] D. Holm and M. Yuanyao, "English Translation," The Brigands' Song: Serving in the Army of A Native Chieftain, pp. 85–150, 2021, doi: https://doi.org/10.1163/9789004498754_003.

[2] D. Kripper, "In English Translation," Teaching Literature in Translation, pp. 102–110, 2022, doi: https://doi.org/10.4324/9781003105220-15.

[3] A. Vedelago, "VIRGIL IN ENGLISH TRANSLATION," The Classical Review, vol. 72, pp. 521–523, 2022, doi: https://doi.org/10.1017/s0009840x22000403.

[4] K. Gültan, "Preface to the English Translation," The Purple Color of Kurdish Politics, pp. xiii–xviii, 2022, doi: https://doi.org/10.2307/j.ctv3142tfh.5.

[5] M. L. Miller, "Foreword to the English Translation," The Beginnings of Anti-Jewish Legislation, pp. xiii–xv, 2023, doi: https://doi.org/10.7829/j.ctv2vdbvg5.6.

[6] "Foreword to the English translation," 2022, doi: https://doi.org/10.1515/9783110790900-205.

[7] E. Schrijver, L. Meiboom, S. Arndt, et al., "6. Index of Titles in English or in English Translation," Fables in Jewish Culture, pp. 507–511, 2023, doi: https://doi.org/10.1515/9781501775840-019.

[8] X. Zhang, "Long English Sentence Translation Techniques in Business English Translation," Creativity and Innovation, 2022, doi: https://doi.org/10.47297/wspciwsp2516-252719.20220601.

[9] A. Kaluba, "IT SURPRISES ME (English translation)," Best "New" African Poets 2023 Anthology, p. 6, 2024, doi: https://doi.org/10.2307/jj.13168031.8.

[10] C. Xu and Q. Li, "Machine Translation and Computer Aided English Translation," Journal of Physics: Conference Series, vol. 1881, no. 4, Art. no. 042023, 2021, doi: https://doi.org/10.1088/1742-6596/1881/4/042023.

[11] X. Zhong, "Introduction to the English Translation," Gregory of Nyssa: Homilies on the Our Father. An English Translation with Commentary and SupportingStudies, 2021, doi: https://doi.org/10.1163/9789004463011_005.

[12] S. Mózes and J. Schwartzmann, "English Translation and Hebrew Original," The Path of Moses: Scholarly Essay on the Case of Women in Religious Faith, pp. 41–123, 2022, doi: https://doi.org/10.1163/9789004515000_003.