Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/37/20230463

Syntax-aware bidirectional decoding Neural Machine Translation model

Jia Dong

JCGS High School, Zhenjiang City, Jiangsu Province, 212499, China

dongjiajack@foxmail.com

Abstract. The mainstream model in neural machine translation, the Transformer, relies heavily on self-attention mechanisms for translation operations. This approach has significantly improved both accuracy and speed. However, there are still some challenges. For instance, it lacks the incorporation of linguistic knowledge and the ability to leverage syntactic structure information in natural language for translation, leading to issues such as mistranslation and omission. Addressing the limitations of the Transformer's autoregressive decoding, which decodes from left to right without fully utilizing context information and is prone to exposure bias, this paper proposes a syntax-aware bidirectional decoding neural machine translation model. By employing both forward and backward decoders, the generated decoding results can encompass contextual information. Additionally, the model integrates dependency syntax to generate target language sentences with syntactic guidance. Finally, an optimization strategy involving the Teacher Forcing mechanism is introduced to balance the discrepancies between the Teacher Forcing training phase and the autoregressive testing phase, thus alleviating exposure bias issues.

Keywords: Neural Machine Translation, transformer, autoregressive decoder.

1. Introduction

In recent years, the continuous evolution of machine learning techniques alongside advancements in computer hardware systems has propelled the field of machine translation through distinct phases: from rule-based machine translation and statistical machine translation to the current era of neural machine translation (NMT). Throughout this progression, both the accuracy and efficiency of translation have exhibited gradual improvements. Initially, rule-based machine translation relied on linguists to summarize and deduce transformation rules between different languages, which were subsequently utilized as translation knowledge and executed by computers. Given its inherent reliance on manual labor, rule-based machine translation suffered from drawbacks such as low translation efficiency, challenges in rule extraction, and elevated human resource costs.

With the growth of internet technologies, statistical machine translation emerged as the dominant paradigm for machine translation towards the end of the 20th century. Statistical machine translation employed data models to describe transformation rules between languages, guiding the construction of latent structures to achieve translation across different languages. However, this approach retained several challenges, including difficulties in managing reordering that affected the fluency of generated translations and the inability to capture global dependencies solely through local features.

In 2013, Kalchbrenner and Blunsom introduced a neural network-based approach to machine translation [1]. Subsequently, numerous scholars developed fully neural-based neural machine translation models, resulting in substantial improvements in translation performance. This evolution was driven by the ability of neural networks to capture complex patterns and relationships within linguistic data, consequently overcoming some limitations of previous paradigms. As we navigate the landscape of NMT in this paper, we delve into its underlying mechanisms, address challenges, and explore potential avenues for further advancements. Through this exploration, we contribute to the ongoing discourse on the enhancement of translation technologies in an era characterized by intensified global communication.

The Transformer decoder employs an autoregressive approach, decoding words from left to right. It generates the current word based on the previous word's decoding output and the encoder's output. However, decoding target language sentences using an autoregressive decoder lacks access to information beyond the current word. Additionally, due to the disparity between the Teacher Forcing training phase and the autoregressive testing phase, exposure bias issues can arise. To address these challenges, this paper proposes a syntax-aware bidirectional decoding neural machine translation model. This model leverages contextual information around the current word to predict it, thereby enhancing decoding accuracy. Moreover, the model incorporates the Scheduled Sampling mechanism, which probabilistically replaces reference translations with candidate translations, mitigating the environmental differences between training and testing phases.

2. Related work

The Transformer's autoregressive decoder decodes from left to right during testing, limiting its ability to fully leverage complete context information and potentially slowing down the decoding process. Moreover, the disparity between the Teacher Forcing training and the Autoregressive testing environments gives rise to exposure bias issues. To tackle these concerns, researchers have made various optimizations to the decoder. These optimizations can be broadly categorized into building bidirectional decoding models to exploit contextual information and enhancing the efficiency of decoding through improved beam search strategies.

In 2016, Liu et al. [2] proposed a joint training approach involving bidirectional decoders, aiming to find target words that both decoders assign high probabilities to during testing. This strategy ensures consistency between left-to-right and right-to-left decoders, thus enhancing overall translation quality. In 2017, Freitag et al. [3] accelerated decoding speed by enhancing the beam search strategy. This approach adjusted the candidate window size at each time step based on changes in candidate scores. Additionally, they introduced four pruning methods, including relative threshold pruning, absolute threshold pruning, relative local threshold pruning, and maximum candidate number per node, to boost decoding efficiency. In 2018, Zhang et al. [4] introduced a backward decoder to the encoder-decoder framework. This backward decoder decoded from right to left based on the hidden state sequence generated by the encoder, providing contextual information. The forward decoder then decoded from left to right, ensuring the consideration of context information in every decoding time step to enhance translation quality. In 2019, Zhou et al. [5] introduced Synchronous Bidirectional Decoding in Neural Machine Translation (SDB-NMT), employing synchronized bidirectional decoding. This model simultaneously predicts outputs by interacting between left-to-right and right-to-left decoding. Importantly, the generation from left-to-right (right-to-left) depends not only on previously generated output but also on the future word information predicted by right-to-left (left-to-right) decoding.

Further advancements followed in 2019 when Fu et al. [6] introduced a reference network that integrated the reference process into the translation decoding process. Utilizing local coordinate encoding, they acquired a global context vector encompassing monolingual and bilingual context information. This vector was then employed in the decoding process. In 2021, Feng et al. [7] introduced a Seer decoder in the encoder-decoder framework to access future word information. The Seer decoder guided the original decoder's behavior through knowledge distillation. When the traditional decoder's predicted distribution closely resembled the Seer decoder's, the traditional decoder's performance was considered akin to the Seer decoder's. During testing, the translation model solely reasoned with the traditional decoder to achieve accurate translation results. Additionally, more recent works consider the character-level decoder [8] or model cost [9-10] for the neural machine translation.

3. Method

In natural language, each word in a sentence maintains inherent connections with other words. When predicting a word, using contextual information significantly improves predictive performance compared to relying solely on preceding or subsequent context. For instance, in the sentence "I like eating rice," when considering the known portion of the source sequence, "I like ... rice." and "I like ...," the predictive effect for "eating" is notably better in the former case. Therefore, when decoding target language translations, this paper employs both forward and backward decoders to decode the current word from two directions, effectively utilizing the current word's contextual information. Simultaneously, when predicting a word, the contribution of predictions to unmasked words is assessed, further enhancing decoding accuracy. During the training phase, decoding follows the Teacher Forcing mechanism, while in the testing phase, autoregressive decoding is used. To balance the differences between training and testing environments, the training phase aims to simulate the testing environment as closely as possible, thus reducing this environmental disparity.

3.1. Backward decoder

In this model, the backward decoder decodes the target language from right to left, serving two primary purposes. Firstly, it provides the contextual information decoded from right to left to the forward decoder. This allows the forward decoder to accurately utilize contextual cues for decoding the final target language output. Secondly, the backward decoder's decoding results are probabilistically substituted for the ground truth, thus reducing the environmental disparity between the training and testing phases. The training and decoding flow of the translation model described in this paper is outlined as follows:

First, the natural language sequence, which translates to "I love China." is fed into the encoder. After undergoing word embedding and positional embedding, the encoder generates semantic encoding vectors. These vectors are then input into the backward decoder for decoding. In the backward decoder, training is carried out using the Teacher Forcing method, which involves substituting the decoding results with the reference translation "I love China." The decoding process involves a stack of multiple decoder layers, with each decoder layer comprising three sub-layers.

In the multi-head masked attention mechanism sub-layer, three matrices WQ, WK, and WV are defined. These matrices are utilized to perform linear transformation operations on the input sequence, yielding three new vectors qt, kt, and vt respectively. The collection of all qt vectors is concatenated into a large matrix denoted as the query matrix Q, while the collection of all kt vectors is concatenated into a large matrix called the key matrix K, and the collection of all vt vectors is concatenated into a large matrix referred to as the value matrix V. At this stage, matrices K, Q, and V are all sourced from the decoder. The first word's query vector q is then multiplied with the key matrix K to obtain attention weights for the first word. Subsequently, these weights are subjected to a softmax operation to ensure their sum equals 1. The obtained weights are multiplied with the corresponding word's value vector vt and summed accordingly, resulting in the output of the first word. A similar procedure is carried out for subsequent input vectors to obtain all outputs following the multi-head masked attention mechanism sub-layer. Before applying softmax, the input sequence's word information is masked by adding the initial masking matrix to the associative matrix. This step aims to maintain consistency between training and testing environments possible. The initial masking matrix M for the backward decoder is depicted in Figure 1.

	I	like	eating	rice	
rice	-inf	-inf	-inf	0	t=1
eating	-inf	-inf	0	0	t=2
like	-inf	0	0	0	t=3
I	0	0	0	0	t=4

Figure 1. Initial masking matrix of backward decoder.

3.2. Forward decoder

The proposed model's overall architecture encompasses two forward decoders. To distinguish between the two decoders, the decoder responsible for the first stage of decoding is referred to as Decoder 1, and the decoder for the second stage is termed the forward decoder. Decoder 1 performs synchronous decoding with the backward decoder during the first stage. After decoding, the obtained results are probabilistically replaced with the reference translation to alleviate exposure bias. The forward decoder, operating in the second stage, generates the final target sequence. Once results are generated by the backward decoder and Decoder 1, the forward decoder employs a probability-based substitution of the reference translation. Specifically, the forward decoder's input is the reference translation with the first-stage decoding results substituted in. Additionally, the forward decoder integrates prior contextual information acquired from the backward decoder as prior knowledge to assist in generating the final target language sequence. To facilitate this integration, the forward decoder employs a cross-attention sub-layer to incorporate the contextual information generated by the backward decoder. This is mathematically expressed as shown in Equation (1).

$$Attention\left(Q^{fd},K^{bd},V^{bd}\right)=softmax\left(\frac{Q^{fd}K^{bd}}{\sqrt{d_k}}\right)V^{bd} \tag{1}$$

Where $\mathrm{Q^{fd}}$ represents the query matrix from the forward decoder, $\mathrm{K^{bd}}, \mathrm{V^{bd}}$ represent the key and value matrices from the backward decoder. The decoding process of the forward decoder is analogous to that of the backward decoder, with the key difference being that the forward decoder operates from left to right.

3.3. Teacher forcing decoding optimization

Currently, the majority of neural machine translation models employ the Teacher Forcing mechanism during the training phase. This technique efficiently utilizes the target language translations from bilingual datasets to enhance the model's ability to fit the desired translations accurately. The essence of Teacher Forcing lies in using the true target language translations as inputs to the decoder at each time step during training, instead of relying on the model's own generated outputs. This approach accelerates model convergence and facilitates the learning of meaningful translation patterns.

However, when transitioning to the testing phase, a discrepancy arises. The model must generate translations autonomously without access to the ground truth target language translations. Each generated word becomes the foundation for predicting the next word, creating a chain of dependencies. If the initial predictions deviate slightly from the actual intended translations, these errors can propagate and accumulate, leading to a significant divergence between the model's outputs and the true translations.

To address this challenge and enhance the model's robustness, this paper introduces an innovative optimization to the Teacher Forcing mechanism within a bidirectional decoding Transformer model. This model capitalizes on the strengths of both left-to-right and right-to-left decoding strategies. In contrast to traditional sequential models like RNNs, Transformers process input sequences in parallel, making the incorporation of sequential dependencies more intricate. To tackle this, the proposed model incorporates an additional forward decoder, labeled Decoder 1 in section 3.2. Decoder 1 contributes a left-to-right decoding approach, generating translations as if guided by the true translations from the ground truth. This generated output from Decoder 1 serves as a type of planned sample, aiding the subsequent optimization steps.

By introducing bidirectional decoding and planned sampling within the Transformer framework, this paper contributes to mitigating the issues caused by the discrepancies between training and testing environments, thereby enhancing the translation quality of the model. The innovative utilization of Decoder 1, combined with the synergy between forward and backward decoders, provides a comprehensive solution to address the challenges associated with the Teacher Forcing mechanism in machine translation.

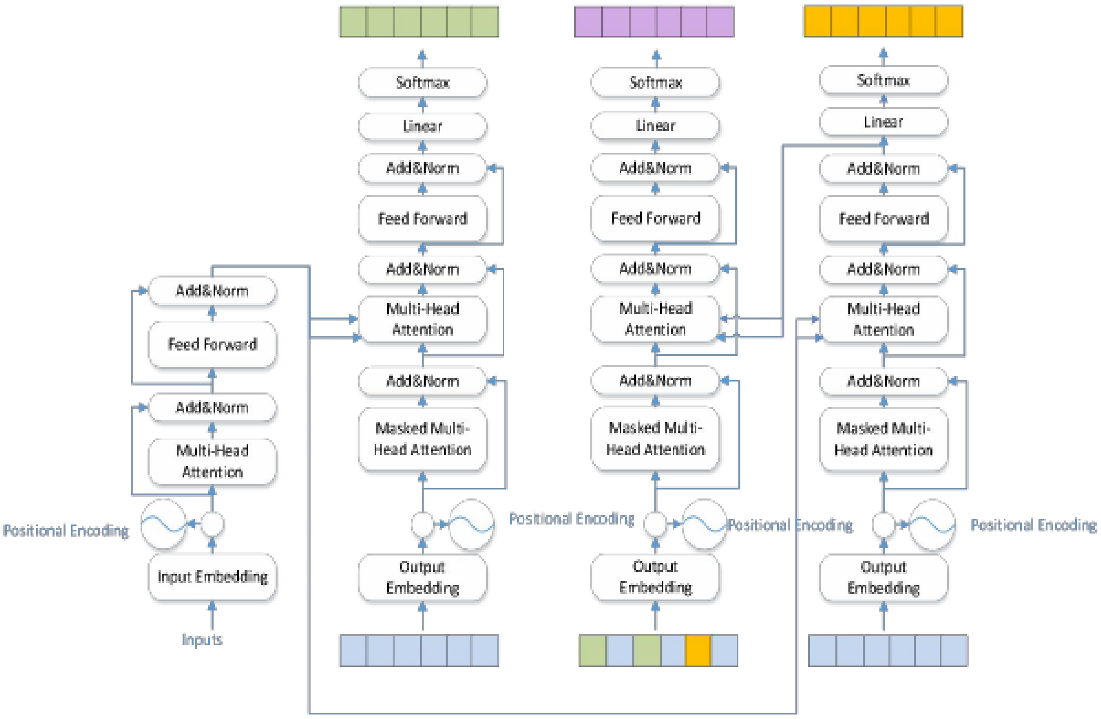

Figure 2. Bidirectional decoding of the Transformer model architecture.

4. Experiment

4.1. Dataset

This paper builds upon the Transformer model proposed by Ashish Vaswani et al. and introduces enhancements. The experiments are conducted on the Chinese-English language pair from the WMT17 dataset and the English-German language pair from the WMT14 dataset. The dataset used for the Chinese-English translation task is the same as described in Chapter 3. For the English-German translation task, the WMT14 dataset is used as the training set, newstest2013 is employed as the validation set, and newstest2014 serves as the test set. Out-of-vocabulary words are represented as "UNK." The scale of the experimental data is illustrated in Table 1.

Table 1. Experimental data scale statistics.

Language Pair	Training Set Sentence Pairs	Validation Set Sentence Pairs	Test Set Sentence Pairs
Chinese-English	227k	2k	2k
English-German	4.5M	3k	3k

4.2. Setting of parameters

The experimental setup of this paper is presented in Table 2 as follows:

Table 2. Experiment environment configuration.

Environment	Specifications
Operating System	Windows 10
Memory	32GB
Disk Space	1TB
CPU	Intel Core i7
GPU	NVIDIA GeForce RTX 2060
Programming Language	Python 3.6 (64-bit)
Deep Learning Framework	TensorFlow

Relevant studies have indicated that when the size of Beam_size exceeds 5, it significantly impairs translation performance, a phenomenon commonly referred to as the "beam search curse," which is listed among the six major challenges in Neural Machine Translation (NMT). In light of this, the testing phase of this paper employs a beam search strategy with different values for beam_size, specifically, setting it to 1, 3, and 5 respectively, for experimental comparison.

4.3. Experimental results and analysis

In the scope of this study, the focus is on enhancing the established Transformer baseline model. Two modified versions are introduced: the Transformer (+BD+DE) model, which incorporates bidirectional decoding and incorporates dependency information, and the Transformer (+BD) model, which omits dependency information integration. A benchmark for comparison is provided by the ABD-NMT model, as proposed by Zhang et al. For the testing phase, a beam search strategy is applied, aligning with prior research that has highlighted the impact of different beam sizes (k) on decoding performance. To systematically address this concern, the beam size parameter is tested at three levels: 1, 3, and 5. The choice of k = 1 corresponds to employing a greedy search approach for decoding. Evaluation of the models is conducted using the BLEU metric, which assesses the quality of translations. The outcome of these experiments is summarized in Table 3 for reference and analysis.

Table 3. Experimental comparison results.

Model	Beam Size	Chinese-English	English-German
Transformer	1	22.33	26.25
	3	23.15	26.98
	5	23.21	27.10
ABD-NMT	1	23.04	26.83
	3	23.67	27.32
	5	23.85	27.51
Transformer (+BD)	1	23.23	26.92
	3	23.96	27.77
	5	24.06	27.72
Δ1		+0.85	+0.62

Table 3. (continued).

Transformer (+BD+DE)	**1**	**23.58**	**27.16**
	3	24.31	27.93
	5	24.38	28.06
Δ2		+1.17	+0.9

From Table 3, it is evident that under the same dataset conditions, the proposed bidirectional decoding translation model outperforms the baseline model and the ABD-NMT model. Translation accuracy using beam search is superior to that of greedy search. For the Transformer(+BD) model, in the English-German dataset, the BLEU score for k=5 is lower than that for k=3. However, for all other translation results, the BLEU score for k=5 is higher than that for k=3 and k=1. Therefore, this paper primarily compares the translation performance of the model at k=5 with the baseline model. The bidirectional decoding translation model proposed in this paper achieves a BLEU score of 24.06 in the Chinese-English translation task, surpassing the baseline model by 0.85 BLEU points. In the English-German translation task, the BLEU score reaches 27.72, which is an improvement of 0.62 BLEU points over the baseline model. Additionally, by introducing dependency syntax in the decoding process based on bidirectional decoding, a further enhancement of 0.32 BLEU points in the Chinese-English translation task and 0.34 BLEU points in the English-German translation task is observed. Overall, it can be concluded that the proposed approach effectively enhances the translation performance of the Transformer model.

5. Conclusion

Addressing the issue of inadequate utilization of contextual information by the autoregressive decoder, this paper proposes a novel approach: the Syntax-Aware Bidirectional Decoding Neural Machine Translation model. By introducing a backward decoder, this model augments the forward decoder with contextual information, ensuring more informed decoding and consequently enhancing translation accuracy. Moreover, the integration of sentence syntax structure during the decoding process aids the model in comprehending semantic nuances. To mitigate the exposure bias problem that arises during decoding, the paper introduces the Scheduled Sampling mechanism, which bridges the gap between training and testing decoding environments. Experimental findings demonstrate that the proposed model outperforms traditional Transformer models in both the Chinese-English and English-German translation tasks. While the inclusion of the backward decoder in the Transformer model resolves the underutilization of contextual information during decoding, it does introduce a time cost to the decoding process. In future research, attention will be directed towards enhancing decoding efficiency. Strategies such as beam search optimization through pruning will be explored to reduce decoding time while maintaining translation quality.

References

[1] Kalchbrenner N, Blunsom P. Recurrent Continuous Translation Models[C]. Conference on Empirical Methods in Natural Language Processing, Washington, D.C., USA, 2013: 1700-1709.

[2] Liu L, Utiyama M, Finch A, et al. Agreement on Target-Bidirectional Neural Machine Translation[C]. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Santiago, Chile, 2016: 411-416.

[3] Freitag M, Al-Onaizan Y. Beam Search Strategies for Neural Machine Translation[C]. the First Workshop on Neural Machine Translation, Vancouver, Canada, 2017: 56-60.

[4] Zhang X, Su J, Qin Y, et al. Asynchronous Bidirectional Decoding for Neural Machine Translation[C]. the AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, 2018: 5698-5705.

[5] Zhou L, Zhang J, Zong C. Synchronous Bidirectional Neural Machine Translation[J]. Transactions of the Association for Computational Linguistics, 2019, 7(5): 91-105.

[6] Fu H, Liu C, Sun J. Reference Network for Neural Machine Translation[C]. Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 3002-3012.

[7] Feng Y, Gu S, Guo D, et al. Guiding Teacher Forcing with Seer Forcing for Neural Machine Translation[C]. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing, Online, 2021: 2862-2872.

[8] Chung J , Cho K , Bengio Y .A Character-level Decoder without Explicit Segmentation for Neural Machine Translation[J]. 2016.DOI:10.18653/v1/P16-1160.

[9] Tu Z , Lu Z , Liu Y ,et al.Modeling Coverage for Neural Machine Translation[J]. 2016.DOI:10.18653/v1/P16-1008.

[10] Duan C , Chen K , Wang R ,et al.Modeling Future Cost for Neural Machine Translation[J]. 2020.DOI:10.48550/arXiv.2002.12558.