The metamorphosis of machine translation: The rise of neural machine translation and its challenges

Yuduo Chen

Sichuan University, School of Mathematics, Unit 2, Building 21, Jinko Ziyuan, 53 East Huangshan Avenue, Yubei District, Chongqing, China

chenyuduo1999@gmail.com

Abstract. Machine translation refers to the process of using computers to translate source language into target language, which has undergone significant transformations since its inception, with the current mainstream neural machine translation achieving satisfactory translation performance. This paper overviews the three developmental stages of machine translation: rule-based machine translation, statistical machine translation, and neural machine translation, with a focus on neural machine translation. It introduces the key models that emerged in the development process of neural machine translation, namely the recurrent neural network encoder-decoder model, recurrent neural network search model, and Transformer, and compares their strengths and limitations. Other relevant technologies and models developed alongside neural machine translation are also discussed. Addressing the current challenges of neural machine translation, the paper delves into issues of overfitting, low-resource translation, structural optimization of Transformer models, and enhancement of neural machine translation interpretability. Finally, the paper explores the prospects of applying neural machine translation to multimodal translation.

Keywords: Machine Translation, Neural Machine Translation, Attention Mechanism, Model Training, Domain Adaptability.

1. Introduction

The issue of translation between different human languages has always received considerable attention. People from diverse regions have variations in language and writing, creating a need for translation when they engage in social activities. Traditionally, human translators have handled translation work, ensuring high accuracy but requiring substantial time and effort. However, with increased global communication the demand for translation has grown exponentially, leading to the exploration of tools to aid in the process. Machine Translation (MT) involves using computers to translate the source language into the target language while preserving semantic equivalence. In 1949, Weaver proposed the concept of MT, suggesting the use of cryptographic methods to tackle the task of translating human languages [1].

Early machine translation technology was based on rule-based methods. Rule-based Machine Translation (RBMT) relied on dictionaries and manually created rule sets to perform translations using a combination of rules [2]. However, this approach had limitations in coverage and was sensitive to noise in the rules or templates. As a result, MT was being questioned and entered a decline.

In the 1980s, as paper-based texts became digitized, there was an increase in computationally readable language data. This led to the possibility of using mathematical models to analyse and infer patterns from the data. In the 1990s, Brown et al. [3] introduced word alignment- based MT method. Statistical Machine Translation (SMT) emerged, using statistical models to automatically learn translation knowledge from language data without the need for manual rule creation. SMT quickly became a prominent approach in MT research and application. In 2006, the release of Google Translate as a free service marked the practical use of MT.

With the development of deep learning, a new approach called Neural Machine Translation (NMT) has emerged. Kalchbrenner and Blunsom [4] introduced an encoder-decoder model using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Their work can be considered the beginning of NMT. Sutskever et al. [5] introduced the sequence to sequence (seq2seq) learning approach with Long Short Term Memory (LSTM) structures. Bahdanau et al. [6] pioneered the use of the attention mechanism in MT. Vaswani et al. [7] proposed the Transformer model which has become the dominant framework in NMT. To this day, the Transformer remains the mainstream architecture for NMT, with many state-of-the-art systems being variants of the Transformer [8-9]. In addition, the exploration and advancements in MT are ongoing.

MT research has identified three key user demands. Firstly, it assists in reading foreign language materials and enables barrier-free communication. Secondly, it enhances human translation efficiency and cost-effectiveness through computer-assisted translation. Lastly, MT enables the processing of multilingual textual data through data analysis. Manual translation alone cannot handle large-scale translation tasks effectively, making MT the only viable solution. Thus, research in MT holds profound significance.

This paper is divided into four sections. Firstly, it reviews the history of MT, tracing the evolution of translation methods from early stages to the present. Then, this paper places an emphasis on explaining the development journey of NMT, analysing the advantages and limitations of various NMT techniques. Next, this paper outlines the current challenges and problems that NMT is facing and discusses the existing solutions to them. Finally, a summarization of the main content is provided, along with a prospective outlook on the future of NMT.

2. Literature review

Neco et al. [10] and Castano et al. [11] have previously attempted to utilize neural network technology in MT tasks. However, due to the limitations of computer processing power and available language resources at that time, these works did not receive much attention. Nevertheless, these works share similarities with many subsequent neural machine translation NMT methods.

Kalchbrenner and Blunsom [4] introduced an encoder-decoder model using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Their work can be considered the beginning of NMT. However, this model did not demonstrate excellent translation performance. The most significant reason is the severe issues of gradient vanishing and gradient explosion encountered during the training process of the neural translation model [12-13].

Beyond machine translation tasks, word embeddings play a crucial role in the entire field of natural language processing (NLP). Word embedding refers to the process of transforming natural language into numerical representations embedded in a mathematical space. Mikolov et al. [14] introduced the Word2Vec method for training word embeddings using neural network models to obtain vector representations of words.

To address the issues of gradient vanishing and gradient explosion in the model proposed by Kalchbrenner and Blunsom [4], Sutskever et al. [5] introduced the sequence to sequence (seq2seq) learning approach with LSTM structures. LSTM units are a variant of recurrent neural network (RNN) units, which introduced gating mechanisms to alleviate the problems of gradient vanishing and gradient explosion in neural networks. Unlike the neural machine translation model proposed by Kalchbrenner and Blunsom [4], the model proposed by Sutskever et al. [5] incorporates recurrent neural networks in both the encoder and decoder, hence referred to as RNNencdec.

The main issue with RNNencdec lies in the encoder part. During encoding, arbitrarily long source language sentences are mapped to a fixed-dimensional vector. This approach underutilizes the storage space and reduces efficiency for shorter source language sentences, while it fails to capture sufficient information from longer source language sentences.

Luong et al. [15] were the first to introduce the attention mechanism in neural machine translation. Luong et al. proposed two models: a global approach, where all source words are attended, and a local approach, where only a subset of source words is considered at a time.

Bahdanau et al. [6] introduced the RNNsearch model. By incorporating the attention mechanism in the decoder, the RNNsearch model dynamically calculates which information from the source language sentences is relevant at each decoding step, creating a real-time context vector. Compared to RNNencdec, which maps any source language sentence to a fixed-dimensional vector, RNNsearch offers stronger expressive power and resolves the issue of long-distance dependencies that existed in the past.

In neural machine translation, algorithm optimization during training is also an important concern. Algorithm optimization refers to improving the training process to minimize the loss function. Diederik et al. [16] introduced the Adam optimization algorithm. In practical applications, the Adam method performs well, converges faster compared to other adaptive learning rate algorithms, and addresses certain issues present in other optimization techniques.

Wu et al. [17] proposed Google s Neural Machine Translation (GNMT) system, which is built upon the RNNsearch model. Wu et al. compared the phrase based statistical translation model with the RNNsearch model of neural machine translation for six language pairs. The results showed that the neural machine translation model represented by RNNsearch achieved significant improvements in translation performance when trained on large-scale corpora. Since then, neural machine translation gradually replaced statistical machine translation and became the mainstream research direction in machine translation. However, the RNNsearch model also has limitations. LSTM operates sequentially, processing words in a sequence one by one, which makes the model less efficient. Additionally, due to the different distances that words need to propagate in a sequence, RNNsearch fails to fully exploit the texts information features.

Due to the limitations of recurrent neural network models, researchers have attempted to use other neural network models to address machine translation tasks. Gehring et al. [18] proposed a model that is entirely based on CNN. Convolutional neural networks exhibit high parallelism, enabling the translation model to be trained in a fully parallel manner, greatly enhancing efficiency.

Vaswani et al. [7] proposed the Transformer model which solely relies on attention mechanisms and eliminates the need for RNN and CNN. Vaswani et al. introduced variations of the attention mechanism, namely multi-head attention and self-attention, in the Transformer model. The advantage of the Transformer model lies in its use of simple matrix operations, which allows for high parallelization and overcomes the efficiency issues of RNNsearch. Additionally, through the multi-head attention mechanism, the Transformer model can better capture text information features.

Integrating human prior knowledge with data-driven neural network methods is also an important research direction. The introduction of prior knowledge may help increase the interpretability of neural machine translation models and provide valuable information during translation. Gu et al. [19] proposed the Sequence to Doubly-Recurrent Neural Networks (Seq2DRNN) model, which incorporates syntactic tree structures in the decoder and introduces attention mechanisms specific to syntactic information. Wang and Xiong [20] proposed a simple method to incorporate predefined bilingual pairs into NMT, which does not require modifications to the decoding search algorithm or complex modifications to the model.

Current neural machine translation models perform well on sentence-to-sentence translation tasks, but they still have limitations in handling document-level translation tasks. Miao et al. [21] suggested that document-level translation should be based on clauses rather than sentences. They supplemented neural machine translation models with clause knowledge by manually annotating clause alignments. They also introduced a multi-way coordination self-attention mechanism (MC-SefAtt) to enhance the encoder's representation of the semantics of clause in the source language.

During training, neural machine translation models rely on large-scale parallel corpus data. In situations where parallel corpus data is limited, the translation quality of neural machine translation models can significantly decrease, sometimes even falling below that of statistical machine translation models. Therefore low-resource translation problems especially for language pairs with scarce or no parallel data, are important research areas. Zhu et al. [22] addressed this issue by using contextual information from monolingual data to learn the probability distribution of low-frequency words. Zhu et al. [22] then recalculated the word embeddings of low-frequency words based on this distribution and retrained the Transformer model, effectively alleviating the problem of inaccurate representation of low-frequency words.

3. Discussion

Although NMT has greatly improved translation quality compared to rule-based and statistical machine translation approaches, it still faces several challenges. Firstly, during training, NMT models encounter overfitting and the issue of co-adaptation of neural units. Secondly, NMT relies on data-driven methods, so it performs poorly in low-resource domains where data is scarce. Additionally, NMT models have complex mechanisms and structures, making it important to optimize the model architecture and incorporate prior knowledge. Lastly, enhancing the interpretability of NMT and enabling external intervention are essential concerns. These challenges significantly impact the overall effectiveness of NMT. The following sections will provide further discussion on these four aspects.

Firstly, NMT models have complex structures and strong representation capabilities. However, since training samples are limited, the models tend to excessively adjust their parameters to fit the specific training data, resulting in poor generalization performance on much larger test sets. This phenomenon is known as overfitting. To address this, Szegedy et al. [23] proposed a type of regularization technique named Label Smoothing, which has also been applied in Transformer models [7]. Ideally, each neuron in an NMT model should contribute independently to the final translation prediction. However, with increasing neuron numbers and network complexity, the contribution of one neuron to the output becomes correlated with other neurons, leading to the phenomenon of co-adaptation. This phenomenon allows the neural network model to better capture patterns in the training data but also exacerbates overfitting. Dropout, introduced by Hinton et al. [24], is a commonly used method to mitigate overfitting. It randomly deactivates a portion of neurons during training to prevent overfitting caused by co-adaptation. However, in the Transformer model, the effectiveness of Dropout is not very pronounced, and overfitting issues still persist in deep models. Fan et al. [25] proposed the method named Layer Dropout as an alternative. Unlike conventional Dropout, Layer Dropout randomly drops self-attention sub-layers or feed-forward neural network sub-layers during training to reduce co-adaptation between different sub-layers. This approach aims to alleviate overfitting in the Transformer model.

Secondly, NMT models require a significant amount of parallel bilingual data for training. However, many language pairs lack sufficient parallel data, leading to the challenge of low-resource translation. Researchers have proposed various approaches to tackle this issue and improve the translation quality of NMT models under low-resource conditions. One approach is the utilization of semi-supervised methods. Sennrich et al. [26] introduced the back-translation method, which involves training a target-to-source translation model using a small-scale parallel corpus. The target language monolingual data is then translated into the source language to create a large-scale pseudo-parallel corpus. The final NMT model for source-to-target translation is trained using a combination of the small-scale parallel corpus and the generated pseudo-parallel corpus. In addition, unsupervised approaches have been explored. Lample et al. [27] proposed an unsupervised NMT model based on autoencoders, where sentences are mapped to a latent semantic space and reconstructed using a decoder. Artetxe et al. [28] introduced a shared encoder unsupervised NMT model based on cross-lingual word embeddings. Ruiter et al. [29] employed self-supervised learning using non-parallel sentences with similar content as an auxiliary task. Transfer learning techniques have also been investigated. Zoph et al. [30] proposed a transfer learning approach for low-resource NMT, where an NMT model is initially trained on a high-resource language pair and then fine-tuned on the low-resource language pair. Gu et al. [31] explored the use of shared encoder representations in multilingual NMT models to leverage information from high-resource language pairs. When parallel data is scarce between the source and target languages but abundant between each language and a third language, pivot-based methods can be employed. Cheng [32] suggested training separate translation models for the source-to-pivot and pivot-to-target directions. Leng et al. [33] applied pivot-based low-resource translation methods to unsupervised NMT models.

Thirdly, the Transformer has become the most popular neural machine translation model. However. the structure of the Transformer itself presents several issues. On one hand, the attention mechanism in the Transformer disregards the positional relationships between input units, making it difficult for the Transformer to differentiate between local dependencies and long-range dependencies. Shaw et al. [34] attempted to address this by introducing relative positional information to emphasize local dependencies. Gulati et al. [35] replaced the self-attention mechanism in the Transformer with lightweight convolutional or dynamic convolutional networks in both the encoder and decoder while retaining the encoder-decoder attention mechanism, thereby enhancing the model’s ability to model local information to some extent. On the other hand, the attention mechanism increases the computational complexity, resulting in slow translation speeds when dealing with longer text sequences. One approach to tackle this issue is to limit the scope of the self-attention mechanism. Qiu et al. [36] proposed the method of chunk-level attention, which divides the sequence into fixed-sized segments, and the attention model operates only within the corresponding segment. Another approach involves allowing the model to learn the scope of the attention mechanism instead of predefining it. Kitaev et al. [37] introduced the Reformer model, which reduces the computational range of the attention mechanism by incorporating the locality-sensitive hashing attention. In addition to modifications to the Transformer model, incorporating prior knowledge such as syntactic knowledge is also a research direction worth exploring. Translation examples have shown that current neural machine translation models often suffer from issues like over-translation or incoherent translation due to their insufficient learning capacity on syntactic knowledge and semantic structures [38]. However, prior knowledge such as syntactic knowledge and bilingual dictionaries are typically indiscrete forms, making it crucial to investigate how to integrate such discrete knowledge into continuous representations of neural machine translation models. On one hand, one can consider incorporating syntactic knowledge from the source language. Eriguchi et al. [39] used syntactic tree structures in the encoder and employed a tree-structured recurrent neural network to encode the source language sentences. Chen et al. [40] integrated source language syntactic rules into both the encoder and decoder. This approach allows the attention mechanism to focus more on phrase level information, avoiding excessive attention on individual words and the problem of repetitive translations, ensuring translation coherence. On the other hand, one can consider incorporating syntactic knowledge from the target language. Aharoni and Goldberg [41] transformed the target language sentences into linear sequences based on syntactic trees. These sequences included words and syntactic structure tags of the target language sentences and were used to directly train a sequence-to-sequence model based on attention mechanism. Wu et al. [42] introduced an additional decoder in the decoder side to parse the generated translated sentence into a sequence of dependency structure information. The sequence representation of the generated translation on the dependency structure can serve as an additional input to guide the generation of subsequent translations.

Fourthly, a notable concern in neural machine translation is the limited interpretability of the translation models. The high-dimensional real-valued vectors and numerous non-linear functions within the hidden layers of machine translation models make it challenging to comprehend the information contained within them. To achieve reliable and controllable machine translation systems, enhancing the interpretability of these models is crucial. Currently, there is significant research focusing on interpreting deep neural network models in machine translation. One approach involves investigating the influence of the internal structure of the models on their outputs. Voita et al. [43] analysed the roles of different attention heads in Transformer models and studied three attention functions: positional capturing semantic capturing, and low-frequency word capturing. Another approach is to design auxiliary tasks to examine whether the models can capture specific information. Dalvi et al. [44] proposed the linguistic correlation analysis task to interpret the linguistic information represented by specific dimensions in the intermediate states of neural networks. They also introduced the cross-model correlation analysis task to identify the primary functioning units in machine translation models. In the context of Transformer models, Raganato and Tiedemann [45] discovered that the self-attention mechanism allows for the calculation of correlation coefficients between words in a sentence. These coefficients can be used to organize the sentence into a tree structure. In tasks with abundant training data, the extracted tree structure closely resembles the standard dependency tree.

4. Conclusion

In the course of human civilization, translation between different languages has always been an important requirement. With advancements in technology, machine translation has demonstrated remarkable effectiveness in certain translation tasks and has emerged as a significant research direction. This paper provides an overview of the development of machine translation since 1949 and focuses on introducing NMT, which has become the mainstream approach since the 2010s. Subsequently, by reviewing the work of previous researchers, this paper attempts to analyse and compare the strengths and limitations of various NMT methods. Finally, the challenges faced by NMT and potential solutions are discussed. NMT builds upon the data-driven approach of statistical machine translation while introducing distributed representations for modeling textual sequences, offering a fresh perspective on translation. NMT has significantly improved translation quality and continues to be an active area of research and development in the field of machine translation. Moreover, the potential of NMT is still being explored, and it can be anticipated that research on NMT will continue in the coming years.

However, NMT still faces several challenges and difficulties. These challenges can be categorized into four main aspects. Firstly, NMT models often suffer from overfitting during training. Secondly, low-resource translation remains a significant challenge. Thirdly, there is a need for structural optimization of the dominant Transformer models used in NMT. Fourthly, enhancing the interpretability of NMT systems is another important area of focus. Many limitations observed in NMT, such as discourse translation and issues of under translation or over translation, are related to these four aspects. In future research, apart from addressing the theoretical challenges mentioned above, considerations from an application perspective can also be vital. In practical applications, translation tasks involve not only textual data but also speech recognition or video information. This type of problem is known as multimodal translation. Exploring the integration of NMT techniques with technologies like speech recognition and image understanding is an essential research direction. It is believed that with further research and technological advancements machine translation will continue to provide greater assistance to human production and daily life.

References

[1] Weaver W 1952 Translation Proc. of the Conf. on Mechanical Translation (Carlsbad)

[2] Zarechnak M 1979 The history of machine translation Trends in Linguistics: Studies and Monographs vol 11, ed Chiara Gianollo and Daniel Van Olmen (Berlin: De Gruyter Mouton) part 1 pp 3-87

[3] Brown P F, Cocke J, Della Pietra S A, Della Pietra V J, Jelinek F, Lafferty J D, Mercer R L and Roossin P S 1990 A statistical approach to machine translation Comput. Linguistics **16(2)** 79-85

[4] Kalchbrenner N and Blunsom P 2013 October Recurrent continuous translation models Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing (Seattle: Association for Computational Linguistics) pp 1700-09

[5] Sutskever I, Vinyals O and Le Q V 2014 Sequence to sequence learning with neural networks Adv. neural inf. process syst. 27

[6] Bahdanau D, Cho K and Bengio Y 2015 Neural machine translation by jointly learning to align and translate 3rd Int. Conf. on Learning Representations (San Juan)

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is all you need Adv. Neural Inf. Process. Syst. 30

[8] Devlin J, Chang M W, Lee K and Toutanova K 2019 Bert: Pre-training of deep bidirectional transformers for language understanding Proc. of NAACL-HLT (Minneapolis) pp 4171-86

[9] Liu X, Duh K, Liu L and Gao J 2020 Very deep transformers for neural machine translation Preprint arXiv:2008.07772

[10] Neco R P and Forcada M L 1997 June Asynchronous translations with recurrent neural nets Proc. of Int. Conference on Neural Networks (Houston) vol 4 pp 2535-40

[11] Castano A and Casacuberta F 1997 A connectionist approach to machine translation 5th European Conf. on Speech Communication and Technology (Rhodes)

[12] Hochreiter S 1998 The vanishing gradient problem during learning recurrent neural nets and problem solutions Int. J. Uncertain. Fuzz. **6(02)** 107-16

[13] Bengio Y, Simard P and Frasconi P 1994 Learning long-term dependencies with gradient descent is difficult IEEE Trans. Neural Netw. **5(2)** 157-66

[14] Mikolov T, Sutskever I, Chen K, Corrado G S and Dean J 2013 Distributed representations of words and phrases and their compositionality Adv in Neural Inf. Process. Syst. 26

[15] Luong M T, Pham H and Manning C D 2015 Effective approaches to attention-based neural machine translation Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing (Lisbon) pp 1412-21

[16] Kingma D P and Ba J 2014 Adam: A method for stochastic optimization Preprint arXiv:1412.6980

[17] Wu Y et al. 2016 Google's neural machine translation system: Bridging the gap between human and machine translation Preprint arXiv:1609.08144

[18] Gehring J, Auli M, Grangier D, Yarats D and Dauphin Y N 2017 July Convolutional sequence to sequence learning In Int. Conf. on Machine Learning pp 1243-52

[19] Gū J, Shavarani H S and Sarkar A 2018 Top-down tree structured decoding with syntactic connections for neural machine translation and parsing Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing pp 401-13

[20] Wang T and Xiong D 2022 Enhancing Neural Machine Translations with Pre-Defined Bilingual Pairs J. Chin. Inf. Process **6(36)**

[21] Miao G, Liu M, Chen Y, Xu j, Zhang Y and Feng W 2022 Incorporating Clause Alignment Knowledge into Chinese-English Neural Machine Translation Acta Sci. Nat. Univ. Pekin **58(1)** 8

[22] Zhu J, Yang F, Yu Z, Zou X and Zheng Z 2022 Low Resource Neural Machine Translation with Enhanced Representation of Rare Words J. Chin. Inf. Process **6(36)**

[23] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 Rethinking the inception architecture for computer vision Proc. of the IEEE Conf. on computer vision and pattern recognition pp 2818-26

[24] Hinton G E, Srivastava N, Krizhevsky A, Sutskever I and Salakhutdinov R R 2012 Improving neural networks by preventing co-adaptation of feature detectors Preprint arXiv:1207.0580

[25] Fan A, Grave E and Joulin A 2019 September Reducing transformer depth on demand with structured dropout Int. Conf. on Learning Representations (New Orleans)

[26] Sennrich R, Haddow B and Birch A 2015 Improving neural machine translation models with monolingual data Proc. of the 54th Annu. Meeting of the Association for Computational Linguistics (Berlin) vol 1 pp 86-96

[27] Lample G, Conneau A, Denoyer L and Ranzato M A 2018 Unsupervised machine translation using monolingual corpora only Int. Conf. on Learning Representations (Vancouver)

[28] Artetxe M, Labaka G, Agirre E and Cho K 2018 February Unsupervised neural machine translation Int. Conf. o Learning Representations (Vancouver)

[29] Ruiter D, Espana-Bonet C and van Genabith J 2019 July Self-supervised neural machine translation Proc. of the 57th Annu. Meeting of the Association for Computational Linguistics (Florence) pp 1828-34

[30] Zoph B and Le Q V 2016 Neural architecture search with reinforcement learning Int. Conf. on Learning Representations (San Francisco)

[31] Gu J, Hassan H, Devlin J and Li V O 2018 Universal neural machine translation for extremely low resource languages Proc. of NAACL-HLT 2018 (New Orleans) vol 1 pp 344-54

[32] Cheng Y, Yang Q, Liu Y, Sun M and Xu W 2019 Joint training for pivot-based neural machine translation Proc. of the 26th Int. Joint Conf. on Artificial Intelligence (Melbourne) vol 1 pp 41-54

[33] Leng Y, Tan X, Qin T, Li X Y and Liu T Y 2019 July Unsupervised pivot translation for distant languages Proc. of the 57th Annu. Meeting of the Association for Computational Linguistics (Florence) pp 175-83

[34] Shaw P, Uszkoreit J and Vaswani A 2018 Self-attention with relative position representations Proc. of NAACL-HLT 2018 (New Orleans) pp 464-68

[35] Gulati A et al. 2020 Conformer: Convolution-augmented transformer for speech recognition Interspeech (Shanghai)

[36] Qiu J, Ma H, Levy O, Yih S W T, Wang S and Tang J 2020 Blockwise self-attention for long document understanding Findings of the Association for Computational Linguistics: EMNLP 2020 pp 2555-65

[37] Kitaev N, Kaiser L and Levskaya A 2019 September Reformer: The efficient transformer Int. Conf. on Learning Representations (New Orleans)

[38] Li J, Xiong D, Tu Z, Zhu M, Zhang M and Zhou G 2017 Modeling source syntax for neural machine translation Proc. of the 55th Annu. Meeting of the Association for Computational Linguistics (Vancouver) vol 1 pp 688-97

[39] Eriguchi A, Hashimoto K and Tsuruoka Y 2016 August Tree-to-sequence attentional neural machine translation Proc. of the 54th Annu. Meeting of the Association for Computational Linguistics (Berlin) pp 823-33

[40] Chen H, Huang S, Chiang D and Chen J 2017 Improved neural machine translation with a syntax-aware encoder and decoder Preprint arXiv:1707.05436

[41] Aharoni R and Goldberg Y 2017 July Towards string-to-tree neural machine translation Proc. of the 55th Annu. Meeting of the Association for Computational Linguistics (Vancouver) vol 2 pp 132-40

[42] Wu S, Zhang D, Yang N, Li M and Zhou M 2017 July Sequence-to-dependency neural machine translation Proc. of the 55th Annu. Meeting of the Association for Computational Linguistics (Vancouver) vol 1 pp 698-707

[43] Voita E, Talbot D, Moiseev F, Sennrich R and Titov I 2019 Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned Proc. of the 57th Annu. Meeting of the Association for Computational Linguistics (Florence) pp 5797-808

[44] Dalvi F, Durrani N, Sajjad H, Belinkov Y, Bau A and Glass J 2019 July What is one grain of sand in the desert? analyzing individual neurons in deep nlp models Proc. of the AAAI Conf. on Artificial Intelligence vol 33 pp 6309-17

[45] Raganato A and Tiedemann J 2018 An analysis of encoder representations in transformer-based machine translation Proc. of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP pp 287-97