applied sciences

MDPI

Article

Towards Understanding Neural Machine Translation with Attention Heads' Importance

Zijie Zhou [1,2], Junguo Zhu [1,2,*] and Weijiang Li [1,2]

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; zhouzijie@stu.kust.edu.cn (Z.Z.); lwj@kust.edu.cn (W.L.)
[2] Yunnan Provincial Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China
* Correspondence: jg.zhu@kust.edu.cn

Abstract: Although neural machine translation has made great progress, and the Transformer has advanced the state-of-the-art in various language pairs, the decision-making process of the attention mechanism, a crucial component of the Transformer, remains unclear. In this paper, we propose to understand the model's decisions by the attention heads' importance. We explore the knowledge acquired by the attention heads, elucidating the decision-making process through the lens of linguistic understanding. Specifically, we quantify the importance of each attention head by assessing its contribution to neural machine translation performance, employing a Masking Attention Heads approach. We evaluate the method and investigate the distribution of attention heads' importance, as well as its correlation with part-of-speech contribution. To understand the diverse decisions made by attention heads, we concentrate on analyzing multi-granularity linguistic knowledge. Our findings indicate that specialized heads play a crucial role in learning linguistics. By retaining important attention heads and removing the unimportant ones, we can optimize the attention mechanism. This optimization leads to a reduction in the number of model parameters and an increase in the model's speed. Moreover, by leveraging the connection between attention heads and multi-granular linguistic knowledge, we can enhance the model's interpretability. Consequently, our research provides valuable insights for the design of improved NMT models.

Keywords: neural machine translation; interpretability; linguistics

Citation: Zhou, Z.; Zhu, J.; Li, W. Towards Understanding Neural Machine Translation with Attention Heads' Importance. *Appl. Sci.* **2024**, *14*, 2798. https://doi.org/10.3390/app14072798

Academic Editor: Stefan Fischer

Received: 5 March 2024
Revised: 22 March 2024
Accepted: 24 March 2024
Published: 27 March 2024

1. Introduction

With the advent of attentional mechanisms [1] and sequence-to-sequence learning [2], neural machine translation (NMT) has rapidly advanced, narrowing the gap between machine and human language capabilities, particularly in understanding and generating translations. The Transformer model [3] excels across a spectrum of machine translation tasks, with its attention mechanism being a pivotal component that offers exceptional parallelism, facilitating efficient large-scale data processing. Despite these advancements, the intricacy of the model's structure and the opaque nature of decision making in NMT models render them akin to black boxes. This obscurity hinders a clear understanding of the specific decisions made and the precise attributes of attention learned. Moreover, the inherent complexity of machine translation tasks, coupled with the sophisticated model architectures designed to tackle them, exacerbates this challenge [4].

For humans, NMT models often appear as "black boxes", while interpretability serves as a "crystal ball" that elucidates the models' internal decision-making processes. This clarity not only aids researchers in debugging the models but also provides a theoretical foundation for trustworthy AI. The interpretability of NMT models has been tackled from two primary, complementary perspectives. One approach emphasizes visualizing the decision-making processes and simulating the models' internal mechanisms to cultivate

genuine trust between humans and the models [5]. The other perspective aims to thoroughly understand the decisions made by the model from a human standpoint, identifying the model's learned content through language knowledge that is accessible to people [6].

We chose the state-of-the-art NMT model, the Transformer model, as our object of study. This model is a neural network based on the self-attention mechanism and includes multiple layers of encoders and decoders. Each layer is composed of several attention mechanism modules and feed-forward neural network modules. In our experiments, we employ the standard Transformer model configuration, where both the encoder and decoder consist of six layers, with each layer having eight attention heads.

In this paper, we delve into the importance of attention heads in deciphering the decision-making processes of the model. Employing the saliency method enhances the model's interpretability [7]. Specifically, we evaluate the impact of each attention head on the translation performance of NMT models by employing the Masking Attention Heads method, which serves as a metric to ascertain the importance of each attention head [8]. The verification of the importance of attention heads is achieved by masking them, thereby confirming their importance. An analysis is conducted on the distribution of important attention heads from diverse perspectives. Through the application of two correlation evaluation techniques, Pearson's correlation coefficient and KL divergence, our research indicates a significant relationship between the importance of attention heads and their contribution to part-of-speech (POS), revealing varying levels of relevance across different POS categories to the attention heads [9].

To enhance our comprehension of the decision-making processes within the model, we translate the obscure decisions of the model into linguistic knowledge that is accessible to humans. Our analysis examines the correlation between the importance of attention heads and linguistic knowledge across three distinct levels: POS, dependency relations, and syntactic trees. This analysis sheds light on how the Transformer model processes information at the word, phrase, and sentence levels. We posit that linguistic knowledge can facilitate our understanding of model decisions. Our experiments reveal that important attention heads acquire more knowledge about nouns and adjectives. Moreover, prepositions are predominantly learned by the most important attention head, whereas verbs are distributed more evenly across all attention heads.

In summary, the organization of this paper is outlined as follows: Section 1 introduces the concept of interpretability in NMT, setting the stage for our research objectives and forthcoming experimental investigations. Section 2 provides an overview of the current landscape in interpretability and linguistics research, pinpointing prevalent issues and suggesting potential remedies. Section 3 delves into the importance of attention heads utilizing the masking method, affirms its validity, and explores the distribution of attention heads. Section 4 delineates the relationship between attention heads and linguistic knowledge, interpreting model behaviors through understandable linguistic insights. Finally, Section 5 offers conclusions and outlines future directions for this research.

Contributions

Our primary contributions are as follows:

- We offer a comprehensive examination of the three types of attention mechanisms in the Transformer model: encoder self-attention, decoder self-attention, and encoder–decoder attention. This thorough analysis enriches our understanding of the model's attention mechanisms;
- Our study investigated the distribution and importance of attention heads. This helps researchers retain important attention heads and remove unimportant ones, thereby reducing model parameters and improving model speed;
- Our findings reveal a link between attention heads and the understanding of specific POS features, offering theoretical insights for future studies. Specifically, we suggest that focusing on the acquisition of noun and verb knowledge in Chinese–English machine translation could potentially improve model performance;

- We analyze the decision-making processes of attention heads based on POS, dependency relations, and syntactic trees, which may inform and inspire model design. Our research indicates that certain linguistic elements, such as nouns, adjectives, and adjective modifiers, are pivotal, whereas others, like determiners, are less critical. This discovery suggests that in Chinese–English machine translation, the model focuses more on certain information, which enhances the model's interpretability.

2. Background

2.1. Interpreting Attention

Interpretability lacks a unified definition. Some researchers perceive it as the extent to which people can understand the reasons behind decisions, while others believe that interpretability refers to an individual's ability to consistently and accurately predict the outcomes of model predictions [10]. Due to the complexity of widely used neural network models, their internal logic and working principles are difficult for users to understand. Users are unclear about what comprehensible linguistic knowledge the model has acquired through training and how the model processes language information. Consequently, we propose that interpretability involves bridging the gap between models that are unintelligible to humans and the integration of human-comprehensible linguistic knowledge to elucidate model behavior.

In the realm of NMT interpretability, numerous studies have concentrated on explaining the importance of various model components. These studies enable an intuitive observation of the elements crucial to model predictions, thereby enhancing the interpretability of models. Recent inquiries into attention interpretability have honed in on the functionalities of attentional mechanisms and the implications of attention weights, with research on attention heads gaining increasing traction [11]. There has been a significant emphasis on dissecting the influence of diverse components within models for enhancing interpretability [12]. Attention as the driving force behind the Transformer, the state-of-the-art NMT model, is garnering escalating scholarly attention. Previous investigations underscored the intrinsic value of attention. For example, Serrano [13] employed intermediate representation erasure to ascertain the impact of attentional mechanisms, while Li [14] explored the alignment effect of attention weights. However, as recent studies indicate, relying solely on attention weights to interpret attentional processes is inadequate [15]. The debate over the explicability of attention also constitutes a major discourse [16,17]. Therefore, in our approach to interpreting attention, we pivot towards leveraging linguistic knowledge that is understandable to humans.

The core of the attention layer is the multi-head attention, formed by concatenating the outputs of multiple attention heads. This architecture allows for the learning of diverse feature aspects. Recent studies have scrutinized the redundancy within multi-head attention mechanisms, revealing that attention heads vary in importance [8]. Voita [18] utilized Layer-wise Relevance Propagation (LRP) [19] to ascertain the contribution of individual attention heads in each layer towards model predictions, further investigating the functions of critical attention heads. The importance of attention heads has also been explored in multilingual and cross-linguistic contexts [20]. Motivated by research into the information processed by attention mechanisms [21], we aim to analyze the multi-grain linguistic knowledge acquired by important attention heads, thereby elucidating the factors influencing model decisions [22].

In our experiments, we analyze the relationship between attention heads and multi-granularity linguistic knowledge by masking each attention head in every layer of three different types of attention mechanisms: encoder self-attention, decoder self-attention, and encoder–decoder attention. We aim to explain the reasons and content behind the model's decisions using linguistic knowledge that humans can understand.

2.2. Linguistic Knowledge

Linguistics delves into the structure, function, and essence of human language, encompassing areas such as language structure, phonology, word formation, syntax, and semantics.

In the realm of interpretability, linguistic knowledge is particularly valued for its ability to render model decisions understandable to humans. Many researchers advocate for using linguistic knowledge as a basis for elucidating models [23]. Recent studies have increasingly interpreted NMT models using linguistic insights, including aspects like morphology [6] and POS [24]. However, these studies primarily focus on the impact of individual word information in NMT. Consequently, we have explored the utilization of inter-word interactions and multi-word information to shed light on model decisions [25]. Understanding the interplay between words and leveraging multiple-word information can make translations smoother, achieving the "faithfulness, expressiveness, and elegance" standard of translation. Additionally, this approach can improve the accuracy and fluency of translations involving complex sentence structures, such as idioms and colloquialisms.

This paper investigates the decisions made by the model's important heads at three different granularities of linguistic knowledge: POS, dependency relations, and syntactic trees. POS reflects the role of individual word information in a sentence in NMT and helps us understand how the model handles individual words. Dependencies reflect the dominant relationship between two words in a sentence and help us understand how the model handles inter-word relationships. The syntax tree reflects the relationships between words in a sentence and helps us understand how the model handles multiple words as well as syntactic structure.

Finally, in Table 1, we compared two categories of interpretability: visualization and linguistic interpretation. We explained their strengths and limitations and provided explanations for some references and their main contributions in order to better understand the main contributions of our study.

Table 1. An overview of the strengths and limitations of visualization and linguistic interpretation, as well as some references and their main contributions.

Categories	Strengths	Limitations	References	Contributions
Visualization	1. Intuitive representation of textual data and processes; 2. Provides visually intuitive charts; 3. Accessible to non-experts.	1. Difficulty in explaining complex grammatical structures; 2. May overlook subtle nuances in text; 3. Limited in-depth linguistic analysis.	Visualizing and understanding neural machine translation [5].	Understanding the internal workings of NMT models through attention-based visualization techniques.
			Rethinking the value of transformer components [12].	Understanding the importance of each component of the Transformer model through visualization.
			Towards Understanding Neural Machine Translation with Word Importance [26].	Understanding the importance of words in generating sentences through visualization.
Linguistic interpretation	1. Provides in-depth analysis and interpretation based on linguistic theories; 2. Explains reasons and patterns behind language phenomena; 3. Aids understanding of language structures and semantics.	1. Requires linguistic expertise for interpretation; 2. Relies heavily on corpora and linguistic resources; 3. May not cover all language phenomena, especially non-structural and abstract features.	Interpreting language models with contrastive explanations [23].	Explaining the model's behavior by contrasting through word replacement.
			What do neural machine translation models learn about morphology [6]?	Explaining the internal processing of the model through morphology.
			Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks [24].	Explaining the internal processing of the model through POS and semantic.

Numerous studies have explored the interpretability of NMT. In this paper, we link the importance of attention heads to linguistic knowledge, using visualization to understand the importance of attention heads in the model, and employing linguistic knowledge to understand the content of the model's decision making.

3. Attention Heads' Importance

3.1. Experimental Setup

To perform a robust and reproducible evaluation of the disclosed approach, we chose a large-scale, publicly available dataset for our study. Specifically, for the Chinese–English translation task, we use the CWMT21 Chinese–English dataset, which contains 9.0M sentence pairs [27]. As detailed in Table 2, this dataset is segmented into training, testing, and validation sets in the ratio of 0.9:0.05:0.05. Following the standard NMT procedure, we adopt the standard byte pair encoding (BPE) [28] for all language pairs. We use the standard SacreBLEU score as the evaluation metric for translation performance. The BLEU score is a widely used metric for evaluating the quality of machine translation. It measures the similarity between the machine-generated translation and one or more reference translations based on n-gram precision and a brevity penalty [29].

Table 2. The dataset for Chinese–English.

ZH-EN	casia2015 corpus	1 million	9 million
	casict2011 corpus	2 million	
	casict2015 corpus	2 million	
	datum2015 corpus	1 million	
	datum2017 corpus	1 million	
	neu2017 corpus	2 million	

We employ the state-of-the-art Transformer, which is composed of two parts: an encoder and a decoder. Each part is divided into six layers, each layer has eight attention heads. A particularity of this model is that it features three distinct attention mechanisms: encoder self-attention (Enc-Enc), decoder self-attention (Dec-Dec), and encoder–decoder attention (Enc-Dec), all of which use multi-headed attention (MHA). Our study analyzed the importance of the attention head to all of these three attention mechanisms. We analyzed the learning of attention heads in the Transformer model, which includes information on individual words (part of speech), inter-word relationship information (dependency relations), and overall sentence structure information (syntax tree). Based on the importance ranking of the attention heads, we established a relationship between these three types of human-understandable, multi-granular linguistic knowledge (part of speech, dependency relations, syntax tree) and the attention heads. Most of the experiments were carried out on the test set, and the model was trained by fairseq [30] in standard settings.

3.2. Multi-Headed Attention

NMT is a technique for translating one natural language $x = \{x_1, \ldots, x_M\}$ into another natural language $y = \{y_1, \ldots, y_N\}$, In autoregressive NMT, each word of the target sentence is generated step by step. Therefore, the generation of the nth target word y_n is influenced by the source sentence x and the part of the target sentence that has already been generated, denoted as $y_{<n}$:

$$P(y|x) = \prod_{n=1}^{N} P(y_n|y_{<n}, x) \tag{1}$$

where x is the source sentence, y is the target sentence, y_n is the generation of the nth target word, and $y_{<n}$ is the $n-1$ target words that have already been generated. The model generates an output word based on the source sentence x and the partial translation $y_{<n}$.

The output values of numerous attention heads are combined to produce MHA. The following is the attention mechanism formula:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, and V are three linearly transformed matrices, and d_k is the vector dimension of Q, K. First, calculate the degree of similarity between each *query* and each *key*. Second, $softmax$ is employed to derive the weights. Third, the weights are then averaged with the corresponding $value$.

Finally, the output values of MHA are then obtained by concatenating the output values of each layer's attention heads:

$$MultiHead(Q, K, V) = Concat(h_1, \ldots, h_i, \ldots, h_n)W^O \tag{3}$$

where h_i denotes the output value of the ith attention head in the MHA. W^O is a linear transformation matrix used to transform the concatenated output of all attention heads.

In machine translation, masking is a crucial technique that helps the model learn and predict data more effectively. The application of this technique markedly boosts the model's efficiency and accuracy, rendering it an essential component of contemporary NMT systems. Masking enables the model to better understand and process sentences of varying lengths and to utilize contextual information more effectively, thereby enhancing the quality of translation. It is worth mentioning that in our research, we investigated the importance of attention heads by employing masking to obscure them.

In this paper, we attribute the importance of attention heads to the effect on translation performance. We modified the attention formula to implement the masking operation of the attention heads:

$$h = \varphi Attention(Q, K, V) \tag{4}$$

where φ is a discrete value, which can be 0 or 1. When φ is 0, it indicates that the attention head is masked; when φ is 1, it means that the attention head is retained in the complete model.

Our experiments were implemented in the Transformer, which has three different attention mechanisms: Enc–Enc, Dec–Dec, and Enc–Dec. The impact of each attention head on translation performance serves as the attention heads' importance. We masked each attention head in each attention mechanism in turn [31]. We used the heatmap to display the information regarding the importance of attention heads that we obtained. It provides an intuitive representation of the distribution of attention head importance. We used the Matplotlib library to implement the drawing of the heatmap. We used the change in BLEU score values due to masking each attention head as an input and set the parameter minimum value to -1.0.

Figure 1 illustrates the heatmap of the attention heads' importance in the Transformer, which reflects the impact of each attention head on the translation performance. In contrast to the findings of Michel [7], not all attention heads are important; some are and some are not.

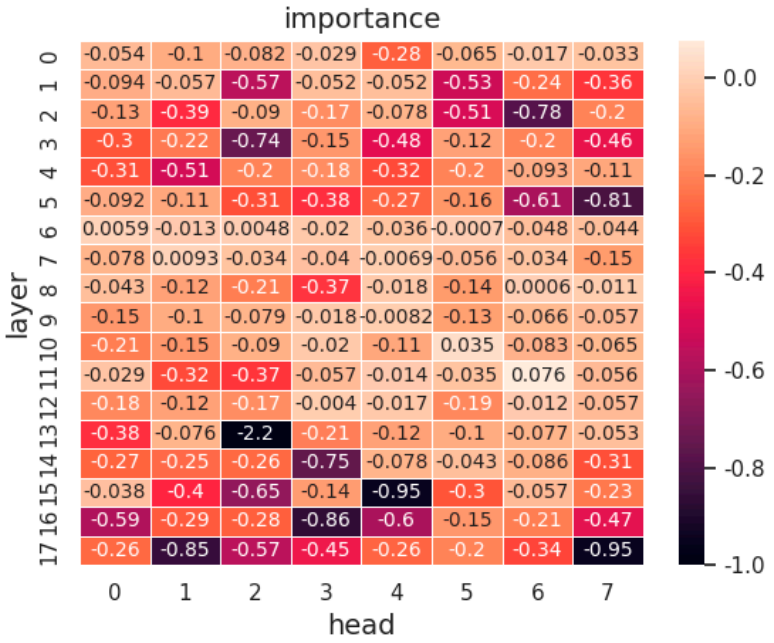


Figure 1. Attention heads' importance. The heatmap reflects the impact of each attention head in the model on translation performance. The x-axis from zero to seven represents the eight attention heads in each layer of the model. The y-axis represents the different layers of the three attention mechanisms in the model. Specifically, 0–5 represent the 1–6 layers of encoder self-attention, 6–11 represent the 1–6 layers of decoder self-attention, and 12–17 represent the 1–6 layers of encoder–decoder attention. The numbers represent the impact of each attention head on translation performance. A higher value indicates that the attention head is less important, while a lower value indicates that the attention head is more important. The color reflects the importance of the attention head, with the darker one indicating more importance.

3.3. Attention Head Analysis

We mask several important attention heads, unimportant attention heads, and random attention heads based on the attention heads' importance we acquired to show that it is accurate. By looking at the degradation in translation performance, we evaluate the validity of attention heads' importance.

Figure 2 illustrates the effect of masking multiple attention heads on translation performance. There is a significant degradation in translation performance when masking several important attention heads, while masking multiple unimportant attention heads results in minimal impact. This outcome validates the accuracy of our attention head importance ranking.

To ensure the reliability of our experimental results, we assessed the distribution of attention head importance and its correlation with the contribution of attention heads to POS.

Wang [12] posits that certain components of the Transformer model are more important than others, particularly within the attention layer. They suggest that the decoder self-attention layers are the least critical, whereas the upper encoder-attention layers in the decoder are more significant than the lower ones. We conducted two experiments to investigate the distribution of attention head importance and verify if it aligns with these suggested properties. First, we want to know which layers in the model are more important and which are less important. This can help us understand the importance of the attention layers. Therefore, the initial experiment aims to assess the importance of the attention heads in each layer. Specifically, the translation performance is evaluated as a function that aggregates the contributions of all attention heads within a layer. If this function demonstrates a significant influence, the layer is deemed crucial. Otherwise, it is considered non-essential. Then, we want to know whether the important attention heads are concentrated in a particular layer and whether retaining only a few attention

heads in a certain layer can allow the model to function normally. Therefore, the second experiment examines the distribution of important and unimportant attention heads. We specifically count the 30 most important and the 30 least important attention heads in the model, based on their importance. A layer is considered important if it contains a high number of crucial attention heads and a low number of non-essential ones. Conversely, it is deemed unimportant if this is not the case.

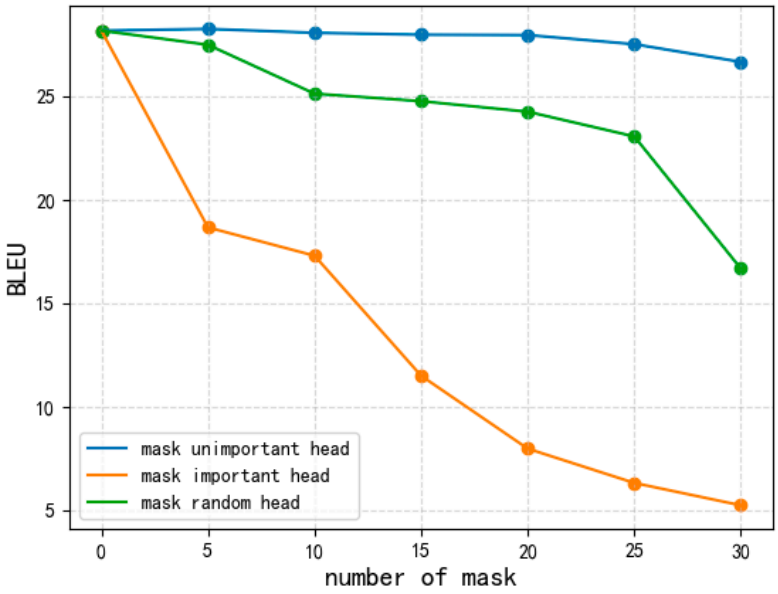


Figure 2. Evolution of BLEU score when heads are pruned. Respectively masking important head (yellow), unimportant head (blue), and random head (green).

Figure 3 represents the importance of attention layers, which is the sum of the impacts of all attention heads in each layer of the three types of attention mechanisms in the model on translation performance. We find that the encoder self-attention (0–5) and encoder–decoder attention (12–17) are more important, while the decoder self-attention (6–11) is less important. Overall, we observe that higher layers are more crucial than lower layers across different attention layers. Particularly, in encoder–decoder attention, we find that the higher layers (15, 16, 17) have a more significant impact on translation performance, indicating that higher layers in encoder–decoder attention are more important than lower layers. Figure 4 represents the distribution of important and unimportant attention heads in the model. We observe that important attention heads are primarily concentrated in encoder self-attention (0–5 in Figure 4a) and encoder–decoder attention (12–17 in Figure 4a), while unimportant attention heads are mainly found in decoder self-attention (6–11 in Figure 4b). This finding indicates that encoder self-attention and encoder–decoder attention are more crucial to the model than decoder self-attention. Moreover, from Figure 4a's 12–17, we can see that important attention heads are more concentrated in higher layers than in lower layers, implying that higher layers are more important than lower layers in the encoder–decoder attention mechanism. According to the results of two experiments, Figures 3 and 4, we find that the encoder self-attention layers (0–5) and the encoder–decoder attention layers (12–17) are important, while the decoder self-attention layers (6–11) are unimportant. And in the encoder–decoder attention layer (12–17) the higher layers are more important than the lower layers. These results support Wang's [8] findings, confirming the validity of the attention heads' importance we obtained.

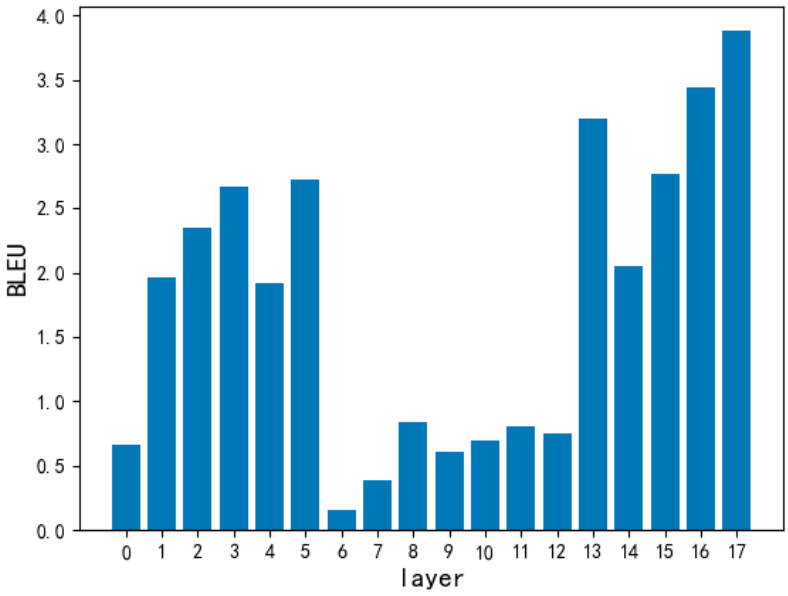


Figure 3. The importance of each layer's attention head. X-axis denotes the attention layer. Y-axis denotes the impact on the BLEU score (absolute value).

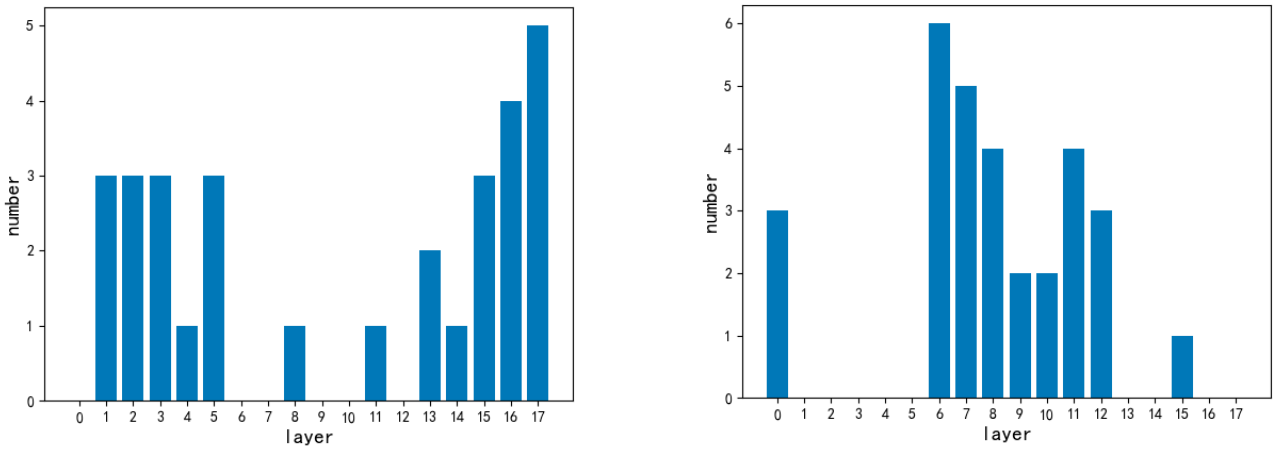


Figure 4. The distribution of attention head importance. (**a**) The number of important attention heads. (**b**) The number of unimportant attention heads.

3.4. Correlation Analysis

Previous research has identified that important attention heads frequently possess specific and interpretable functions [18]. To verify whether the attention heads' importance is consistent with the interpretable function, we analyzed the correlation between attention head importance and attention head POS contribution. The attention head contributes to the accurate translation of each POS word, which is the attention head POS contribution. Specifically, we evaluate the impact on the translation accuracy of each POS word when masking the attention head. For more details see Section 4.1. We employed various statistical methods to evaluate these correlations, including Pearson correlation and KL divergence. The Pearson correlation measures the linear correlation between two variables and the KL divergence is an asymmetric measure of the difference between two probability distributions. The value of the Pearson correlation is easy to understand and directly reflects the strength of the linear relationship between variables. The KL divergence can sensitively capture subtle differences between two probability distributions and is widely used in machine learning. Since the data contain negative values,

we must normalize it to [0, 1] before computing the KL divergence. We use two normalization methods. The first is $KL_{Norm} = (x_i - x_{min})/(x_{max} - x_{min})$. The second method takes the logarithm of the data $a = [x_1, \ldots, x_i, \ldots, x_n]$ in the interval $(-\infty, +\infty)$ to obtain the data $b = [\ln x_1, \ldots, \ln x_i, \ldots, \ln x_n]$ in the interval $(0, +\infty)$, then normalizes the data $KL_{Softmax} = x_i/sum(x)$.

The correlation between attention head importance and attention head POS contribution among three distinct relevance evaluation criteria is shown in Table 3. We have established a connection between human-understandable POS information and the importance of attention heads, significantly enhancing the model's interpretability and aiding in our understanding of why some attention heads are important. Moreover, the experiment employs statistical analysis of a large number of sentences to ensure the accuracy and validity of the results. Overall, there was a discernible positive correlation between attention head importance and their contributions to various POS categories, indicating that the important attention heads we identified possess specific and interpretable functions. This aligns with the research findings of Voita [18]. A more substantial correlation was observed between attention head importance and the contributions to nouns and verbs. This could be attributed to the pivotal role of nouns and verbs in Chinese–English machine translation, suggesting that important attention heads have a deeper understanding of these POS. Such insights motivate us to refine the model design to facilitate enhanced learning of nouns and verbs by the attention heads, potentially improving the model's overall performance.

Table 3. The correlation between attention head importance and attention head POS contribution.

	Pearson	KL_{Norm}	$KL_{Softmax}$
None	0.85586148	0.00288859	0.01695689
Verb	0.81776615	0.00448712	0.01009835
Adjective	0.6634828	0.12858195	0.05381574
Adverb	0.59999701	0.12642623	0.07041694
Preposition	0.61585583	0.10229620	0.12294885
Determiner	0.36217017	0.11767329	0.34239231

4. Linguistic Knowledge

In this section, we analyze how the attention head learns linguistic information at various granularities, which better reveals how NMT makes decisions. Through various linguistic knowledge granularities—POS, dependency relations, and syntax trees—we explore how NMT models process word information, inter-word relations, multi-word information, and syntactic structure information. We strive to elucidate model decisions that are beyond human comprehension by leveraging linguistic knowledge that is understandable to humans [32].

We masked several important and unimportant attention heads according to the attention heads' importance to identify the various linguistic knowledge that these attention heads had acquired. Three steps comprised our investigations. In the first step, the reference translation was annotated by StanfordCoreNLP 4.5.4. StanfordCoreNLP is an open-source natural language processing toolkit that integrates various NLP tools, enabling comprehensive linguistic analysis. Therefore, we chose it for our experiments. The accuracy of the reference translation and the translation generated from the complete model are compared in the second step. In the third step, we compare the accuracy of the translation generated by the pruning model and the reference translation. We believe that when the important attention heads are masked, the performance of the model is not as good as the complete model, leading to a decrease in the accurate translation rates of various types of linguistic information and the occurrence of negative accuracy. This experiment helps us explore which key linguistic information the model has learned and which pieces of linguistic information are more important. On the other hand, when the less important attention heads are masked, it may have a positive effect on some linguistic information,

providing new insights for model pruning. The following is a detailed explanation of specific linguistic knowledge.

4.1. Part-of-Speech

In this experiment, we masked several important attention heads and unimportant attention heads according to attention heads' importance to identify different POS knowledge learned by these attention heads. Specifically, we evaluate the POS information learned by the attention head in the translation accuracy of each POS in the translations generated by the masking model.

In the first step, according to the POS types in Penn Treebank [33], we used Stanford-CoreNLP to POS annotate the reference translation. Penn Treebank is a large-scale corpus that provides precise POS and syntactic annotations, which are extremely useful for the training and evaluation of language models. It is widely referenced in natural language processing research and serves as a benchmark dataset for numerous algorithms and models. Therefore, we chose it for our experiments. POS information represented the function of the individual word information in a sentence for NMT. For this study, we focused on the six POS types that are most frequently used: the noun (N), verb (V), adjective (ADJ), adverb (ADV), preposition or subordinating conjunction (IN), and determiner (DT). Since various POS are divided into POS subclasses, for example, general adjectives, adjectival comparatives, and adjectival maxims are three categories of adjectives, we cluster nouns, verbs, adjectives, and adverbs individually [34]. We statistically counted the number of words of each POS type in the reference translation, as shown in Table 4, and the total number of words of POS type p in the reference translation was indicated as $total_{p,ref}$.

Table 4. Number of words in each POS category, number of combinations of dependent words, and number of syntactic tree path patterns in the reference translation.

	Types	Number
Part-of-speech	None	1209801
	Verb	639937
	Adjective	325112
	Adverb	168872
	Preposition	521772
	Determiner	421079
Dependency	Root	253109
	Nsubj	300879
	Obj	200308
	Advmod	183645
	Det	401812
	Amod	285112
	Nmod	229550
	Compound	242266
Syntax tree	NP-NP-NN	185474
	NP-PP-IN	139905
	PP-NP-NN	148864
	NP-NP-DT	126033
	VP-PP-IN	131974

In the second step, we compare the number of correct translations of each POS word in the translation generated by the complete model with the reference translation. The total number of correctly translated words of POS type p in the translations generated by the complete model is as follows:

$$N_{p,model} = \Sigma_{j=1}^{s}\Sigma_{i=1}^{m} min\left(n_{w_i,ref}, n_{w_i,model}\right) \quad (5)$$

where s denotes the number of sentences in the dataset and m denotes a non-repeating word of POS p in a sentence. For the ith non-repeating word w_i of POS p in the jth sentence of the reference translation, the number in the reference sentence $n_{w_i,ref}$ and the number in the complete model-generated sentence $n_{w_i,model}$ are counted.

In the third step, we use the masking model for translation and count the total number of correctly translated words of POS type p in the translation generated by the masking model as

$$N_{p,head} = \Sigma_{j=1}^{s}\Sigma_{i=1}^{m} min\left(n_{w_i,ref}, n_{w_i,mask_{head}}\right) \quad (6)$$

In the end, we obtain the insights of the attention head learning POS information:

$$acc_{p,model} = \frac{N_{p,model}}{total_{p,ref}} \quad (7)$$

$$acc_{p,head} = \frac{N_{p,head}}{total_{p,ref}} \quad (8)$$

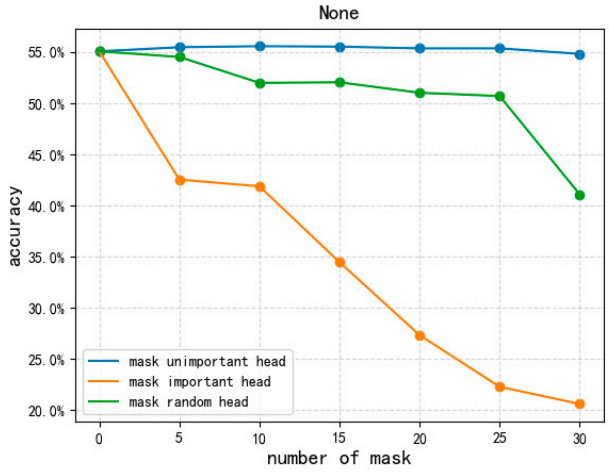
These formulas represent the accuracy of the complete model and the model with masked attention heads in correctly translating words of each POS, relative to the reference translation, where $acc_{p,model}$ represents the accuracy of the complete model in correctly translating words of POS p, serving as the baseline for POS experiment. $acc_{p,head}$ represents the accuracy of the model with masked attention heads in correctly translating words of POS p.

Figure 5 displays the experimental results, which show how masking important attention heads, unimportant attention heads, and random attention heads affects various POS information. Overall, for each POS, the accuracy of the words' correct translations decreased when important heads were masked, demonstrating that important attention heads have the propensity to engage in a lot of information about the POS. The accuracy of the words' correct translations of each POS remained stable when masking the unimportant attention heads. When randomly masking attention heads, the change in the correct translation accuracy for words of different parts of speech falls between the accuracy changes caused by masking important and unimportant attention heads, a phenomenon that aligns with general patterns.

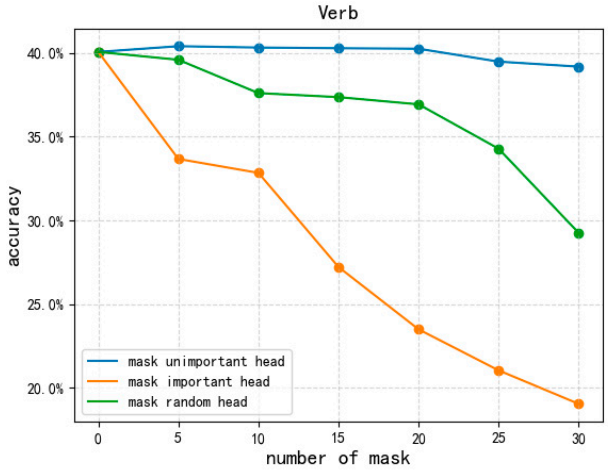
To better understand the findings of our experiment, we selected an example from the dataset we used. As shown in Table 5, we display the reference translation, the translation generated by the complete model, the translation generated when important attention heads are masked, and the translation generated when unimportant attention heads are masked. It can be observed that when important attention heads are masked, words of various POS, such as the verb "provides", the adjective "theoretical", and the noun "checking", are not translated correctly. In addition, when unimportant attention heads are masked, there is almost no change in the translation results.

Table 5. The example of reference translation, the translation generated by the complete model, the translation generated when important attention heads are masked, and the translation generated when unimportant attention heads are masked.

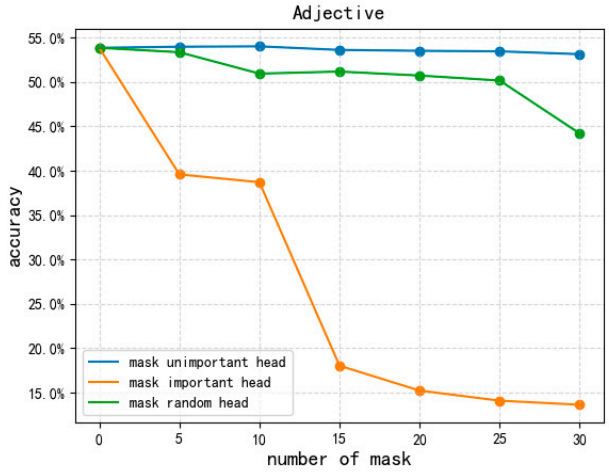
Reference	The analysis method provides some theoretical foundation for the safe assessment on welded joint strength in a rotation blade.
Transformer	This method provides a theoretical basis for the strength-checking of rotary welded joints.
Mask Important Attention heads	The analysis method is used to verify the welding strength of the joints.
Mask Unimportant Attention heads	This method provides a theoretical basis for the strength-checking of the rotary joint.



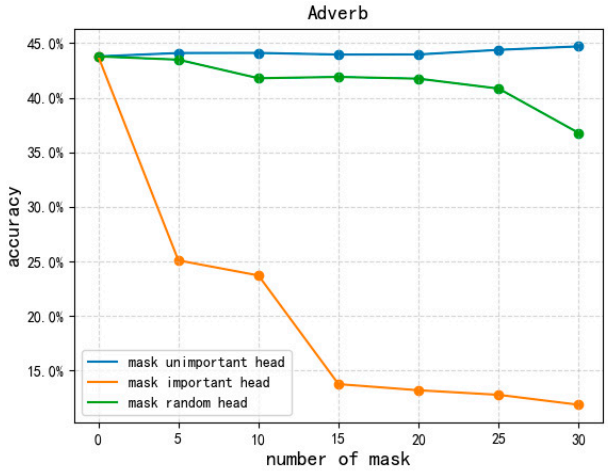
(**a**)



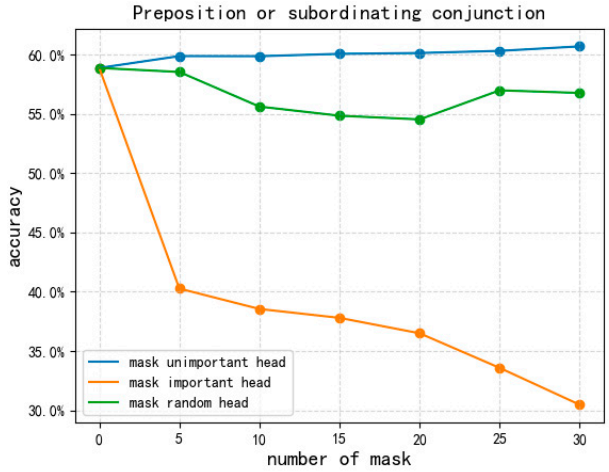
(**b**)



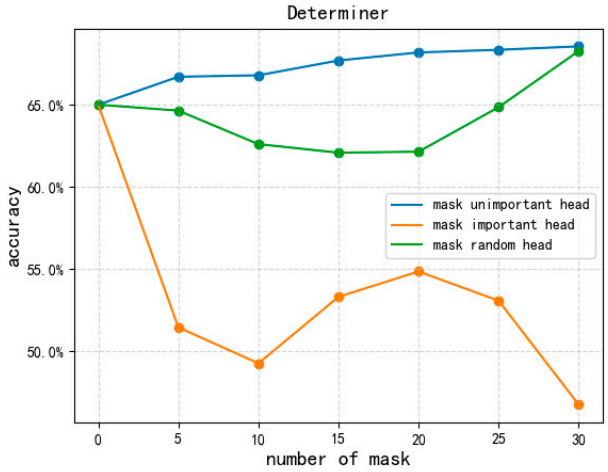
(**c**)



(**d**)



(**e**)



(**f**)

Figure 5. The effect of masking attention heads on each piece of POS information. *x*-axis denotes the number of masked attention heads. *y*-axis denotes the accuracy of correct translation. (**a**–**f**) denote different types of POS.

We discovered that the decreasing trend of the words' correct translations was different for each type of POS information while masking important attention heads. We examine the alteration in the accuracy of the words' correct translations when masking the important attention head to better understand how the model learns various POS information, as shown in Figure 6. It can be seen that all POS information is important for model decisions. However, due to the different baseline values for each part of speech, we cannot intuitively determine which part of speech experiences the greatest decrease in correct translation accuracy. Therefore, we have presented in table format the correct translation accuracies of words for each part of speech generated by the complete model, as well as the decrease in these accuracies when important attention heads are masked, as shown in Table 6. This method allows for an effective observation of the changes in correct translation accuracy when multiple attention heads are masked [14]. We observed that the attention head learned more about adjectives and nouns when the important attention head was masked. This is presumably due to the reason that nouns and adjectives play a more significant part in sentence structure in English. When masking the most important attention heads (e.g., five), we found that the accuracy of prepositions reduced the fastest, indicating that preposition-related knowledge primarily exists in the most important attention heads.

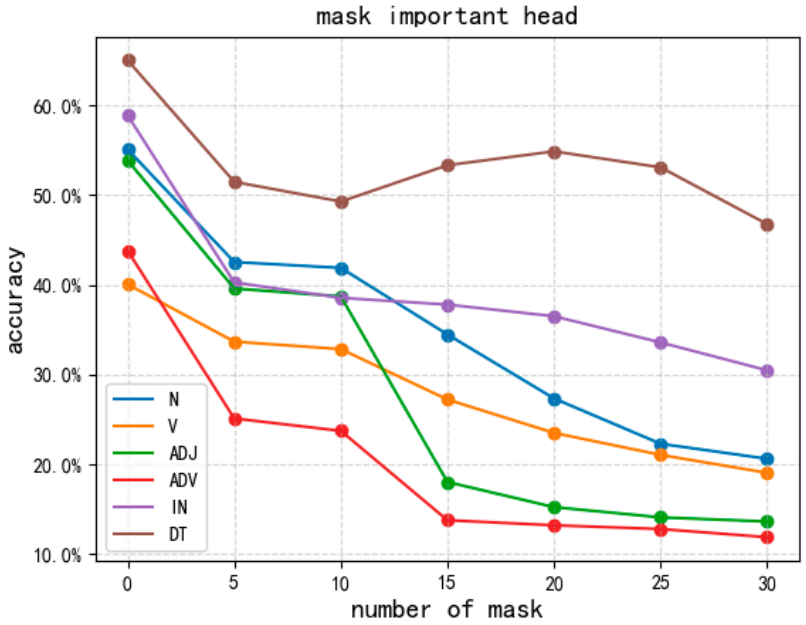


Figure 6. Masking important attention heads on each piece of POS information.

Table 6. The changes in the accuracy of correct translations for words of each POS between the complete model and the pruned model. All data in the table are represented as percentages. The leftmost column lists the POS. The "Complete Model" column shows each part's translation accuracy, while the "Pruned Model" column, labeled "5–30", displays the percentage decrease in accuracy after masking 5 to 30 attention heads.

	Complete Model	**Pruned Model**					
		5	**10**	**15**	**20**	**25**	**30**
N	55.07	12.54	13.19	20.61	27.74	32.79	34.46
V	40.05	6.39	7.22	12.84	16.56	19.01	21.01
ADJ	53.85	14.27	15.15	35.83	38.63	39.77	40.23
ADV	43.79	18.69	20.07	30.03	30.59	31.01	31.92
IN	58.87	18.61	20.33	21.08	22.38	25.30	28.41
DT	65.01	13.55	15.74	11.69	10.15	11.93	18.26

4.2. Dependency

In this section, we explored the knowledge of various dependencies learned by the attention heads by masking several important attention heads and unimportant attention heads to better understand the information about inter-word relationships learned by each attention head in the model, as in the experiments on POS.

In the first step, we used stanfordCoreNLP to extract dependencies from reference translations, and this tool can extract all dependencies of a sentence. Dependency is linguistic knowledge that reveals the dominant relationship between two words. A sentence has a root that does not depend on other words, all other words depend on one word of the sentence, and no one word can depend on two or more words. For this study, we focused on the eight dependency types that are most frequently used: root, nsubj, obj, advmod, det, amod, nmod, and compound. We statistically counted the number of word combinations of each dependency type in the reference translation, as shown in Table 4, and the total number of word combinations of dependency type dep in the reference translation was indicated as $total_{dep,ref}$.

In the second step, we extracted the dependencies from the translations generated by the complete model. The number of correct translations for the word combination of each dependency in the translation is counted as $N_{dep,model}$, and the formula is as follows:

$$N_{dep,model} = \Sigma_{j=1}^{s}\Sigma_{i=1}^{m}\varphi\left(DEP_{ref,j}, DEP_{model,j,i}\right) \tag{9}$$

where s denotes the number of sentences in the dataset and m denotes a word combination of dependency dep in a sentence. $\varphi()$ is an indicator function, $DEP_{ref,j}$ denotes all word combinations where dependency is dep in the jth sentence of the reference translation, and $DEP_{model,j,i}$ denotes the ith word combinations where dependency is dep in the jth sentence of the translation generated by the complete model. If the word combination of $DEP_{model,j,i}$ exists in $DEP_{ref,j}$, the word combination is considered to be translated correctly and the output of the indicator function is 1, otherwise the output is 0.

In the third step, we extracted the dependencies from the translations generated by the masking model and calculated the total number of correctly translated word combinations whose dependencies are dep as

$$N_{dep,model} = \Sigma_{j=1}^{s}\Sigma_{i=1}^{m}\varphi\left(DEP_{ref,j}, DEP_{mask_{head},j,i}\right) \tag{10}$$

In the end, we obtained the insights of attention head learning dependency information:

$$acc_{dep,model} = \frac{N_{dep,model}}{total_{dep,ref}} \tag{11}$$

$$acc_{dep,head} = \frac{N_{dep,head}}{total_{dep,ref}} \tag{12}$$

These formulas represent the accuracy of the complete model and the model with masked attention heads in correctly translating word combinations for each dependency, relative to the reference translation, where $acc_{dep,model}$ represents the accuracy of the complete model in correctly translating word combinations for dependency dep, serving as the baseline for dependency experiment. $acc_{dep,head}$ represents the accuracy of the model with masked attention heads in correctly translating word combinations for dependency dep.

The effect of masking the attention heads on the accuracy of word combinations that correspond to each dependency in the translation reveals the attention head's learning of knowledge of various dependencies. The experimental results are shown in Figure 7. When the important attention heads were masked, the accuracy of the correct translation of each dependent word combination decreased, demonstrating that the attention head had become acquainted with the dependence. Similarly, we analyzed the variation in the accuracy of each dependency when masking important attention heads, as shown in Figure 8 and

Table 7. We discovered that attention heads learned more inter-word information about adjectival modifiers (amod). This supports the findings of POS experiments, which suggest that adjectives are essential to English sentence construction.

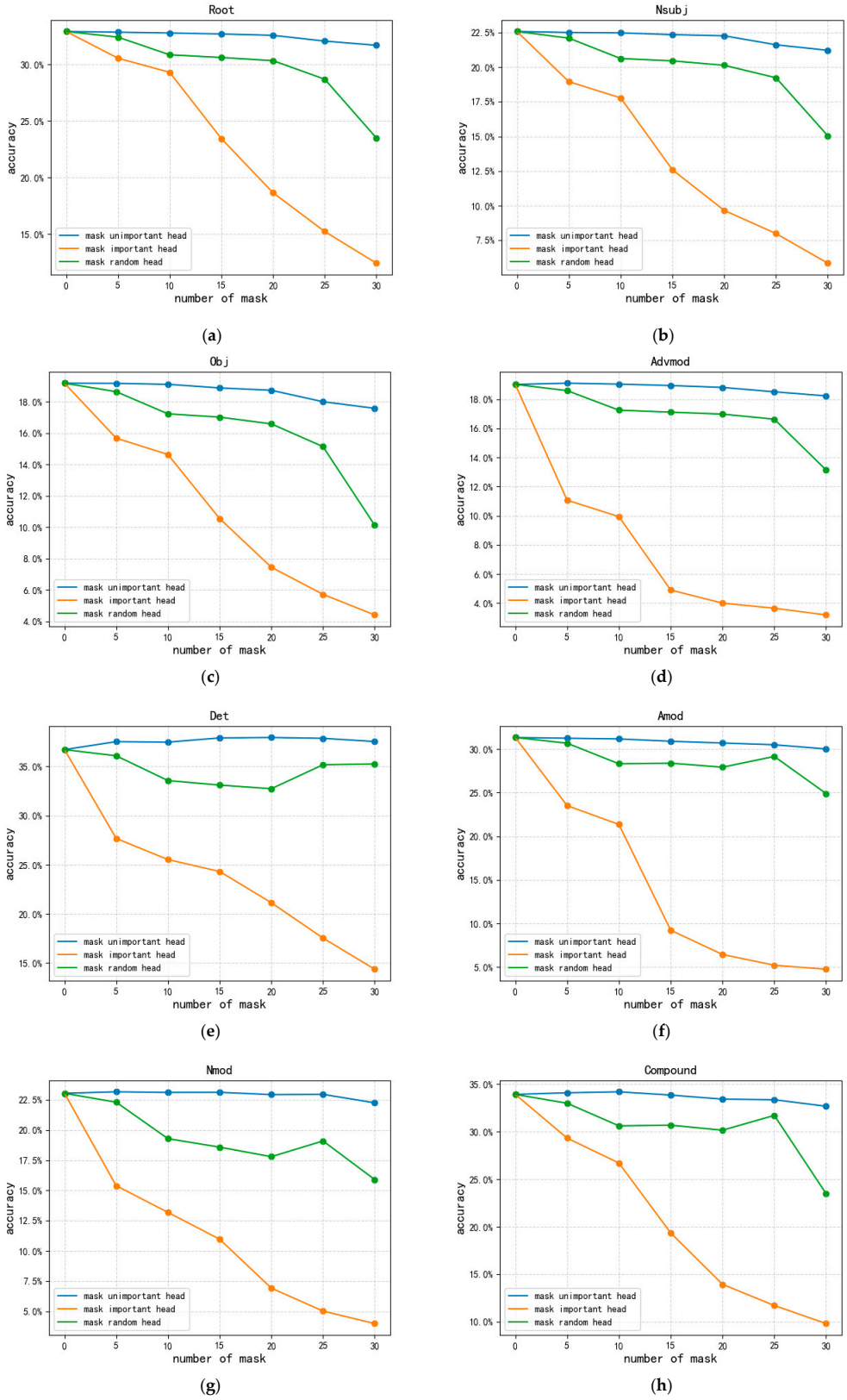


Figure 7. The effect of masking attention heads on each dependency information. (**a**–**h**) denote different dependency information.

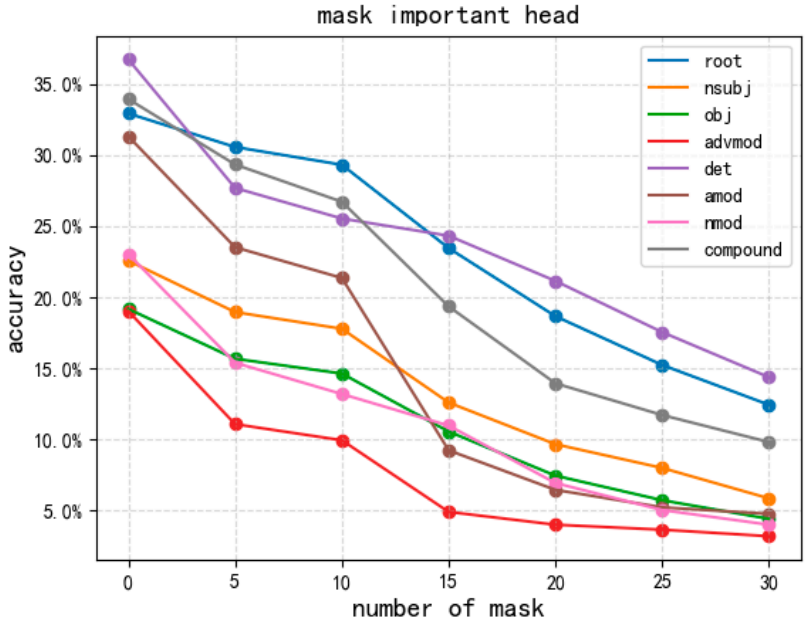


Figure 8. Masking important attention heads on each piece of dependency information.

Table 7. The changes in the accuracy of correct translations for word combinations for each dependency between the complete model and the pruned model. All data in the table are represented as percentages. The leftmost column lists the dependency. The "Complete Model" column shows each part's translation accuracy, while the "Pruned Model" column, labeled "5–30," displays the percentage decrease in accuracy after masking 5 to 30 attention heads.

	Complete Model	**Pruned Model**					
		5	**10**	**15**	**20**	**25**	**30**
root	32.92	2.36	3.61	9.47	14.26	17.69	20.49
nsubj	22.57	3.62	4.79	9.98	12.91	14.58	16.72
obj	19.17	3.50	4.54	8.62	11.72	13.45	14.76
advmod	19.02	7.95	9.08	14.12	15.02	15.37	15.83
det	36.72	9.04	11.19	12.40	15.58	19.18	22.35
amod	31.31	7.80	9.96	22.08	24.86	26.09	26.55
nmod	23.03	7.64	9.84	12.06	16.10	18.00	19.03
compound	33.93	4.60	7.22	14.58	19.99	22.22	24.12

4.3. Syntax Tree

In this section, we studied the change in the accuracy of syntactic tree information in the translation while masking several important attention heads and unimportant attention heads to evaluate the impact of attention heads on the model's processing of multiple-word information and overall syntactic structure [35]. A syntactic tree is an abstract syntactic structure that represents the structure of a sentence.

In the first step, we constructed the syntactic analysis of the reference translation using stanfordCoreNLP, a tool that constructs syntactic trees of sentences. We summarize the patterns of the bottom three nodes in the syntactic tree paths of each word to better examine the syntactic structure knowledge acquired by the attention head because there are plenty of ways to combine syntactic tree paths. For this study, we focused on the five syntactic tree path patterns that are most frequently used: NP-NP-NN, NP-PP-IN, PP-NP-NN, NP-NP-DT, and VP-PP-IN. We statistically counted the number of words of each syntactic tree path pattern in the reference translation, as shown in Table 4, and the total number of words of each syntactic tree path pattern *path* in the reference translation was indicated as $total_{path,ref}$.

In the second step, we constructed syntactic trees for the translations generated by the complete model. We counted the total number of words in the translation generated by the complete model and the reference translation whose word paths have the same syntactic tree path pattern, denoted as $N_{path,model}$, with the following formula:

$$N_{path,model} = \Sigma_{j=1}^{s}\Sigma_{i=1}^{m}\varphi\left(PATH_{ref,j}, PATH_{model,j,i}\right) \quad (13)$$

where $PATH_{ref,j}$ denotes the word where the syntactic tree path pattern is *path* in the *j*th sentence of the reference translation, and $PATH_{model,j,i}$ denotes the *i*th word where the syntactic tree path pattern is *path* in the *j*th sentence of the translation generated by the complete model. The other parameters were consistent with the experimental settings of dependency.

In the third step, we constructed syntactic trees for the translations generated by the masking model and calculated that the total number of correctly translated words with the syntactic tree path pattern is *path*:

$$N_{path,model} = \Sigma_{j=1}^{s}\Sigma_{i=1}^{m}\varphi\left(PATH_{ref,j}, PATH_{mask_{head},j,i}\right) \quad (14)$$

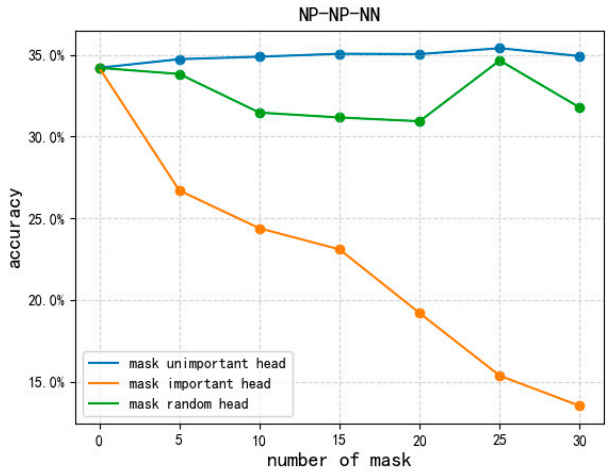
In the end, we obtained the insights of attention head learning syntactic tree information:

$$acc_{path,model} = \frac{N_{path,model}}{total_{path,ref}} \quad (15)$$

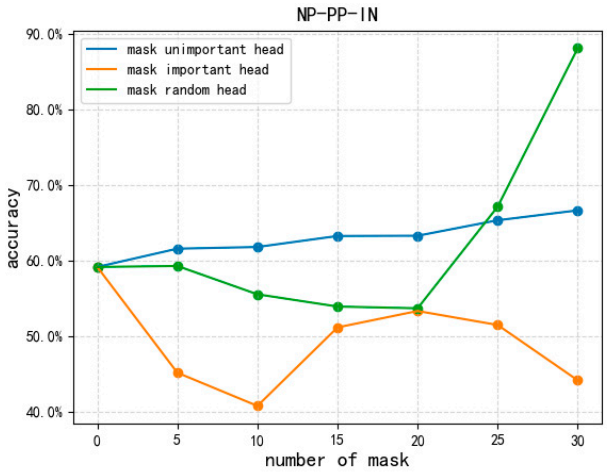
$$acc_{path,head} = \frac{N_{path,head}}{total_{path,ref}} \quad (16)$$

These formulas represent the accuracy of the complete model and the model with masked attention heads in correctly translating words for each syntactic tree path pattern relative to the reference translation, where $acc_{path,model}$ represents the accuracy of the complete model in correctly translating words for the syntactic tree path pattern *path*, serving as the baseline for the dependency experiment. $acc_{path,head}$ represents the accuracy of the model with masked attention heads in correctly translating words for the syntactic tree path pattern *path*.

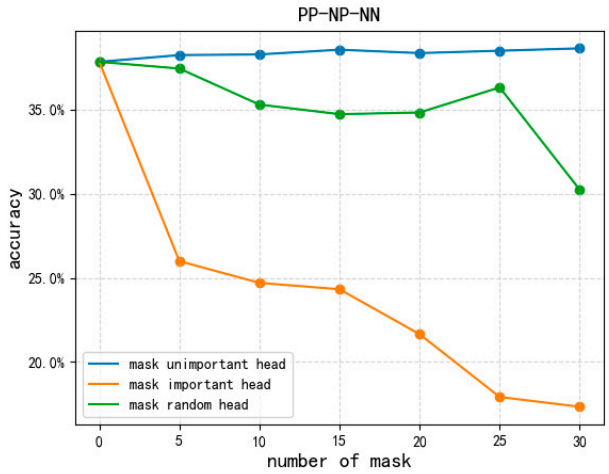
Figure 9 displays the variety of attentional head learning syntactic information that we discovered. The attention head commonly learns knowledge related to syntax tree information. Meanwhile, we find that some syntactic tree path patterns benefit when some attention heads are masked. Below, we examine the situation by masking the important attention heads on each syntactic tree path pattern. As shown in Figure 10 and Table 8, when the most important heads (e.g., 5, for example) are masked, the accuracy of all syntactic tree path patterns rapidly decreases. Additionally, when more attention heads are masked, the accuracy of some syntactic tree path patterns (e.g., NP-PP-IN and NP-NP-DT) increases before decreasing again. We hypothesize that the concentration of syntactic knowledge in a few crucial attention heads during the learning of syntactic information might be the reason for this observation. This idea provides a direction for future research.



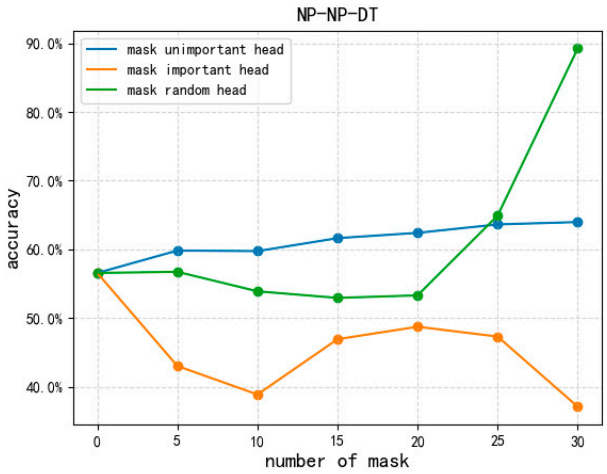
(**a**)



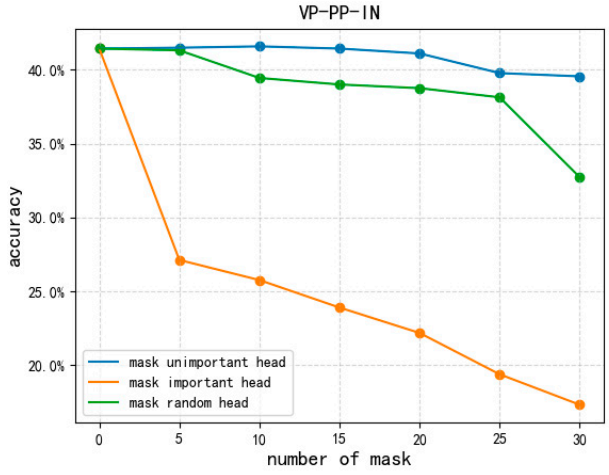
(**b**)



(**c**)



(**d**)



(**e**)

Figure 9. The effect of masking attention heads on each syntactic tree information. (**a**–**e**) denote different syntactic tree information.

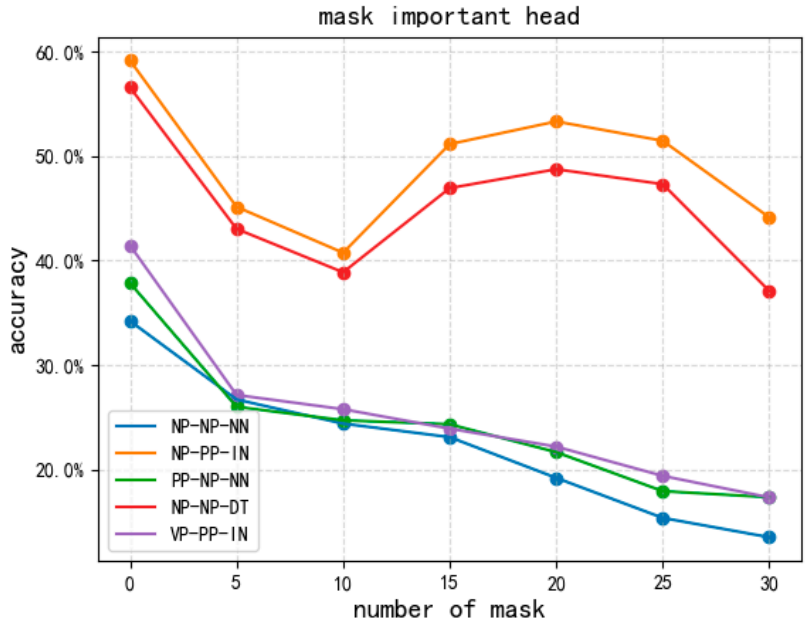


Figure 10. Masking important attention heads on each piece of syntactic tree information.

Table 8. The changes in the accuracy of correct translations for words for each syntactic tree path pattern between the complete model and the pruned model. All data in the table are represented as percentages. The leftmost column lists the syntactic tree path pattern. The "Complete Model" column shows each part's translation accuracy, while the "Pruned Model" column, labeled "5–30", displays the percentage decrease in accuracy after masking 5 to 30 attention heads.

	Complete Model	**Pruned Model**					
		5	**10**	**15**	**20**	**25**	**30**
NP-NP-NN	34.20	7.51	9.82	11.11	15.01	18.84	20.69
NP-PP-IN	59.12	14.00	18.39	8.00	5.83	7.68	15.01
PP-NP-NN	37.83	11.84	13.13	13.51	16.18	19.91	20.49
NP-NP-DT	56.55	13.54	17.69	9.64	7.83	9.26	19.48
VP-PP-IN	41.43	14.31	15.67	17.53	19.25	22.05	24.11

5. Conclusions

In this paper, we explore the importance of attention heads in understanding NMT. We confirmed the correctness of the importance of the attention head and investigated its distribution. The correlation between attention head importance and their contribution to POS was assessed using three evaluative criteria. Our investigation covered how attention heads acquire language knowledge at three different granularities: POS, dependency relationships, and syntactic trees, enhancing our understanding of their learning processes.

We find that not all attention heads are important, and the important attention heads are more distributed in the encoder self-attention layer and the encoder–decoder attention layer. These findings have propelled us towards a deeper understanding of the internal structure of NMT models and have opened up new research directions for subsequent related studies. Specifically, these discoveries can guide us in conducting in-depth analyses of critical components within the model to enhance its performance. Additionally, they also support pruning non-critical parts of the model to reduce parameters and improve operational efficiency. In addition, the important attention heads in Chinese–English machine translation contribute more to the generation of nouns and verbs in their translations. Finally, we discovered that the attention heads had varying degrees of POS, dependence, and syntax tree learning. These findings offer valuable guidance for future model design. For instance, when designing Chinese–English machine translation models, we should

pay closer attention to the processing of linguistic elements such as nouns and adjectives. This ensures that the attention mechanism functions more effectively, thereby enhancing translation quality. Overall, we investigated the multigranularity of linguistic knowledge within the context of attention heads making decisions.

For future work, we plan to explore the role of attention heads in learning other linguistic knowledge, such as morphology. Additionally, we will conduct comparisons using other salience methods, including layer-wise relevance propagation (LRP) and gradient-based approaches, to examine the decisions made by important attention heads from different perspectives.

Author Contributions: Conceptualization, J.Z.; methodology, Z.Z. and J.Z.; validation, Z.Z.; resources, J.Z.; writing—original draft, Z.Z. and J.Z.; writing—review and editing, J.Z. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: Supported by the National Natural Science Foundation of China (grant No. 62166022; 62066022) and the General Project of Yunnan Fundamental Research Programs (grant No. 202101AT07 0077; 202301AT070015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
2. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
4. Jack, H.; Ana, M.; Jena, D.H.; Lillian, L.; Jeff, D.; Rowan, Z.; Robert, M.; Yejin, C. Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from the New Yorker Caption Contest. In Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023.
5. Ding, Y.; Liu, Y.; Luan, H.; Sun, M. Visualizing and understanding neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1150–1159.
6. Belinkov, Y.; Durrani, N.; Dalvi, F.; Sajjad, H.; Glass, J. What do neural machine translation models learn about morphology? *arXiv* **2017**, arXiv:1704.03471.
7. Jing, L.; Yong, Z. An Algorithm for Finding Optimal k-Core in Attribute Networks. *Appl. Sci.* **2024**, *14*, 1256. [CrossRef]
8. Michel, P.; Levy, O.; Neubig, G. Are sixteen heads really better than one? *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14037–14047.
9. Nikita, M.; Tom, S.; Mark, S.; Alexandra, B. Extrinsic Evaluation of Machine Translation Metrics. In Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023.
10. Lipton, Z. The Mythos of Model Interpretability. *Commun. ACM* **2016**, *61*, 36–43. [CrossRef]
11. Wu, W.; Jiang, C.; Jiang, Y.; Xie, P.; Tu, K. Do PLMs Know and Understand Ontological Knowledge? In Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023.
12. Wang, W.; Tu, Z. Rethinking the value of transformer components. *arXiv* **2020**, arXiv:2011.03803.
13. Serrano, S.; Smith, N.A. Is attention interpretable? *arXiv* **2019**, arXiv:1906.03731.
14. Li, X.; Li, G.; Liu, L.; Meng, M.; Shi, S. On the word alignment from neural machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1293–1303.
15. Kobayashi, G.; Kuribayashi, T.; Yokoi, S.; Inui, K. Attention module is not only a weight: Analyzing transformers with vector norms. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020.
16. Jain, S.; Wallace, B.C. Attention is not explanation. *arXiv* **2019**, arXiv:1902.10186.
17. Wiegreffe, S.; Pinter, Y. Attention is not not explanation. *arXiv* **2019**, arXiv:1908.04626.
18. Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv* **2019**, arXiv:1905.09418.

19. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef] [PubMed]
20. Ma, W.; Zhang, K.; Lou, R.; Wang, L.; Vosoughi, S. Contributions of transformer attention heads in multi-and cross-lingual tasks. *arXiv* **2021**, arXiv:2108.08375.
21. Ghader, H.; Monz, C. What does attention in neural machine translation pay attention to? *arXiv* **2017**, arXiv:1710.03348.
22. Chen, Z.; Jiang, C.; Tu, K. Using Interpretation Methods for Model Enhancement. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language, Singapore, 6–10 December 2023.
23. Yin, K.; Neubig, G. Interpreting language models with contrastive explanations. *arXiv* **2022**, arXiv:2202.10419.
24. Belinkov, Y.; Màrquez, L.; Sajjad, H.; Durrani, N.; Dalvi, F.; Glass, J. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *arXiv* **2018**, arXiv:1801.07772.
25. Ekin, A.; Dale, S.; Jacob, A.; Tengyu, M.; Denny, Z. What learning algorithm is in-context learning? Investigations with linear models. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023.
26. He, S.; Tu, Z.; Wang, X. Towards Understanding Neural Machine Translation with Word Importance. 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), Hong Kong, China, 3–7 November 2019. [CrossRef]
27. Qiang, J.; Liu, K.; Li, Y.; Zhu, Y.; Yuan, Y.H.; Hu, X.; Ouyang, X. Chinese Lexical Substitution: Dataset and Method. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language, Singapore, 6–10 December 2023.
28. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
29. Papineni, K.; Roukos, S.; Ward, T. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002. [CrossRef]
30. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1243–1252.
31. Tan, S.; Shen, Y.; Chen, Z.; Courville, A.; Gan, C. Sparse Universal Transformer. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language, Singapore, 6–10 December 2023.
32. Müller, M.; Jiang, Z.; Moryossef, A.; Rios, A.; Ebling, S. Considerations for meaningful sign language machine translation based on glosses. In Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023.
33. Marcus, M.; Santorini, B.; Marcinkiewicz, M.A. Building a large annotated corpus of English: The Penn Treebank. *Comput. Linguist.* **1993**, *19*, 313–330.
34. Kai, V.; Frank, K. Cluster-Centered Visualization Techniques for Fuzzy Clustering Results to Judge Single Clusters. *Appl. Sci.* **2024**, *14*, 1102. [CrossRef]
35. Woosik, L.; Juhwan, L. Tree-Based Modeling for Large-Scale Management in Agriculture: Explaining Organic Matter Content in Soil. *Appl. Sci.* **2024**, *14*, 1811. [CrossRef]