# Do Multilingual Neural Machine Translation Models Contain Language Pair Specific Attention Heads?

Zae Myung Kim, Laurent Besacier, Vassilina Nikoulina, Didier Schwab

# Do Multilingual Neural Machine Translation Models Contain Language Pair Specific Attention Heads?

**Zae Myung Kim[1,2], Laurent Besacier[1], Vassilina Nikoulina[1], Didier Schwab[2]**
[1]NAVER LABS Europe
[2]Univ. Grenoble Alpes, CNRS, LIG
{zae-myung.kim,laurent.besacier,vassilina.nikoulina}@naverlabs.com
didier.schwab@univ-grenoble-alpes.fr

## Abstract

Recent studies on the analysis of the multilingual representations focus on identifying whether there is an emergence of language-independent representations, or whether a multilingual model partitions its weights among different languages. While most of such work has been conducted in a "black-box" manner, this paper aims to analyze individual components of a multilingual neural translation (NMT) model. In particular, we look at the encoder self-attention and encoder-decoder attention heads (in a many-to-one NMT model) that are more specific to the translation of a certain language pair than others by (1) employing metrics that quantify some aspects of the attention weights such as "variance" or "confidence", and (2) systematically ranking the importance of attention heads with respect to translation quality. Experimental results show that surprisingly, the set of most important attention heads are very similar across the language pairs and that it is possible to remove nearly one-third of the less important heads without hurting the translation quality greatly.

## 1 Introduction

Recent work on analyzing the internals of Transformer-based models (Vaswani et al., 2017) sheds some light on how different components within the models affect the final performance (Bogoychev, 2020; Behnke and Heafield, 2020), and are closely related to playing linguistically interpretable roles (Voita et al., 2019; Jo and Myaeng, 2020). Moreover, studies on the analysis of multilingual representations (Conneau et al., 2020b; Dufter and Schütze, 2020; Wang et al., 2020b) focus on identifying whether there is an emergence of language-independent representations in multilingual models, or whether multilingual models partition their weights among different languages.

In this paper, we investigate if similar analysis can be made for pretrained multilingual neural machine translation (NMT) models regarding language pair specificity. More precisely, we analyze multi-head attention in a many-to-one (Transformer-based) NMT model and try to find, through an extensive ablation method on selection of the attention heads, whether some heads are more specific to the translation of a certain language pair than others.

Our contributions are the following: (1) we examine the effectiveness of different attention-based metrics on pruning encoder self-attention and cross attention heads; (2) we find that while it is possible to discover rare heads that are specific to a language pair by using a proposed head selection method, most important heads are language-independent; (3) we also show that around 30% of heads can be removed with very little loss of performance.

## 2 Related Work

Recent studies analyzed the roles of attention heads in the Transformer models either in language modeling (LM) (Michel et al., 2019; Clark et al., 2019; Jo and Myaeng, 2020) or NMT (Voita et al., 2019; Behnke and Heafield, 2020; Michel et al., 2019). It has been shown that a set of attention heads might be redundant at inference and can be pruned with almost no loss in performance. In addition, some studies (Voita et al., 2019; Clark et al., 2019) suggested a linguistic interpretation of self-attention heads. However, most of these analyses were carried out for a single language (in case of LM) or a single language pair (in case of NMT).

In the meantime, efficiency in the cross-lingual transfer of recently released pretrained multilingual language models (Devlin et al., 2019; Conneau et al., 2020a) has boosted an active line of research trying to analyze their representations to

understand what favors the emergence of an *interlingua*. For instance, Pires et al. (2019); Dufter and Schütze (2020); Karthikeyan et al. (2020) tried to decouple the effect of shared "anchors"[1] from the rest of the model. Very recently, Muller et al. (2021) performed a more fine-grained analysis, examining representations at each layer of the model.

Despite the success of massively multilingual NMT models (Johnson et al., 2017; Bapna and Firat, 2019; Aharoni et al., 2019; Zhang et al., 2020), less effort has been made in analyzing multilingual NMT representations. Kudugunta et al. (2019) clustered the representations of different languages learned by multilingual NMT models showing that common representations emerge in the encoder. Mareček et al. (2020) found that while RNN models (Attention Bridge architecture) (Cífka and Bojar, 2018; Lu et al., 2018) learn to capture certain linguistic properties with an increasing number of target languages, Transformer models are largely unaffected. Recent work of Zhang et al. (2021) introduced a conditional routing layer in a form of gate selection between language-specific and language-independent projection, providing some insights on which components allow for the emergence of *interlingua*.

Our work builds on the findings from the attention heads analyses (Voita et al., 2019; Michel et al., 2019) but attempts to extend them to multilingual NMT, investigating whether it is possible to discover attention heads that are language pair specific. Also, we experimented with a set of attention-based metrics and analyzed how effective they are in pruning under different language pairs and types of attention.

## 3 Methodology

As our goal was to identify "important" attention heads for different language pairs, we first needed to define a metric or a procedure that can capture the notion of "importance" of an attention head, and selected heads based on this importance.

In Section 3.1, we present a set of metrics that quantify certain aspects of attention weights, which to some extent, can be considered as the importance. Section 3.2 illustrates a more direct approach where the importance of a head is defined as the extent of decrease in BLEU scores (Papineni et al., 2002) resulted in pruning the head.

[1] either shared vocabulary or shared special tokens such as ⟨SEP⟩, ⟨EOS⟩, etc.

### 3.1 Metrics Based on Attention Weights

We experimented with three types of metrics that are defined for each attention head, $\texttt{head}_{l\in L,h\in H}$, where $l$ and $h$ are the indices of layer and multi-head, respectively. In what follows we define how the metrics were computed for one sentence. Each metric was computed and averaged over a set of development sentences, then normalized to zero mean and unit standard deviation for ease of comparison. We note that $|I|$ and $|J|$ were the number of source tokens and/or target tokens, depending on whether we looked at the self-attention of encoder or the encoder-decoder cross attentions.

**Confidence** Voita et al. (2019) defined the notion of confidence of a head to be the mean of its maximum attention weights, and showed that only a small set of heads are confident and responsible for most of the model's performance.

$$\text{conf}(\text{head}) \coloneqq \frac{1}{|I|}\sum_{i\in I}\max_{j\in J}\alpha_{i,j}$$

**Variance** Inspired by Vig and Belinkov (2019), we computed the expected position of attention for token $i$ as $\mu_i \coloneqq \mathbb{E}[j|i] = \sum_{j\in J} j\cdot\alpha_{i,j}$, and measured how much each individual position was away from it:[2]

$$\text{var}(\text{head}) \coloneqq -\sum_{i\in I}\sum_{j\in J}\alpha_{i,j}\,(\mu_i - j)^2$$

**Coverage** Tu et al. (2016) defined the notion of coverage for encoder-decoder attentions which computes the amount of attention a source token has received. We extended the idea to the self-attentions in encoder as well.

$$\text{cov}(\text{head}) \coloneqq \sum_{j\in J}\left(\sum_{i\in I}\alpha_{i,j}\right)^2$$

More details on the metrics are provided in Appendix C.

### 3.2 Sequential Backward Selection of Heads

Intuitively, a head can be considered as important if its removal results in a drastic decrease in the BLEU scores. As different combinations of heads can affect the performance differently, we followed the sequential backward selection (SBS) algorithm (Aha and Bankert, 1996), which is a top-down algorithm starting from a feature set of all features (in

[2] As we wanted the important heads to have lower variance, we multiplied the score with $-1$.

**Algorithm 1:** SBS for Head Selection

```
selections ← ∅;
while |selections| < |L| · |H| do
    bleuMin ← ∞;
    headMin ← ∅;
    for ∀ head_{l∈L,h∈H} ∉ selections do
        masks ← selections ∪ head_{l,h};
        trans ← Translate(masks);
        bleuDrop ← Evaluate(trans);
        if bleuDrop < bleuMin then
            bleuMin ← bleuDrop;
            headMin ← head_{l,h};
        end if
    end for
    selections ← selections ∪ headMin;
end while
return selections;
```

our case, a set of all heads) and sequentially removing the most irrelevant features that maximize the evaluation metric (in our case, the BLEU score).

The pseudo-code for the head selection procedure is illustrated in Algorithm 1. The algorithm first selects a head that, when masked, results in the smallest decrease in the BLEU score; and adds it to selections. For subsequent iterations, it proceeds similarly, but the masks now include the heads in selections as well as the candidate head. The procedure terminates when all heads are selected. Note that the time complexity of the algorithm is $\mathcal{O}(|L|^2|H|^2)$, where $L$ and $H$ denote the set of layers and attention heads, respectively. It is a computationally intensive procedure as for each iteration, a test set is translated and evaluated.

## 4 Experiments and Results

### 4.1 Preliminary Experiment

We conducted a preliminary experiment using a many-to-one multilingual model trained on a TED talk dataset (Qi et al., 2018), covering top-20 source languages with the most data. We observed that patterns of attention heads (measured with the "confidence" metric) for both encoder self-attention and encoder-decoder attention were very similar among the language pairs.

For the main experiment, we decided to use a larger and stronger multilingual model for the following reasons: (1) the TED dataset is quite small and the model trained on it achieves lower BLEU scores and may not be regularised very well; (2) the network capacity of the TED model could be too limited for the language-pair-specific patterns to emerge (if any). According to a study on BERT's multilinguality (Dufter and Schütze, 2020), the increased network capacity (i.e., over-parameterization) is shown to lead to more decoupled representations between languages.

As the multilingual model described in Section 4.2 is trained on much larger datasets, and has a network capacity larger than the initial TED model while covering fewer language pairs, we expect that the language-pair specificity (if any) is more likely to emerge.

### 4.2 Experimental Settings

For the sake of reproducibility, all experiments were conducted using a strong publicly available many-to-one multilingual NMT model released by Bérard et al. (2020). The model can translate French, German, Italian, Spanish, and Korean sentences into English. It is trained with standard open-accessible datasets, including biomedical corpora where available. The model uses a variant of the Transformer-Big architecture (Vaswani et al., 2017) with a shallower decoder: 16 attention heads, 6 encoder layers, and 3 decoder layers. The model produces SOTA- or near-SOTA-level results for news, IWSLT, and biomedical translation tasks.

As the model is many-to-one, we could set up a controlled experiment where the BLEU scores were directly compared among the language pairs. We employed the development and test sets from the TED talk dataset, and utilized only the multilingual sentence pairs where both source and reference sentences were present for all five language pairs.[3] After the filtering, the development and test sets contained 1771 and 2137 pairs, respectively.

As we were using a many-to-one model, we conducted experiments on both encoder self-attentions and encoder-decoder attentions. In our experiments, we did not re-train or fine-tune the model when masking each head, making procedure lighter than other approaches involving the re-training.

### 4.3 Results

**Heads importance across languages** Figure 1 shows the heatmaps for each importance metric (Sect. 3.1) for the self-attention and cross-attention

[3]Note that the English reference sentences were the same across the language pairs.

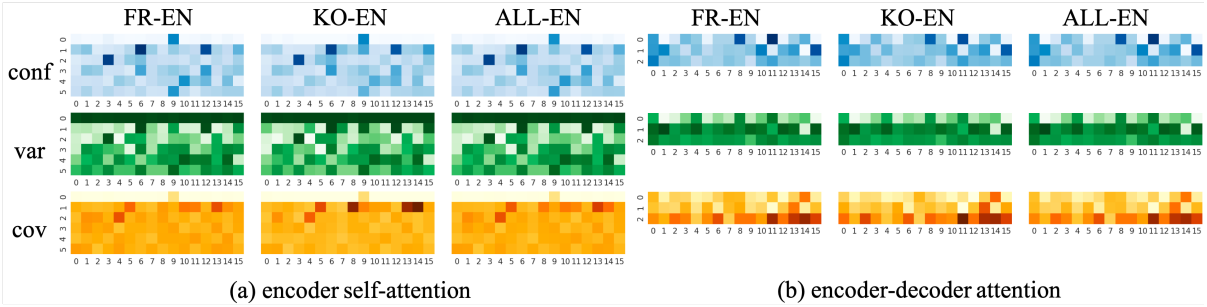



Figure 1: Each heatmap of a language pair shows the corresponding normalized metric scores for every **(a) encoder self-attention** and **(b) encoder-decoder attention** head, broken out by layer (vertical axis) and head (horizontal axis). For each metric, the color scales are identical across language pairs.

heads, respectively. The heatmaps were computed for each language pair separately (FR-EN, KO-EN, etc.) or jointly for all pairs (ALL-EN).[4] The main finding is that even if each metric displayed a different heatmap, the important heads were the same for all language pairs according to these metrics. In other words, the metrics did not highlight the emergence of language pair specific (encoder or cross) attention heads. Comparing among the metrics, *variance* and *confidence* tended to emphasize the same heads (with the exception of the first self-attention heads of each layer which were systematically rated as important by the *variance* metric). On the other hand, *coverage* highlighted different heads compared to the other two metrics.

**Impact of head selection on NMT performance** In the previous paragraph, we explored several metrics that could help capturing the importance of an attention head. We now analyze if these metrics could be used to prune heads and the corresponding impact on MT performance. We also investigate a more direct (but more costly) approach to measure how heads contribute to MT performance.

Figure 2 shows the evolution of BLEU curves as more and more heads were pruned. Head pruning was based on the importance metrics (removing least important heads first according to the metrics presented in Sect. 3.1) or on the SBS algorithm (Sect. 3.2). Head selection was conducted separately for each language pair,[5] and the curves were drawn from fitting polynomial regressions. First, we observed that, for both encoder self-attentions and cross attentions, it was possible to remove around 30% of the less important heads without much decrease of BLEU. Next, we noted that for cross attention head pruning, *coverage* seemed to be a better alternative than *confidence* and *variance*, while for encoder self-attention pruning *confidence* remained the most efficient. Intuitively, *coverage* metric is complementary to *confidence* in case of cross attention as it measures whether the whole input has been attended to. On the other hand, self-attention heads seemed devoted to specific phenomena (Voita et al., 2019; Clark et al., 2019) and there was no need to attend to the whole sentence for this matter. Finally, we also display the BLEU curves for randomly ranking (`rand-ranking`) the attention heads, confirming that the metrics proposed can be used as a proxy to measure the importance of heads and prune the least important ones. However, the exhaustive (but costly) SBS algorithm logically led to the best results.

**Is there really no emergence of language-specific heads?** We verified how statistically significant the BLEU differences were between *language-specific* and *language-independent* heads selection processes according to various metrics with Mann–Whitney U tests (Mann and Whitney, 1947).[6] We found no significant difference between language-specific and language-independent head rankings, even if some differences emerged for results obtained by SBS ranking.

Finally, we looked at how the individual head rankings were varied according to the SBS algorithm. Figure 3 illustrates the standard deviation of each head position among the rankings of the five different language pairs. We observed that there were a few heads whose relative importance varied greatly among the language pairs. For example,

[4] We only display FR-EN and KO-EN, reader should refer to Appendix B for all language pairs.

[5] The language-independent selection of heads led to a very similar plot as Fig. 2 and is provided in Appendix E

[6] We report p-values for Mann–Whitney U tests in the Appendix D.

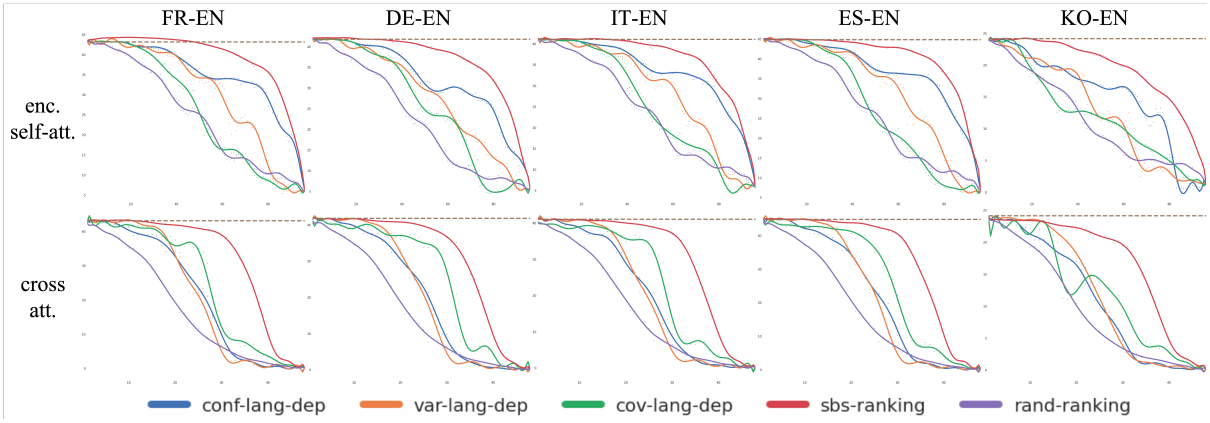
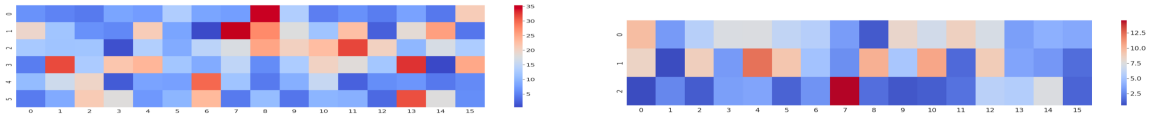



Figure 2: BLEU curves on test set when pruning subsequent **self-attention** and **cross attention** heads based on different importance metrics (or SBS) computed from dev set (language pair dependently). To be seen in color.



(a) encoder self-attention (b) encoder-decoder attention

Figure 3: Standard deviation of each **(a) encoder self-attention** and **(b) encoder-decoder attention** head ranking with SBS algorithm. SBS rankings range from 1 to 96 for self attention and from 1 to 48 for cross attention and standard deviation is calculated for each head, using these scores, among the five language pairs.

the $\texttt{head}_{2,7}$ of the encoder-decoder attention was ranked as least important for KO-EN but quite important for the other four language pairs.[7] Similarly, $\texttt{head}_{1,7}$ for encoder self-attention was ranked as not important for ES-EN while very important for KO-EN. This analysis showed that, even though the majority of important heads seemed to be language-independent, certain heads may capture different linguistic phenomena.

## 5 Conclusion and Future Work

We investigated if there are attention heads that are language pair specific within a many-to-one multilingual NMT model. We examined different metrics for heads selection process and found that *confidence* is a good proxy for self-attention heads "prunability", and *coverage* is a better indicator for cross attention heads "prunability".

We showed that, although it is possible to find the rare heads specific to a language pair via the extensive SBS procedure, the most important heads are language-independent; and it is possible to prune around 30% of the heads with no retraining and almost no loss in BLEU.[8]

As the findings from the SBS procedure indicated that some language pair specific heads do exist, a promising future direction is to perform pruning at different level of granularity (Frankle and Carbin, 2019; Zhao et al., 2020) (as opposed to single scalar values computed by the metrics) in order to identify which part of the model is more language-specific. Such analysis could help us to deploy multilingual models with better efficiency / performance trade-offs.

## Acknowledgments

This work was done as part of the Multidisciplinary Institute in Artificial Intelligence MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). Authors would like to thank Jaesong Lee and Hyun Chang Cho from NAVER Corp., Kwang Hee Lee from KAIST for insightful comments on the experiments. We would also like to extend our gratitude to the anonymous reviewers for their valuable feedback.

[7] Masking this single head alone, resulted in an increase in BLEU for KO-EN by 0.03, while for others, a decrease in BLEU up to 0.5.

[8] It is possible that fine-tuning the model after pruning the heads may lead to better BLEU scores, and therefore more aggressive pruning (Behnke and Heafield, 2020; Wang et al., 2020a).

## References


David W. Aha and Richard L. Bankert. 1996. *A Comparative Evaluation of Sequential Feature Selection Algorithms*. Springer New York, New York, NY.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Maximiliana Behnke and Kenneth Heafield. 2020. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2664–2674, Online. Association for Computational Linguistics.

Alexandre Bérard, Zae Myung Kim, Vassilina Nikoulina, Eunjeong Lucy Park, and Matthias Gallé. 2020. A multilingual neural machine translation model for biomedical data. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Nikolay Bogoychev. 2020. Not all parameters are born equal: Attention is mostly what you need.

Ondřej Cífka and Ondřej Bojar. 2018. Are BLEU and meaning representation in opposition? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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. Association for Computational Linguistics.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium. Association for Computational Linguistics.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*.

David Mareček, Hande Celikkanat, Miikka Silfverberg, Vinit Ravishankar, Jörg Tiedemann, et al. 2020. Are multilingual neural machine translation models better at capturing linguistic features? *The Prague Bulletin of Mathematical Linguistics*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, pages 14014–14024. Curran Associates, Inc.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. In *EACL 2021*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Yong Wang, Longyue Wang, Victor Li, and Zhaopeng Tu. 2020a. On the sparsity of neural machine translation models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020b. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

## A Experimental Details

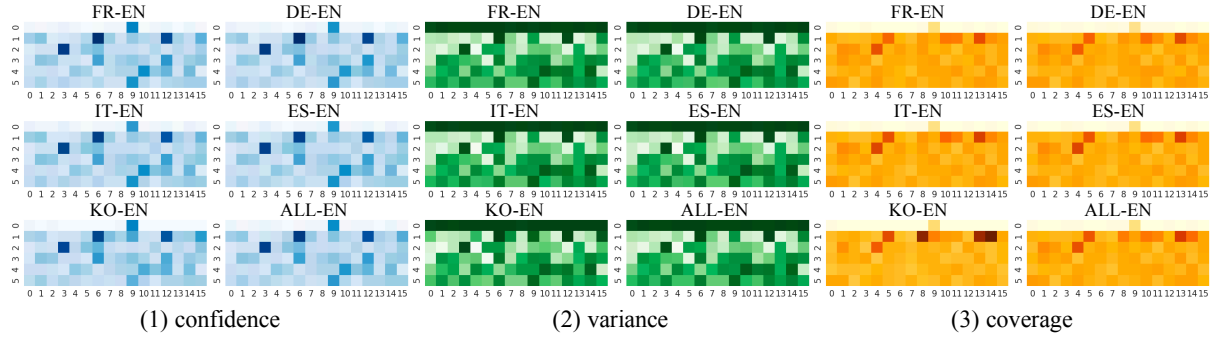
All the experiments were conducted using PyTorch (Paszke et al., 2019) and Fairseq (Ott et al., 2019) toolkit. The multilingual NMT model used in the experiments can be downloaded online.[9] Please refer to Bérard et al. (2020) for more details on the model.

When running an experiment for each language pair, a single V100 GPU was used. We note that computing the SBS rankings for encoder self-attention was the most computationally intensive part, where almost $96^2$ translations of the development set were conducted.

When computing the BLEU curves for the `rand-ranking`, we ran the procedure with a randomly created ranking five times, and averaged the resulting BLEU scores.

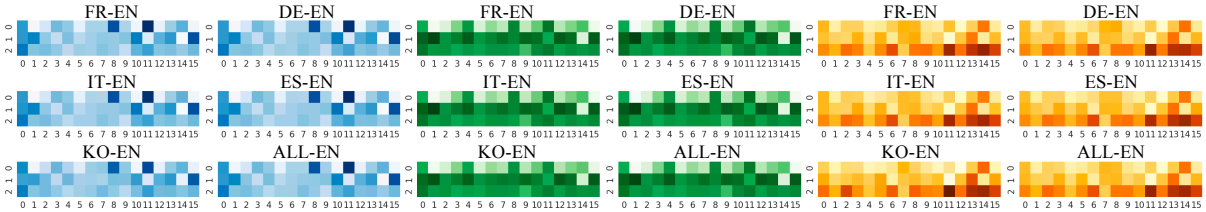
## B Heatmaps of Metric Scores for All Language Pairs

Figure B illustrates the normalized metrics scores for every attention head. We observed that for each metric, the patterns are consistent across all language pairs.




(1) confidence (2) variance (3) coverage

(a) encoder self-attention




(1) confidence (2) variance (3) coverage

(b) encoder-decoder attention

Figure B: Each heatmap of a language pair shows the corresponding normalized metric scores for every **(a) encoder self-attention head** and **(b) encoder-decoder attention head**, broken out by layer (vertical axis) and head (horizontal axis). For each metric, the color scales are identical across language pairs.

## C Remarks on the Importance Metrics

We denote $|L|$ and $|H|$ to be the number of layers and heads, respectively, while $|S|$ and $|T|$ represent the number of source and target tokens. When calculating each metric, we began with the tensor shape, $(|L|, |H|, |S|, |S|)$ or $(|L|, |H|, |T|, |S|)$, depending on whether we were computing for the encoder self-attention or the cross attention. After the computation, the shape of the outcome tensor was: $(|L|, |H|)$.

### C.1 Confidence

We noted that the patterns of the confidence scores for each head tended to vary depending on the length of sentences we used to compute the scores. This was due to the fact that the metric was calculated by averaging over the maximum attention, which was inversely proportional to the length of sentences.

[9] https://github.com/naver/covid19-nmt

### C.2 Variance

The variance metric was defined so that heads with a small variance were considered to be important. A small variance was achieved when most of the attention weights were focused on one or a few positions. While this intuition came initially from encoder-decoder attention (interpreting attention as a source-target alignment), it is less clear if it holds for encoder self-attention as well (our results seemed to suggest that it is not the case).

### C.3 Coverage

While the notion of coverage was initially proposed for encoder-decoder attention, we extended it to the encoder self-attention. We may consider it as how much a source token has been attended from its neighbouring source tokens. Similar to the variance metric, for the encoder self-attention, the importance of high coverage is less clear where a head may play a specific role as discussed in Voita et al. (2019); Clark et al. (2019). This probably accounts for the reason that the head pruning of the encoder self-attention was not as effective as that of the cross attention.

## D P-Values for Mann–Whitney U tests

As the BLEU curves obtained from language-specific pruning and language-independent pruning were very similar, we performed a non-parametric statistical test, namely, Mann-Whitney U test, to compare the outcomes. The test checks whether two samples are likely to derive from the same population (i.e. that the two populations have the same shape).

Table D shows the p-values for the two-sided tests between BLEU curves computed using language-specific and language-independent metrics for encoder self-attention and cross attention.

The high p-values ($> 0.05$) across all language pairs suggest that the differences in the BLEU scores computed from the two scenarios were statistically insignificant.

| | FR-EN | DE-EN | IT-EN | ES-EN | KO-EN |
|---|---|---|---|---|---|
| conf | 0.938 | 0.849 | 0.871 | 0.878 | 0.902 |
| var | 0.927 | 0.995 | 0.959 | 0.939 | 0.573 |
| cov | 0.570 | 0.555 | 0.865 | 0.927 | 0.850 |
| sbs | 0.137 | 0.189 | 0.293 | 0.878 | 0.375 |

(a) encoder self-attention

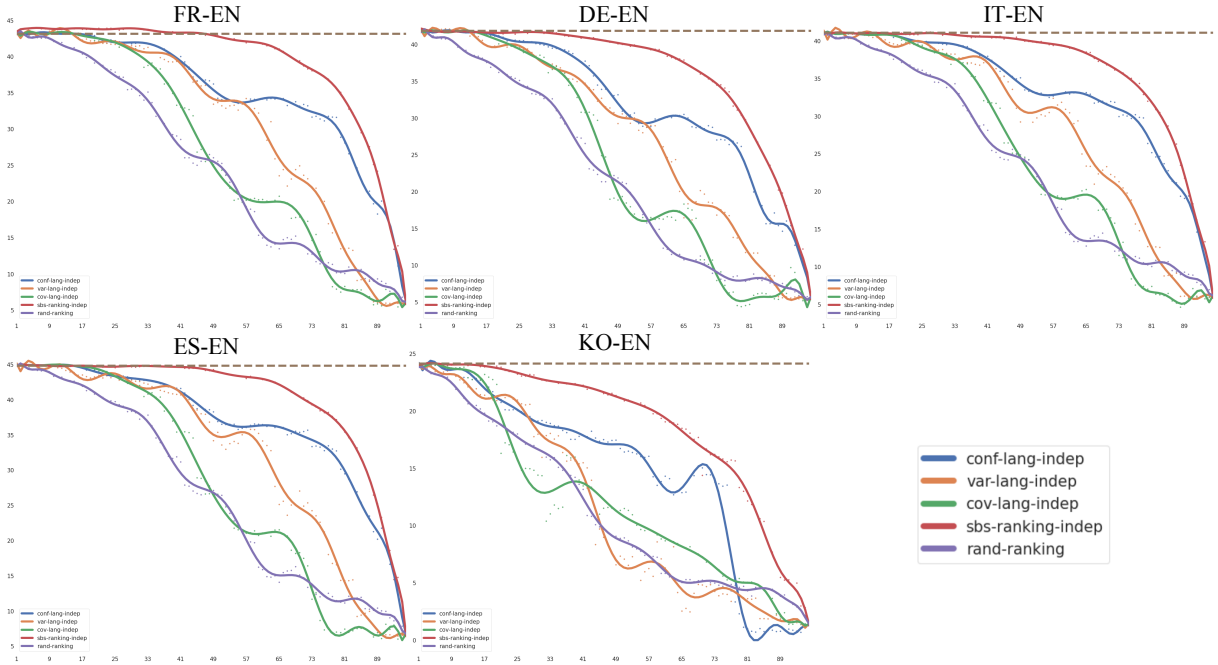
| | FR-EN | DE-EN | IT-EN | ES-EN | KO-EN |
|---|---|---|---|---|---|
| conf | 0.918 | 0.988 | 0.965 | 0.968 | 0.881 |
| var | 0.991 | 0.994 | 0.997 | 0.985 | 0.936 |
| cov | 0.772 | 0.907 | 0.912 | 0.889 | 0.621 |
| sbs | 0.901 | 0.631 | 0.936 | 0.918 | 0.404 |

(b) encoder-decoder attention

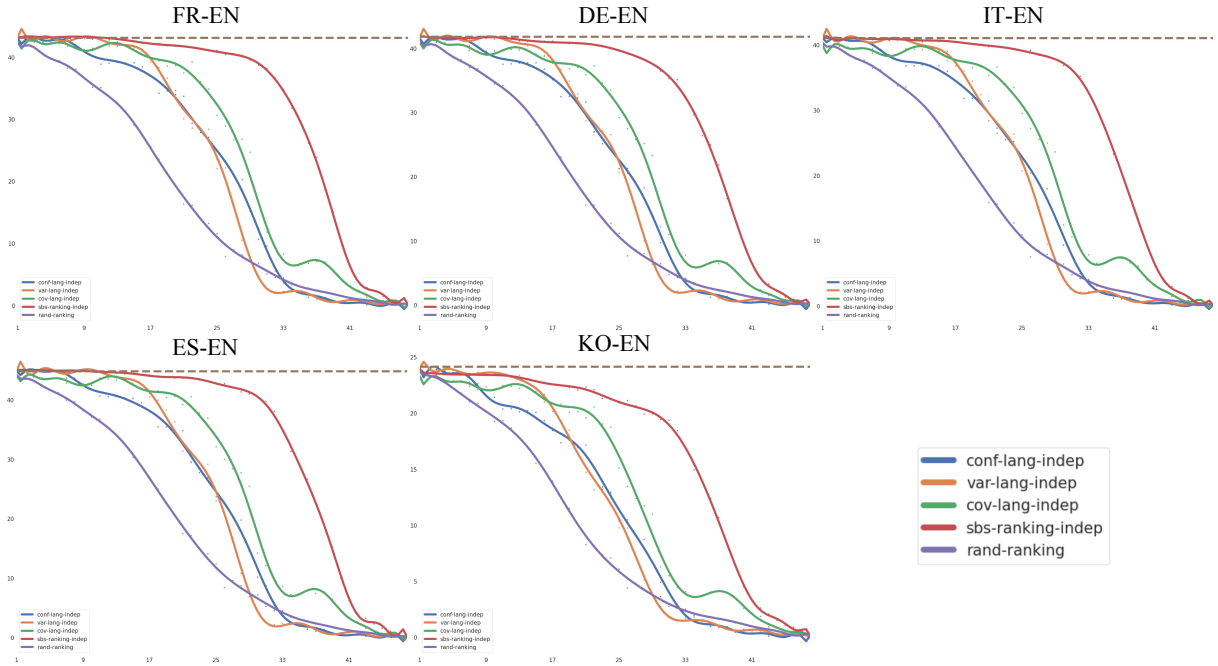
Table D: P-values for Mann–Whitney U tests between BLEU scores computed using language-specific and -independent metrics for **(a) encoder self-attention** and **(b) encoder-decoder attention**.

## E BLEU Curves (Language-Independent Head Selection) for All Language Pairs

In Figure E, we present the BLEU curves obtained from pruning the encoder self-attention heads and cross attention heads according to the importance metrics (and SBS) computed over *all language pairs* (i.e. language-independent). We observed that the curves were very similar to those presented in Figure 2 of the main paper, where the computation was conducted over the *specific language pairs*.




(a) encoder self-attention




(b) encoder-decoder attention

Figure E: BLEU curves on test set when pruning subsequent **(a) encoder self-attention heads** and **(b) encoder-decoder attention heads** based on different importance metrics (or SBS) computed from the development set (language pair independently).