Quantifying the Overlap: Attribution Maps and Linguistic Heuristics in Encoder-Decoder Machine Translation Models

Aria Nourbakhsh[1] **Salima Lamsiyah**[1] **Christoph Schommer**[1]
Department of Computer Science, Faculty of Science, Technology and Medicine,
University of Luxembourg
{aria.nourbakhsh,salima.lamsiyah,christoph.schommer}@uni.lu

Abstract

Explainable AI (XAI) attribution methods seek to illuminate the decision-making process of generative models by quantifying the contribution of each input token to the generated output. Different attribution algorithms, often rooted in distinct methodological frameworks, can produce varied interpretations of feature importance. In this study, we utilize attribution mappings derived from three distinct methods as weighting signals during the training of encoder-decoder models. Our findings demonstrate that Attention and Value Zeroing attribution weights consistently lead to improved model performance. To better understand the linguistic information these mappings capture, we extract part-of-speech (POS), dependency, and named entity recognition (NER) tags from the input-output pairs and compare them with the XAI attribution maps. Although the Saliency method shows greater alignment with POS and dependency annotations than Value Zeroing, it exhibits more divergence in places where its attributions do not conform to these linguistic tags, compared to the other two methods, and it contributes less to the models' performance[1].

1 Introduction

The remarkable advancements in machine learning have led to increasingly sophisticated and robust models. However, this progression towards larger and more complex architectures, often referred to as 'black boxes', has come at the cost of transparency (Xu et al., 2019; Arya et al., 2019; Vieira and Digiampietri, 2022; Saeed and Omlin, 2023). As a result, interpreting the internal decision-making processes of these systems has become progressively more challenging (Jacovi and Goldberg, 2020). To improve our understanding of these models, researchers have turned to XAI methods, which aim to provide greater transparency and facilitate more interpretable outcomes (Lipton, 2018). A key approach in XAI involves attribution methods, which seek to identify and quantify the contribution of individual input features to a model's prediction (Sundararajan et al., 2017a; Wallace et al., 2019; Madsen et al., 2022).

In this work, we explore XAI attribution methods in Transformer-based (Vaswani, 2017) Sequence-to-Sequence (seq2seq) models, which are predominantly based on the encoder-decoder architecture (Sutskever, 2014). Encoder-decoder architecture is a common pipeline in many NLP tasks, such as machine translation, text summarization, and dialogue generation. The literature has focused more on attribution methods in simpler classification tasks (Lal et al., 2021; Attanasio et al., 2023), and comparatively less attention has been given to generative seq2seq models. In the context of seq2seq models, attribution methods aim to identify which parts of the input sequence were most influential in generating each segment of the output sequence (Sarti et al., 2023). In machine translation (MT), several studies have sought to quantify and compare the word alignments produced by statistical models such as GIZA++ (Och and Ney, 2003) or annotated alignments (Ding et al., 2019; Li et al., 2019; Chen et al., 2020).

To obtain the attribution of features or tokens, several methods have been proposed. Approaches such as Saliency maps (Simonyan et al., 2013), Integrated Gradients (Sundararajan et al., 2017b), and DeepLIFT (Shrikumar et al., 2017) analyze the gradients of the output with respect to the input embeddings. These approaches are called post-hoc methods, in which the model behaviour is extracted after the training process (Arrieta et al., 2020). In these methods, higher gradient values indicate greater importance of the corresponding input feature. Perturbation-based methods involve

[1] https://github.com/ariana2011/xai_att_ling.git

Proceedings of Recent Advances in Natural Language Processing,pages 822–831
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-094

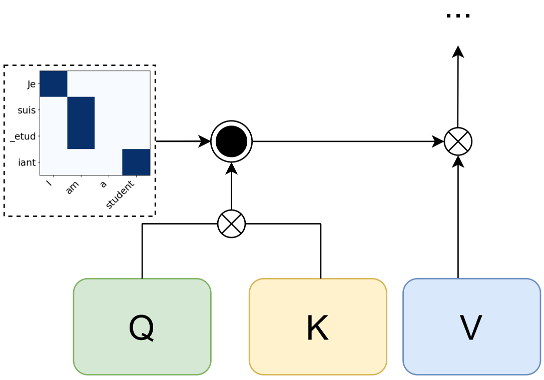

Figure 1: After the extraction of the attribution maps, we integrate this information into the attention mechanism throughout the training and testing processes.

systematically altering parts of the input sequence (e.g., masking or replacing) and observing the impact on the output (Ivanovs et al., 2021). Significant changes indicate the importance of the perturbed elements. On the other hand, the attention mechanism is an inherent, model-based interpretable component (Bahdanau et al., 2016), although some studies have questioned whether its representations truly reflect the model's decision-making process (Serrano and Smith, 2019).

Having obtained attribution maps from XAI methods, evaluating their faithfulness and validity remains a challenging task (Nielsen et al., 2023; Kamath et al., 2024). How do we know if an attribution map is truly reflecting the model's reasoning? This work is based on the premise that distinct attribution methods generate unique sets of scores. We consider these scores to be informative weights, signifying that each method captures a distinct perspective on how input tokens contribute to the model's output. We conjecture that these weights can be integrated with the model as external knowledge, and their effects can be measured in terms of the output performance. Additionally, our approach raises questions about the nature of the extracted relations. Do attribution maps capture surface-level statistical associations, or do they correspond to deeper, linguistically meaningful relationships, such as those indicated by POS and dependency structures? Considering these questions, we outline our contributions:

- We evaluate and compare three attribution methods, each representing a distinct category, to measure their effects on four machine translation tasks.
- We propose an approach to construct semantic and syntactic mappings between source and target sequences.
- We use the extracted linguistic information to compare XAI-based attribution methods in terms of the knowledge they encode.

2 Related work

2.1 Evaluation of XAI methods

There are various ways to evaluate XAI approaches. One common strategy is human judgment, which assesses whether the explanations produced by XAI methods align with human intuition and expectations (Kim et al., 2024; Lopes et al., 2022). However, human evaluation is costly and time-consuming. Notable automatic evaluation techniques for attribution maps have been developed in the context of image classification (Ribeiro et al., 2016; Hooker et al., 2019; Nauta et al., 2023) as well as NLP (Madsen et al., 2022; Moradi et al., 2021), where researchers mask or isolate or keep only the highlighted regions during training or testing. These approaches operate under the assumption that the marked attributions correspond to important features that affect the model's performance.

In the context of MT Li et al. (2019); Chen et al. (2020) have tried to understand if NMT models capture traditional word alignment between source and target words. Key findings indicate that while NMT models capture alignment information and dedicated alignment objectives, hybrid models often enhance both translation accuracy and alignment reliability. Zenkel et al. (2020) They address the challenge of word alignment induction in NMT, specifically with Transformer architectures. While previous research indicated that Transformer attention weights yield poor alignments, this study demonstrates that attention can produce accurate alignments when extracted at the appropriate decoding step. Ding et al. (2019) more interestingly, they use Saliency and SmoothGrad methods to induce the word alignment. Experimental results reveal that, especially under force decoding, their proposed methods can yield higher-quality alignments than traditional tools like fast-align (Dyer et al., 2013), highlighting the latent alignment capabilities of NMT systems. On the other hand, Ferrando and Costa-jussà (2021) present an in-depth analysis of attention weights in Transformer-based NMT models. Their study finds that encoder-decoder attention weights systematically make alignment errors and often focus on uninformative source tokens rather than accurately aligning corresponding words between sequences.

A closer work to ours is done by (Li et al., 2020), where they train surrogate models to predict words based on a window of highest attribution tokens extracted from XAI methods. The current work differs in that we measure the influence of attribution methods directly within the NMT task.

2.2 Attribution Methods

Saliency: Saliency computes gradients of the output with respect to the input to generate a Saliency map, highlighting influential features (Simonyan et al., 2013). For an input x and a model $f(x)$ that outputs a score for a class c, the Saliency $S(x)_i$ for the i-th input feature is often computed as the magnitude of the gradient of the output with respect to that input feature:

$$S(\mathbf{x})_i = \left| \frac{\partial f_c(\mathbf{x})}{\partial x_i} \right|$$

Attention: In the Transformer model, attention is a mechanism that allows the model to weigh the importance of different tokens in an input sequence when encoding or decoding at each position (Vaswani, 2017). Specifically, for each token, the model computes three vectors of query, key, and value by projecting the token's embedding through learned weight matrices. Attention scores are obtained by taking the dot product of each query with all keys, scaling and normalizing via a softmax to produce weights that reflect how much one token should "attend" to another:

$$\text{Attention} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Value Zeroing: Value Zeroing is a novel method to analyze how information is integrated across tokens within Transformer models. This approach overlooks the significant roles of other components, such as feedforward networks, in the encoder block. Value Zeroing addresses this by zeroing out the value vectors of specific tokens during the forward pass, which in turn allows for an assessment of each token's contribution to the model's output. The effect is calculated by determining the distance between a changed output representation of token $\tilde{x}_i^{\neg j}$ where $\tilde{x}_i^{\neg j}$ is calculated by removing token j's value vector (Mohebbi et al., 2023):

$$C_{i,j} = \text{cosine}(\tilde{x}_i^{\neg j}, \tilde{x}_i)$$

Inseq is a Python library that provides a comprehensive tool for analyzing and comparing different explainability methods of generative language models (Sarti et al., 2023). The library offers a range of gradient-based, perturbation-based, and internal representations of the encoder-decoder transformer models.

3 Methodology

To evaluate each of these XAI methods, we use a dataset of source and target pairs and apply the attribution technique to a pre-trained model, deriving a 2D matrix that captures the contribution of each input token to the output tokens. $\left(s_{\text{src}}^{(i)},\ s_{\text{tgt}}^{(i)}\right) \xrightarrow{\text{Attribution}} \left(s_{\text{src}}^{(i)},\ s_{\text{tgt}}^{(i)},\ e^{(i)}\right)$. Then the result is a triplet of source, target, and explanation mapping, which we then use to train and test an untrained model.

For the Saliency method, this gradient-based approach yields a tensor $e \in \mathbb{R}^{j \times k \times l}$, where j is the input sequence length, k is the output sequence length, and l is the hidden dimension of the model as the gradient is calculated for each dimension of the input token vector. We compute the ℓ_2-norm along the last dimension to reach $e \in \mathbb{R}^{j \times k}$ for the attribution of each input token with respect to the output token at position k. ℓ_2-norm measures the length or magnitude of a vector in a Euclidean space.

For Attention, the scores are calculated for each head and for each layer of the cross-attention weights. As a result, the output of the attention score is $e \in \mathbb{R}^{j \times k \times n \times h}$, where n and h are the number of layers and heads respectively. We get the average over the last two axes of the representation to reach the same $e \in \mathbb{R}^{j \times k}$ representation. Likewise, Value Zeroing is calculated by measuring the effect of replacing the value vector of x_i with a zero vector on all layers, so the output $e \in \mathbb{R}^{j \times k \times n}$. We get the average of the scores along the last dimension. In all cases, we reach the mapping of values indicating the weight of the connection between the output tokens and each individual input token.

To identify the single most 'attributing' token in each row of the attribution map, we first apply a row-wise softmax:

$$S_{jk} = \frac{\exp(X_{jk})}{\sum_{m=1}^{K} \exp(X_{jm})}$$

So that each row J of S sums to 1 across columns. We then form a one-hot mask $e \in \{0,1\}^{j \times k}$ by setting

$$E_{jk} = \begin{cases} 1, & \text{if } k = \arg\max_{m} S_{jm}, \\ 0, & \text{otherwise.} \end{cases}$$

Taking the argmax of the attribution factors is a common approach for converting the soft scores of attribution maps into hard scores (Garg et al., 2019; Chen et al., 2020).

In the attention head, the dot product of Q and K determines how relevant each word query is to the key of all other tokens in the sequence. To incorporate this information into the attention mechanism, we utilize the Q, K, V matrices as usual, and inject the attribution masks $\mathbf{E}$ as follows:

$$\text{Attention} = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}} \odot \mathbf{E}\right) V \quad (1)$$

Here $\odot$ denotes element-wise multiplication and $\mathbf{E}$ is broadcasted (i.e., 0 padded) to match the shape of the attention logits. One should pay attention as the attributions are binary numbers; multiplying by binary matrices simply zeroes the product of the dot product at positions where $E_{jk} = 0$ before the softmax. This operation has been done only on the encoder self-attention.

4 Model and Data

Throughout our experiments, we used the Huggingface implementation of Opus-MT[2] to extract the attributions and train the models from scratch. To get the attribution mappings according to the XAI methods, we used the Inseq (Sarti et al., 2023)[3] library to derive the mappings for each language pair.

To evaluate our approach, we used 200,000 samples from the French→English (fr-en), German→English (de-en) (Bojar et al., 2014), Spanish→Italian(es-it), and English→Danish (en-da) (Koehn, 2005) language pairs. We selected samples with fewer than 128 tokens to ensure efficient training and evaluation of our models from scratch. We reserved 15,000 samples for evaluation and testing for each language pair. The models were trained for 20 epochs, with early stopping triggered after three epochs without improvement on the evaluation set. As a baseline, we used models without attribution merging (vanilla models). Our models matched the original Opus-MT configuration, using six attention layers and eight attention heads. In all experiments, the injected attribution information was incorporated into every layer of the encoder self-attention. To evaluate our models, we used the BLEU score throughout the experiments.

5 Results

5.1 Effect of Attribution Injection on Model Output

The results in Table 1 show that incorporating attribution maps into the attention mechanism significantly improves performance across various language pairs. The Attention and Value Zeroing methods consistently yield better results compared to both the Saliency approach and the baseline model. For instance, the fr-en and es-it pairs see approximately a 10-point increase when using the Attention and Value Zeroing attributions merged. The gaps are even wider for the en-da pair for the Attention attributions.

Moreover, the Value Zeroing method (except for en-da) closely matches the Attention method's performance. This suggests that possibly Value Zeroing and Attention may capture similar attribution patterns. In contrast, the Saliency method offers only modest gains over the baseline in fr-en and en-da pairs and generally lags behind the other two methods. These findings highlight the substantial benefit of attribution-informed adjustments to the attention mechanism, which enhance translation quality across multiple language scenarios, as well as a comparison between the methods from which these attributions are derived.

5.2 Encoded Information

As the result suggested, when these attribution maps are merged with sentence pairs during both training and testing, we see a significant improvement in the evaluation metric. This consistent increase across all evaluated language pairs suggests that attribution-guided signals provide the model with valuable cues for sequence alignment and

[2] https://huggingface.co/Helsinki-NLP

[3] https://inseq.org/

Language Pair	Saliency	Attention	Value Zeroing	Baseline
fr-en	**30.02**	**40.00**	**39.95**	28.79
es-it	29.07	**36.74**	**38.40**	29.05
de-en	16.74	**22.41**	**23.69**	16.85
en-da	**15.43**	**38.67**	**26.35**	13.93

Table 1: BLEU score results of the baseline and merging attributions to the encoder-decoder model.

Language Pair	POS	DEP	NER	Baseline
fr-en	**43.46**	**37.67**	27.80	28.79
es-it	**40.57**	**38.23**	29.04	29.05
de-en	**21.05**	15.14	15.41	16.85
en-da	**21.94**	13.82	11.60	13.93

Table 2: BLEU score results of the baseline and merging POS, DEP, and NER tags to the encoder-decoder model.

translation quality. Such improvements highlight the utility of incorporating interpretability tools not just for model understanding but as active components in enhancing model behavior. However, this observation raises an important question regarding the nature of the information captured by these attribution maps. What exactly do these attribution mappings encode that leads to such performance gains? It is plausible that these attribution methods highlight linguistically meaningful correspondences between source and target tokens or capture subtle contextual dependencies. Understanding whether the injected attributions reflect lexical alignments, semantic equivalence, or more abstract relationships could provide insights into the mechanisms underlying improved performance.

To further investigate the nature of the linguistic insights captured by attribution maps, we propose examining their overlap with explicit linguistic annotations. Specifically, we aim to identify whether the information encoded by attribution-based mappings aligns with syntactic and semantic relationships represented through linguistically informed matrices. To accomplish this, we employ SpaCy[4] to annotate our datasets, extracting Part-of-Speech (POS) tags, dependency parsing relationships (DEP), and Named Entity Recognition (NER) labels.

Formally, given a source sentence $S = (s_1, \ldots, s_J)$ and a target sentence $T = (t_1, \ldots, t_K)$, we construct binary matrices to encode these linguistic annotations explicitly. For each annotation type, we define a binary matrix $\mathbf{L} \in \mathbb{R}^{J \times K}$ such that each element is computed as:

$$\mathbf{L}_{jk} = \begin{cases} 1, & \text{if tags of } s_j \text{ and } t_k \text{ match,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

To account for the subword tokenization commonly done in the transformer models, we assign the tag of the untokenized token to each subtoken. As a result, our derived linguistic tags are not one-hot encoded[5]. In the next step, we train our models with the same setting, but instead of injecting the attribution maps, we provide the models with the POS, DEP, and NER mapping of the source and target pairs.

Table 2 presents the outcomes of incorporating linguistic knowledge during both training and testing of the model. For the French-English (fr-en) and Spanish-Italian (es-it) pairs, integrating part-of-speech (POS) and dependency (DEP) tags leads to the most significant improvements over the baseline, as indicated by the bolded scores in those columns. These results suggest that, for Romance language pairs, syntactic information is particularly valuable for enhancing the performance of encoder-decoder models. In the German-English (de-en) setting, improvements are more modest and only happen for the POS tags alignment. For en-da, multiplying DEP and NER information does not yield improvements; in fact, it reduces performance, while POS tagging provides a marginal gain similar to de-en.

[4] https://spacy.io/

[5] We use a heuristic to assign the tag to each subword. The Opus-MT tokenizer does not provide the offset mapping.

German, Spanish, English, and Italian differ in this respect due to their morphological characteristics. German is a morphologically rich language characterized by extensive inflection. Furthermore, German exhibits complex word formation through processes like compounding, where multiple morphemes fuse to create new lexical items (Günther et al., 2019). Assignment of a tag to all the subwords most likely dilutes the fine-grained specific attention weight required for extracting better attributions. We guess that this leads to a lower gain compared to other language pairs.

In all cases, incorporating NER information resulted in a decrease in performance. This is most likely because NER tags are less frequent than POS and DEP tags. When NER information is unavailable for certain tokens, multiplying by a zero matrix leads to a loss of signal, thereby reducing overall model performance.

Quantifying overlap. After observing the impact of linguistic annotation on model performance, we aimed to quantify the extent to which attribution methods overlap with each linguistic tag. To this end, we measure the recall to quantify this overlap. Let $\mathbf{T} \subseteq \{1, \ldots, N\}$ denote the set of positions annotated with a specific linguistic tag, and $\mathbf{A} \subseteq \{1, \ldots, N\}$ denote the set of positions identified by the attribution method in linearized matrices. We calculate the recall as follows:

$$\text{Recall} = \frac{|\mathbf{T} \cap \mathbf{A}|}{|\mathbf{T}|}$$

Table 3 presents the recall scores quantifying the overlap between each attribution method and the linguistic tags. One immediate observation is the consistently low recall for the NER tags across all methods and language pairs, typically below 3.0. This likely reflects the low frequency of named entity tokens relative to POS and dependency tags. Notably, NER also did not contribute to improved model performance and evaluations with this information. However, still, Attention attributions overlap more with the NER tags than the other two XAI methods.

In contrast, POS and DEP tags demonstrate substantially higher coverage by the attribution maps, especially for Attention and Saliency methods. For example, for the fr-en language pair on POS tags, Attention achieves a recall of 26.56, while Saliency achieves 25.30. DEP has the lowest recall scores for de-en compared to other languages. The Value Zeroing method tends to show lower and more varied coverage compared to Attention or Saliency.

However, these recall values merely reflect where the attributions overlap with the annotated indices. Notably, Saliency overlaps more with POS and DEP, yet, as earlier results showed, Saliency does not enhance the model's performance. To further investigate, we also calculated the recall for pairs of attribution methods (reported at the bottom of Table 3). Here, Attention and Value Zeroing exhibit closer recall scores across all languages, whereas the gap between Attention and Saliency, as well as Value Zeroing and Saliency, remains wider. This pattern suggests that while Saliency covers a greater portion of POS and DEP information, possibly where they don't overlap, it has a degrading effect and potentially introduces noise, diminishing its contribution to model performance.

6 Discussion and Conclusion

In this study, we investigated the impact of three attribution mapping methods for seq2seq models, each designed to quantify the contribution of individual input tokens to the generation of output tokens. We also extracted linguistic information, specifically, part-of-speech (POS), dependency (DEP), and named entity (NER) tags, for both the source and target languages and aligned them together. By identifying and aligning tokens with matching tags across source and target sentences, we constructed a heuristic mapping to serve as a proxy for cross-lingual token alignment. This mapping was subsequently used as a form of knowledge injection to train new models from scratch.

Our results demonstrate that both Attention-based and Value Zeroing attribution mappings consistently enhance model performance, as measured by the BLEU score, compared to the gradient-based Saliency method. Additionally, the integration of dependency and POS tag mappings in fr-en and es-it led to an improvement of the results, but for de-en and en-da, only POS tags increased the performance.

While both Attention and Saliency methods exhibited higher overlap with POS and DEP information, the relatively lower correspondence between Saliency and the other two attribution methods suggests that Saliency highlights input regions that may dilute the effectiveness of where it covers POS and DEP information, ultimately resulting in diminished performance compared to Attention and

Pairs	fr-en	es-it	de-en	en-da
Attention-POS	26.56	41.16	32.89	37.44
Attention-DEP	11.30	31.50	4.98	17.06
Attention-NER	1.16	2.73	0.88	1.92
Value Zeroing-POS	14.64	21.31	21.02	24.88
Value Zeroing-DEP	5.99	15.56	4.26	10.68
Value Zeroing-NER	0.75	1.73	0.70	1.46
Saliency-POS	25.30	28.92	31.78	30.17
Saliency-DEP	11.91	22.18	4.94	15.93
Saliency-NER	0.84	1.44	0.78	1.48
Attention-Value Zeroing	66.92	57.02	62.53	66.28
Attention-Saliency	22.38	23.99	23.61	24.79
Value Zeroing-Saliency	15.03	14.84	14.03	16.42

Table 3: Recall scores for different attribution methods across language pairs, indicating the overlap with POS, DEP, NER tags, and other attribution methods.

Value Zeroing approaches.

These results are noteworthy from several perspectives. First, the location where these mappings are merged is itself of interest. While self-attention in the encoder is generally designed to map information from input to input, introducing meaningful information from the output to the input can alter the model's predictive power. Second, in cases where the encoded information is insufficient (e.g., NER or Saliency), the changes in performance compared to the baseline are minimal. This latter point aligns with the observations of Wiegreffe and Pinter (2019), who noted that models can still learn effectively under noisy attention. However, when attention was provided with well-encoded information, it had the potential to significantly enhance the model's predictive performance.

7 Future work

For future work, we suggest exploring the integration of additional linguistic features that capture finer-grained commonalities such as gender, number, or morphological characteristics at a more granular token level. Incorporating these attributes could further refine the mapping between source and target languages. Also, we used one-hot encoded attribution mappings. It would be interesting to conduct the same experiments by assigning 1 to the k highest attribution values.

Finally, in this study, we relied on external models to extract both attribution mappings and linguistic tags, which proved effective in improving model performance in the case of POS. However, our current approach still requires access to these annotations during inference. A promising direction for future research would be to enable the models to internalize and learn these attributions or linguistic patterns during training, hence eliminating the need for explicit extraction and application of these features at test time. Such an approach can be found in Bai et al. (2022).

Acknowledgments

This work is supported by the Luxembourg National Research Fund (FNR) as part of the project C21 - Collaboration 21: IPBG2020/IS/14839977/C21.

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, page 256–266. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Jiangang Bai, Yujing Wang, Hong Sun, Ruonan Wu, Tianmeng Yang, Pengfei Tang, Defu Cao, Mingliang Zhang1, Yunhai Tong, Yaming Yang, Jing Bai, Ruofei Zhang, Hao Sun, and Wei Shen. 2022. Enhancing self-attention with knowledge-assisted attention maps. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–115, Seattle, United States. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors. 2014. *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Javier Ferrando and Marta R. Costa-jussà. 2021. Attention weights in transformer NMT fail aligning words between sequences but largely explain model predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 434–443, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Fritz Günther, Eva Smolka, and Marco Marelli. 2019. 'understanding'differs between english and german: Capturing systematic language differences of complex words. *Cortex*, 116:168–175.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.

Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. 2021. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sandesh Kamath, Sankalp Mittal, Amit Deshpande, and Vineeth N Balasubramanian. 2024. Rethinking robustness of model attributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2688–2696.

Jenia Kim, Henry Maathuis, and Danielle Sent. 2024. Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence*, 7:1456486.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit Proceedings of Conference*, pages 79–86. International Association for Machine Translation.

Vasudev Lal, Arden Ma, Estelle Aflalo, Phillip Howard, Ana Simoes, Daniel Korat, Oren Pereg, Gadi Singer, and Moshe Wasserblat. 2021. InterpreT: An interactive visualization tool for interpreting transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational*

Linguistics: System Demonstrations, pages 135–142, Online. Association for Computational Linguistics.

Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi. 2020. Evaluating explanation methods for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 365–375, Online. Association for Computational Linguistics.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. 2022. Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19):9423.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2022. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1731–1751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Ian E Nielsen, Ravi P Ramachandran, Nidhal Bouaynaya, Hassan M Fathallah-Shaykh, and Ghulam Rasool. 2023. Evalattai: a holistic approach to evaluating attribution maps in robust and non-robust models. *IEEE Access*, 11:82556–82569.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Waddah Saeed and Christian Omlin. 2023. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017a. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017b. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Carla Piazzon Vieira and Luciano Antonio Digiampietri. 2022. Machine learning post-hoc interpretability: A systematic mapping study. In *Proceedings of the XVIII Brazilian Symposium on Information Systems*, pages 1–8.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8*, pages 563–574. Springer.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics.