Training with Adversaries to Improve Faithfulness of Attention in Neural Machine Translation

by

Pooya Moradi

B.Sc., University of Tehran, 2017

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2020

Approval

Name:	**Pooya Moradi**
Degree:	**Master of Science (Computing Science)**
Title:	**Training with Adversaries to Improve Faithfulness of Attention in Neural Machine Translation**
Examining Committee:	**Chair:** Manolis Savva Assistant Professor
	Anoop Sarkar Senior Supervisor Professor
	Angel Chang Supervisor Assistant Professor
	Oliver Schulte Internal Examiner Professor
Date Defended:	**July 17, 2020**

Abstract

Can we trust that the attention heatmaps produced by a neural machine translation (NMT) model reflect its true internal reasoning? We isolate and examine in detail the notion of faithfulness in NMT models. We provide a measure of faithfulness for NMT based on a variety of stress tests where model parameters are perturbed and measuring faithfulness based on how often each individual output changes. We show that our proposed faithfulness measure for NMT models can be improved using a novel differentiable objective that rewards faithful behaviour by the model through probability divergence. Our experimental results on multiple language pairs show that our objective function is effective in increasing faithfulness and can lead to a useful analysis of NMT model behaviour and more trustworthy attention heatmaps. Our proposed objective improves faithfulness without reducing the translation quality and it also seems to have a useful regularization effect on the NMT model and can even improve translation quality in some cases.

Keywords: deep learning; neural network; neural machine translation; interpretability; attention; faithfulness

Dedication

To Mom, Dad, Poorya, and Prof. Anoop Sarkar who treated me like his own son.

Acknowledgements

First and foremost, I would like to express my appreciation toward my dear advisor Prof. Anoop Sarkar. All the contributions in this thesis would not be possible without his supervision, guidance, and full support. His insights have been crucial for completing this work. He also treated me like his own son. I remember I was planing to hitchhike in Mexico and he was threatening to fire me. I am happy that he didn't fire me when I came back alive.

I would like to thank my dear friend and co-author Nishant Kambhatla for his great help in our two papers. I have been lucky to use his experiences, insights, and seniority. Our trip to HongKong for attending EMNLP 2019 became one of my most memorable experiences.

I had an extremely difficult time due to loneliness and depression after coming to Canada from Iran. Surviving from that situation would not be possible without support and companionship of my friends, specially: Hossein Asghari, Saba Akhyani, Pouria Vaziri, Navid Rahimi, Mohsen Katebi, Reza Mirasgar Shahi, and Narges Ashtari.

I shall forever be grateful to my family who have always supported me with their unconditional love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Although neural models have become the standard solutions for solving many challenging tasks including Machine Translation (MT), they are considered as *black-boxes* as their internal computations are not necessarily human-interpretable. In this thesis we have focused on analyzing and improving attention, as an omnipresent component in many neural models, from interpretability perspective. Thus, in Sec. 1.1 we discuss the importance of interpretability in the context of neural models. Then we focus on attention as an interpretation method and briefly mention the notion of faithfulness in Sec. 1.2. In this work we have an encoder-decoder model as our baseline and consequently we provide a brief overview of the encoder-decoder model with attention in Sec. 1.3. In Sec. 1.4 we discuss our contributions and in the end we present an overview of this thesis in Sec. 1.5.

1.1 Interpretability of Neural Models

With advances in sequence-to-sequence (Seq2Seq) models [59], Neural Machine Translation (NMT) systems augmented with attention mechanism [5] have achieved state-of-the-art in many language translation tasks. One shortcoming of NMT models, and neural models in general, is that it is usually difficult for a human interrogator to analyze or understand the true internal reasoning of the neural model for making a particular prediction [15, 18]. The underlying reason behind this difficulty is that information and concepts are represented as real-valued vectors in neural networks. Consequently it's a challenge to interpret these vectors. Why do we want neural models to be interpretable? There can be at least two reasons for this. First of all, to debug a model during error analysis, it is crucial to know how each part of the model is contributing to the prediction and to the error. Moreover, understanding the internal workings of a model is necessary for discovering its deficiencies and improving it. This calls for interpretable neural models and also development of models and methods for understanding and explaining these models. Accordingly, this has led to a wide variety of contemporary NLP research focusing (a) different axes of interpretability including *plausibility* (or interchangeably *human-interpretability*) [19, 31] and *faithfulness* (agreement of an explanation with the internal reasoning of a model) [38, 20], (b) interpretation of the neural model components [7, 12, 64], (c) explaining the decisions made by neural models to humans (using expla-

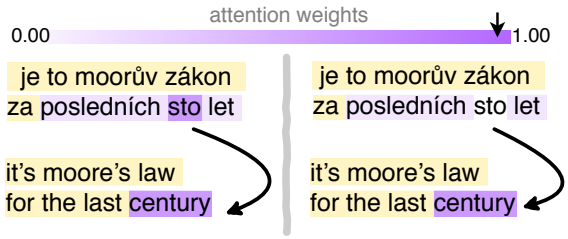


Figure 1.1: An example translation from Cs-En producing *unfaithful* attention weights. The model is generating the token century. In the left attention heatmap, the attention is on the word sto while the decoder generates century. However, in the right heatmap, sto is not attended to at all but century is still produced as the output. This is an example of unfaithful behavior. Yellow words are not attended.

nations, highlights, rationales, etc.) [51, 35, 14, 17, 6, 23], and (d) evaluating different explanation methods from different perspectives [54, 45, 49, 22, 55, 67, 34] which are all discussed in details in Chapter 2.

1.2 Attention Interpretability and Faithfulness

The advent of attention in neural models has improved the overall accuracy in many different tasks. Attention has become an omnipresent component in neural machine translation models and more generally in the architecture of neural models in NLP. The attention mechanism provides a probability distribution over the input with respect to a state variable such as decoding state in NMT. This probability distribution is used to summarize source-side information into a context-vector and is fed to the decoder as an additional signal for prediction. Along with their use in improving performance across different tasks, attention weights are widely implicitly or explicitly used to explain the predictions made by neural models [14, 16, 17]. The general idea is that attention weights can be indicators of the importance of inputs for producing a particular prediction. However, whether or not attention is a reliable source for explanation is often taken for granted. Consequently, there has been a great interest recently in the community in investigating the credibility of attention as explanation [22, 67, 55].

In this work we have focused on the faithfulness of interpretations offered by attention, that is the extent to which the model's internal reasoning process is actually based on that interpretation. Faithfulness of these interpretations are particularly important for NLP practitioners who wish to debug their neural models and improve them. Identification of the faults of a neural model cannot be done if the neural model is not providing a faithful and trustworthy description of what it is doing. Jacovi and Goldberg [20] emphasize distinguishing faithfulness from human-interpretability in interpretability research by providing several clarifications about the terminology used by researchers. They describe the following conditions on the evaluation of how well a research project tackles the notion of faithfulness:

- Be explicit: provide a measurable evaluation of faithfulness.
- Human judgements are not relevant because we are interested in model internals.
- Do not match against gold labels (e.g. AER) because faithfulness of both correct and incorrect decisions made by the model are equally important.
- No model is “inherently" faithful. We need to measure faithfulness not as a binary aspect of a model (it is faithful or not) but rather as a gray-scale measure.
- A more faithful system is a necessary but not sufficient condition for model interpretation by humans, c.f. [21].

Aligned with these criteria, we study faithfulness of NLP neural models, specifically NMT models. We provide a faithfulness measure that is computed based on a variety of stress tests where model parameters are perturbed and measuring how often the model output changes (Figure 1.1). Our findings show that our objective is effective in increasing faithfulness and can lead to a useful analysis of NMT model behaviour and more trustworthy attention heatmaps. We assert that faithfulness is a good property to have in a model whether or not it will be useful for downstream interpretation. A model that is faithful can be more trusted as a component in a larger end-to-end neural model.

1.3 Encoder-Decoder Model with Attention

Given a training sentence pair $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} = [x_1, x_2, ..., x_m]$ is a sentence in the source language and $\mathbf{y} = [y_1, y_2, ..., y_n]$ is its corresponding translation in the target language, the problem of neural machine translation is feeding $\mathbf{x}$ to a neural network and getting $\mathbf{y}$ as the output. The model we use in this work is called "encoder-decoder" model with attention [59, 4]. The idea is that the encoder, which is a recurrent neural network (RNN), runs over the source sentence to calculate the contextualized representation of the source sentence. Then a second neural network which is the decoder "decodes" this information into the target sentence. The attention mechanism is employed to calculate the contextualized representation of the source sentence dynamically based on the decoding step (Figure 1.2).

To be more specific, we use a bidirectional encoder, and concatenate the forward and backward hidden states to build the final representation.

$$\begin{aligned} \overrightarrow{h_t} &= \overrightarrow{f_{enc}}(x_t, \overrightarrow{h_{t-1}}) \\ \overleftarrow{h_t} &= \overleftarrow{f_{enc}}(x_t, \overleftarrow{h_{t+1}}) \\ h_t &= [\overrightarrow{h_t}, \overleftarrow{h_t}] \end{aligned} \tag{1.1}$$

Then the decoder generates output tokens using the following probability distribution:

$$p(y_t|y_{<t}, x) = \textit{softmax}(g_{dec}(s_t, c_t))$$

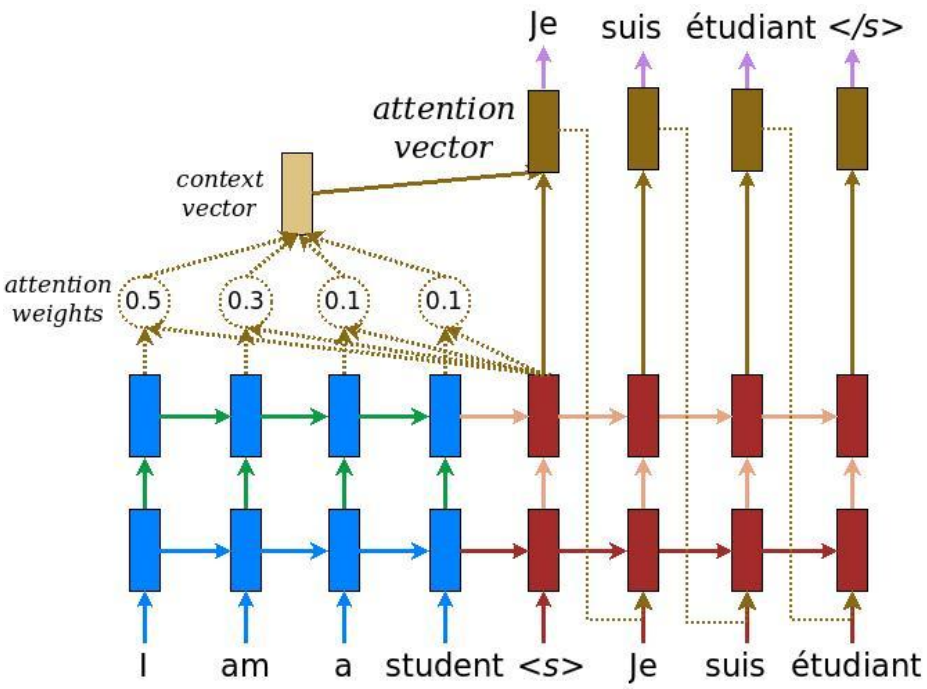


Figure 1.2: An encoder-decoder model with attention mechanism. Image from `https://www.tensorflow.org/tutorials/text/nmt_with_attention`

with g_{dec} being a transformation function that produces a vocabulary-sized vector, and s_t is the hidden unit of the decoder's RNN:

$$s_t = f_{dec}(y_{t-1}, s_{t-1}, c_{t-1})$$

where f_{dec} is a RNN. Here c_t is the context vector calculated by attention mechanisms:

$$c_t = \sum_{i=1}^{m} \alpha_{t,i} h_i$$

where α_t is the normalized attention weights over the source context:

$$\alpha_{t,i} = \frac{e^{a(s_t, h_i)}}{\sum_j e^{a(s_t, h_j)}}$$

Here, a is a scoring function that determines the contribution of each source context vector to the final context vector. Implementation of a depends on the choice of the attention method. In this work, we use general attention [41] as the scoring function:

$$a(s_t, h_i) = s_t^\top W_a h_i$$

1.4 Contributions

We seek to improve faithfulness of NMT models. To this end, we make the following contributions in this work:

- We propose a measure for quantifying faithfulness in NMT.
- We introduce a novel learning objective based on probability divergence that rewards faithful behavior and which can be included in the training objective for NMT.
- We provide empirical evidence that we can improve faithfulness in an NMT model. Our approach results in more a more faithful NMT model while producing better BLEU scores. Most previous work has focused on document or sentence-based classification tasks where attention models are not as directly useful as in NMT models.

We chose to study the impact of faithfulness in NMT because it is under-studied in terms of interpretability. Most previous work has focused on document or sentence-based classification tasks where attention models are not as directly useful as in NMT models. Attention is also more challenging in terms of faithfulness in the context of NMT models due to the substantial impact of the decoder component.[1]

1.5 Overview

This thesis addresses the problem of faithfulness of attention in NMT. First we propose how to measure faithfulness in NMT using proposed adversarial attention as stress tests. Secondly we propose a novel method to improve faithfulness in NMT.

in **Chapter 2** we present previous works on different axes of interpretability, from interpretability methods to interpretability of attention and efforts to build inherently interpretable models.

in **Chapter 3** we demonstrate our proposal for measuring faithfulness in NMT. Our novel method for improving faithfulness in NMT is also discussed. We also mention attention sparsity as an attempt for improving faithfulness.

in **Chapter 4** we discuss our findings. We illustrate behavior of NMT in the presence of stress tests. Moreover, we show the effect of our method and attention sparsity on the faithfulness.

We conclude our work in **Chapter 5**.

[1] We focus on RNN based encoder-decoder models. While Transformers [63] generally produce better NMT models, in order to replace the long distance dependencies in a gated RNN, a Transformer model relies on multiple heads of attention and self-attention. Before we can tackle multi-head attention, we focus on the simpler single-head attention models and try to understand them in terms of faithfulness.

Chapter 2

Related Work

Contemporary research on interpretability of neural models cover diverse topics ranging from different interpretability methods to designing inherently interpretable neural models. In this chapter we present recent research related to our work. In Sec. 2.1 we discuss several methods for analyzing and interpreting neural models. In this work we have focused on interpretability of attention component and thus we review prior studies on understanding the semantics captured by attention in Sec. 2.2. There has been extensive research on investigating attention as an interpretation method from different axes ranging from faithfulness and plausibility to accountability and fairness. Those works are reviewed in Sec. 2.3. In this work we have also analyzed effect of sparsity for improved faithfulness. Consequently we have reviewed works regarding sparsity for better interpretability in Sec. 2.4. Our proposed method for improving faithfulness can be seen as an explanation regularizer. Several previous works have investigated regularizing explanations for better interpretability and we have reviewed them in Sec. 2.5. In the end we review prior attempts on designing inherently interpretable neural models in Sec. 2.6.

2.1 Interpretability Methods

Relevance-based interpretation is a common technique in analyzing predictions in neural models. In this method, inputs of a predictor are assigned a scalar value quantifying the importance of that particular input on the final decision. Saliency methods use the gradient of the inputs to define importance [36, 17, 13]. Layer-wise relevance propagation that assigns relevance to neurons based on their contribution to activation of higher-layer neurons is also investigated in NLP [2, 14, 3]. Another method to measure relevance is by removing the input, and tracking the difference in the network's output [37]. While these methods focus on explaining a model's decision, Shi et al. [56], Kádár et al. [24], Calvillo and Crocker [8] investigate how a particular concept is represented in the network.

2.2 Understanding Information Captured by Attention

Analyzing and interpreting the attention mechanism n NLP is another direction that has drawn major interest. Koehn and Knowles [28] compares alignment in NMT extracted by attention with those of fast-align and argue that attention cannot be reliably used as word alignment between source and target words, at least in the traditional sense in statistical machine translation. Ghader and Monz [16] also verifies this however they show that attention is capturing useful information other than alignment. Tang and Nivre [60] investigates the role of attention in the case of word sense disambiguation (WSD) in NMT models. Counterintuitively, they show that attention pays more attention to the ambiguous noun itself and in fact encoder hidden states are handling WSD. Clark et al. [10] studies the information captured by each head in a BERT model. They conclude that some heads correspond well to linguistic notions of syntax and co-reference. Vig and Belinkov [65] focuses on structure of attention in a GPT-2 model. By visualizing attention for individual instances, they show that attention targets different part-of-speech at different layers and dependency relations are mostly captured in middle layers.

2.3 Attention and different axes of interpretability

While several studies have focused on understanding the semantic notions captured by attention [16, 64, 10], evaluating attention as an interpretability approach has garnered a lot of interest. From the faithfulness perspective, Jain and Wallace [22], Serrano and Smith [55] show that for instances in a data set there can be adversarial attention heatmaps that do not change the output of the text classifier. In other words, adversarial attention leads to no decision flip in each instance. They use this to claim that attention heatmaps are not to be trusted, or unfaithful. Wiegreffe and Pinter [67] argue against per-instance modifications at test time for two reasons: 1) in classification tasks attention may not be useful so perturbing attention is misleading. This is not true for NMT since attention is very useful in NMT. 2) they train an adversarial attention model (e.g. uniform attention) chosen to produce attention weights distant from the original attention weights while at the same time trying to minimize classification error. They show that such adversarial attention models are not as accurate as models with attention. In our work we acknowledge that attention is useful and faithful to some extent and we aim to improve faithfulness of NMT models.

While most of these works provide evidence that attention weights are not always faithful, Moradi et al. [46] confirm similar observations on the unfaithful nature of attention in the context of NMT models. Li et al. [34] is one of the few papers examining attention models in NMT. However, they are focused on the task of identifying relevant source words to explain the output translations selected by the NMT model. They look for optimal proxy models that agree with the NMT model such that the relevant source words picked as an explanation by a proxy model exhibits similar behaviour to the target model. They use the notion of fidelity over proxy models and evaluate several alternative proxy models using empirical risk minimization. Attention weights are evaluated

alongside other proxy models for this task. In contrast, our work is about improving the faithfulness of NMT models and we focus on the internal state of the NMT model rather than proxy models. They use human references, e.g. AER, for evaluating fidelity. As discussed earlier, evaluation of faithfulness cannot involved human judgements or reference data. It is possible that our faithful NMT models are also better at fidelity, but that is an open question.

While prior works have mostly failed to explicitly distinguish faithfulness from plausibility in their arguments, Jacovi and Goldberg [21, 20] focus on formalizing faithfulness and addressing evaluation of faithfulness separately from plausibility respectively.

Subramanian et al. [58] have investigated the concept of faithfulness in neural modular networks (NMN) which are employed for modeling compositionality. They question the faithfulness of the structure of the network modules describing the true abstract reasoning of the model. Similar to us, they attempt to quantify faithfulness and improve upon it. However their contributions like training with an auxiliary atomic-task supervision for improved faithfulness are specific to the context of NMNs.

Pruthi et al. [50] demonstrate that it is possible to train a model that produces a deceptive attention mask, questioning the use of attention weights as explanation from the fairness and accountability perspective.

Alvarez-Melis and Jaakkola [1] investigate the interpretability methods from the robustness perspective. They attempt to quantify robustness and show that current interpretability methods cannot be considered as robust.

2.4 Sparsity for Improved Interpretability

This line of work suggests making attention sparser so that the most contributing input word is more distinguishable over other input words. Martins and Astudillo [43] propose sparsemax as an alternative to the traditional softmax activation function, but able to output sparse probabilities and at the same time being differentiable. Malaviya et al. [42] improves sparsemax by proposing a constrained sparsemax for attention that can model fertilises to the source words and at the same time being sparse and differentiable. Zhang et al. [68] propose sparsity regularization terms such as entropy regularization to promote sparsity in the attention.

2.5 Regularizing Explanations

Recent work on explanations for black-box models has produced tools (e.g. LIME [51]) to show the implicit rules behind predictions, which can help us identify when models are right for the wrong reasons. However, these methods do not scale to explaining entire datasets and cannot correct the problems they reveal. Ross et al. [53] introduce a method for efficiently explaining and regularizing differentiable models by examining and selectively penalizing their input gradients. Rieger et al. [52] follow a similar spirit but they use 'contextual decomposition' [47] to extract explanations offered

by the model. Aligning attention (as explanation) with prior knowledge has also been extensively studied. Mi et al. [44], Liu et al. [40] propose to supervise attention with traditional alignment models so that attention weights match better with alignment. Zhong et al. [69] show that in the task of textual sentiment classification, attention is often misaligned with the words that contribute to attention. They propose to supervise attention with human rationale during training and they observe improved model performance. Cohn et al. [11] demonstrate that by including structural biases from traditional alignment models like positional bias and fertility in attention, it's possible to improve the existing NMT baselines.

2.6 Self-explanatory Neural Models

Contrary to efforts for propose post-hoc explanation methods for neural models, a series of works have attempted to make neural models inherently interpretable or self-explanatory. Stahlberg et al. [57] show that the NMT model can be made self-explanatory by training it to produce the discrete decisions made by the model (from which the translations can be extracted later). Lei et al. [33] propose a model in which first a rationale is selected from the input and then is further used for prediction. Their proposed model consists of a generator and an encoder, which are trained to operate together. The generator specifies a distribution over the input to be used as candidate rationales and these are passed to the encoder for prediction. Previous work proposed to assign binary latent masks to input positions and to promote short selections via sparsity-inducing penalties such as L0 regularisation. Instead Bastings et al. [6] propose a latent model that mixes discrete and continuous behaviour allowing at the same time for binary selections and gradient-based training without REINFORCE. Instead of selecting part of the input as the rationale, Liu et al. [39] propose a generative explanation framework.

Chapter 3

Faithfulness in NMT

In this thesis we have two major contributions: a) quantifying faithfulness in NMT b) improving faithfulness using a novel objective. In Sec. 3.1 we explain our approach for quantifying faithfulness in NMT by putting the model under various stress tests and capturing its behavior. Then in Sec. 3.2 we present our proposed method for improving faithfulness based on a novel objective that rewards faithful behavior. We also talk about our motivation for investigating effect of sparsity on faithfulness.

3.1 Measuring Faithfulness

Intuitively, a faithful explanation should reflect the true internal reasoning of the model. Although there is no formal definition for faithfulness, a common approach in the community is to design stress tests to perturb the most relevant parts of the input, suggested by the explanation, in expectation that the model's decision should change [20]. A common stress test is the erasure test in which the most-relevant part of the input is removed [3]. In the context of NMT, at decoding time step t the attention component assigns attention weights α_t, attending to the source word at position $m_t = \text{argmax}_i \, \alpha_{t,i}$ (or the k-best attended-to words in the source). These weights are often implicitly or explicitly regarded as an interpretation for the model's prediction at the time step t [61, 44, 40, 66, 32, 14, 17]. It is worth noting that erasure is only one of the possible stress tests for evaluating faithfulness. Passing more stress tests implies a more faithful model as it is resilient to more attacks or stress tests of its faithfulness. In this paper we consider three intuitive stress test cases:

- **ZeroOutMax** [3]: Here we remove attention from the most important token according to the attention weights by setting $\alpha_{t,m_t} = 0$.
- **Uniform** [46]: In this stress test all attention weights are set to be equal, $\alpha_t = \frac{1}{m}\vec{1}$, where m is the length of the source sentence. This is to confuse the model about which part of the input is the most important one.
- **RandomPermute** [22]: In this stress test we randomly permute attention weights several times until a change in the model output is observed. We ensure that m_t, the most important

token according to attention, is always changed. We set $\alpha'_t = \texttt{random\_permute}(\alpha_t)$ such that $\text{argmax}_i\, \alpha'_{t,i} \neq m_t$

Many prior studies of attention [22, 67] have used a binary measure: either attention is faithful or it is not. These studies typically are about whether attention has the potential to be useful in terms of accuracy and faithful in terms of model behaviour. In many cases, especially in the case of NMT models, attention is clearly useful and by and large it must be faithful. The question is can we measure the faithfulness and improve faithfulness. It is more natural to have a gray-scale notion of faithfulness for evaluation [20]. Following this reasoning, we define $F(M)$ as faithfulness of attention heatmaps in model M as:

$$F(M) = \frac{\text{\# of tokens passing stress tests}}{\text{\# of tokens}} \tag{3.1}$$

$F(M)$ is a number between 0 to 1 measuring the percentage of output tokens during inference which passed the stress tests, i.e., they changed in the presence of adversarial attention. This metric can also be regarded as a measure of trust we can assign to the attention heatmap to fully reflect the internal reasoning of the NMT model.

3.2 Improving Faithfulness

The conventional objective function in a sequence-to-sequence task is a cross-entropy loss $\mathcal{F}_{acc}$ which should be minimized :

$$\mathcal{F}_{acc}(\theta) = -\frac{1}{|S|} \sum_{(X,Y)\in S} \log p(Y|X;\theta) \tag{3.2}$$

where S is the training data and X and Y are source sentence and the correct translation respectively (We use capital letters for sentences and small letters for single tokens). This training objective does not explicitly model the interpretability aspects (e.g. faithfulness) of the network and it remains unoptimized during training.

Faithfulness Objective In an effort to develop a model that is right for the right reason, Ross et al. [53] change the loss function of their classifier to model both right answers and right reasons instead of only the former. They achieve this by introducing a regularizing term that tends to shrink irrelevant gradients. In a similar spirit, we change our objective to account for the NMT model's faithfulness as well as the cross-entropy score against the reference translations:

$$\mathcal{F} = \mathcal{F}_{acc} + \lambda_{faith}\mathcal{F}_{faith} \tag{3.3}$$

$\mathcal{F}_{faith}$ is an additional component that rewards the model for having more faithful attention. The parameter λ_{faith} regulates the trade-off between between faithfulness and accuracy objectives.

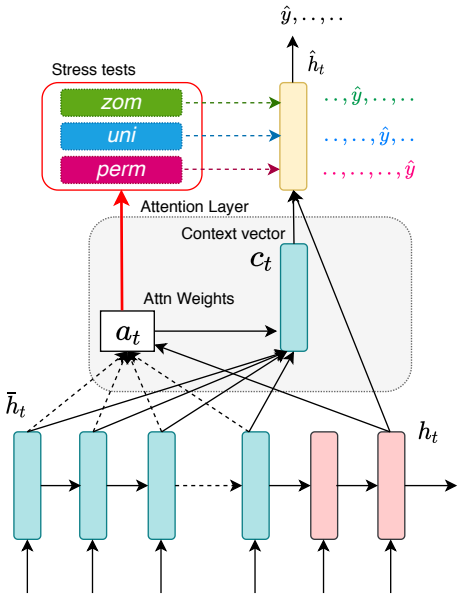


Figure 3.1: We generate adversaries to the attention weights using various stress tests *Uniform, ZeroOutMax,* and *RandomPermute*. When adversarial attention weights are used, in a faithful model we expect the probability of the original output ($\hat{y}$) to drop significantly. We use this criteria to define a faithfulness objective function.

3.2.1 Divergence-based Faithfulness Objective

Consider a predictive model g_θ in which an intermediate calculation is later employed to justify predictions:

$$\hat{y} = \arg\max_y p(y|x) = \arg\max_y g_\theta(x, IC(x), y) \tag{3.4}$$

where $IC(x)$ is the intermediate calculation on the input. A concrete example for $IC(x)$ would be the context vector calculated by the attention mechanism.

Hypothesis If there exists an intermediate calculation $IC'(x)$, as a stress test, that conveys a contradictory post-hoc attention compared to $IC(x)$, then $IC(x)$ cannot be regarded as faithful for predicting $\hat{y}$. If $IC(x)$ is faithful, we expect the model to diverge from predicting $\hat{y}$ when $IC'(x)$ is employed instead.

Based on our hypothesis, we propose a divergence-based objective which mimics behavior of a faithful explanation under stress test:

$$\mathcal{F}_{faith} = \log p(\hat{y}|x, IC'(x)) \tag{3.5}$$

This objective is a negative loss that should be minimized. The minimum of this objective is achieved when the probability of the original prediction approaches zero under the stress test which

is the ideal. Thus, it promotes reduction in output probability under an adversarial intermediate calculation (Figure 3.1). It is worth noting that this objective can be potentially employed in models where outputs are modeled as soft probabilities and thus is not limited to NMT. To put a model under various stress tests we manipulate the context vector during training time by changing the attention weights and feeding it to the decoder to calculate the probability. More precisely:

$$\begin{aligned}\mathcal{F}_{faith} = &\lambda_{zom} \log p(\hat{y}|x, IC'_{zom}(x)) \\ &+ \lambda_{uni} \log p(\hat{y}|x, IC'_{uni}(x)) \\ &+ \lambda_{perm} \log p(\hat{y}|x, IC'_{perm}(x))\end{aligned} \tag{3.6}$$

where IC'_{zom}, IC'_{uni} and IC'_{perm} are *ZeroOutMax*, *Uniform* and *RandomPermute* methods (see Sec. 3) to manipulate attention weights, respectively. $\lambda_{\{method\}}$ parameters regulate the contribution of each objective. We use the term $\mathcal{F}_{all}$ when all $\lambda_{\{method\}}$s in Eq. (3.6) are non-zero. Moreover, we use the term $\mathcal{F}_{\{method\}}$ when $\lambda_{\{method\}}$ is set to 1 and other regularization weights are zero.

3.2.2 On attention sparsity

Do the models trained with the faithfulness objective have sparser attention weights? Sharper attention in a model M might correlate with an intensified contribution of the most-attended source hidden state on the prediction resulting in higher faithfulness.

To measure sparseness of the attention, we take an average over the normalized entropy of attention distribution for each output token during inference on test data. We use normalized entropy which is in range [0,1] to account for the fact that the range of the entropy for each output token depends on the length of the corresponding source sentence.

$$\textit{AvgEnt} = \frac{1}{\sum_{i=1}^{|S|} \left|\hat{Y}_i\right|} \cdot \sum_{i=1}^{|S|} \sum_{j=1}^{\left|\hat{Y}_i\right|} \textit{NormEnt}(\alpha_{i,j}) \tag{3.7}$$

$$\textit{NormEnt}(P) = -\sum_{i} \frac{P_i \log P_i}{\log N} \tag{3.8}$$

Here α_{ij} is the attention distribution for the output token j in the generated translation of source sentence i, and P is a discrete probability distribution. In Eq. (3.8) low entropy indicates a sharper distribution.

Attention Entropy Regularization Alongside investigating sparsity of the models trained by the faithfulness objective, we also train a model in which sparsity in attention is directly optimized. We used attention entropy regularization [68]:

$$\mathcal{F}_{ent} = \mathcal{F}_{acc} + \lambda_{ent} \sum_{i=1}^{|S|} \sum_{j=1}^{\left|\hat{Y}_i\right|} Ent(\alpha_{i,j}) \tag{3.9}$$

where entropy of attention weights is added to the cross-entropy loss (3.2) as a regularization term.

3.2.3 Summary

In this chapter we have presented our approach for quantifying faithfulness and also improving it. We put NMT under different stress tests and define faithfulness based on the percentage of output tokens that have been preserved under those tests. We argue that faithfulness is not optimized in usual NMT objectives and we propose a novel loss based on probability divergence that rewards faithfulness behavior. Moreover we mention our motivation for experimenting with entropy regularization which is investigating if attention sparsity is related to faithfulness. In Chapter 4 we discuss our experimental setup, experiments, and findings.

Chapter 4

Experiments

In this chapter we provide results of our experiments and findings. We first explain our experimental setup including the NMT model employed, datasets used, hyperparameters and training tricks in Sec. 4.1. We also study the faithfulness of our baseline NMT model and its sensitivity to different stress tests in Sec. 4.2. In Sec. 4.3 we analyze our proposed method and its effect on faithfulness and sparsity. We provide a brief summary of this chapter in Sec. 4.4.

4.1 Experimental Setup

Data We use the Czech-English (Cs-En) dataset from IWSLT2016[1] and the German-English (De-En) dataset from IWSLT2014[2]. For the Czech-English dataset we use dev2010, tst2010, tst2011, tst2012, and tst2013 as the test data. For the German-English dataset we use dev2010, tst2010, tst2011, dev2012, and tst2012 as the test data. Table 4.1 shows the number of instances in different sets from the datasets used. We used Moses [29] to tokenize the dataset.

	Cs-En	**De-En**
Train	101225	160239
Valid	4601	7283
Test	5716	6750

Table 4.1: Number of sentences in training, validation, and test sets across Cs-En and De-En datasets.

Architecture and Hyperparameters We use OpenNMT [26] as our translation framework. We employ a 2 layer LSTM-based encoder-decoder [59, 9] model with global attention [41]. Dimension of the hidden states and the word embeddings for both source and target languages are set to 500. Vocabulary size for both the source and target language is set to 50000. We remove sentences with more than 50 tokens from the training data. We use Adam [25] for training our models and we

[1] https://sites.google.com/site/iwsltevaluation2016/

[2] https://sites.google.com/site/iwsltevaluation2014/

	Cs-En	De-En
Number of tokens (+EOS)	115993	139465
% of function words	68%	68%
% of content words	32%	32%

Table 4.2: Percentage of function and content words in the generated translation for test set.

set the learning rate to 0.001. Models are trained until convergence. Our models have around 82M parameters and it took us twelve hours to train each model on two GTX 1080ti GPUs. We optimize the hyperparameters of our models using the validation set. The baseline model is trained using Eqn. (3.2) and we call it $\mathcal{F}_{baseline}$. λ_{ent} in Eq. (3.9) is set to 0.04. We refer to the objective as $\mathcal{F}_{all}$ when λ_{zom}, λ_{uni}, and λ_{perm} are set to 0.5, 0.375, and 0.125 respectively. λ_{faith} is set to 1.

Training Difficulties Our first attempts at using the modified objective function in Eq. (3.3) trained poorly. We observed that it was difficult for the model to learn the faithfulness constraint without having already learned to assign a reasonable probability to correct translations. To address this problem we first train the NMT model using the standard unmodified objective function and then fine-tune this trained model by switching the objective function to Eq. (3.3).

4.2 Analyzing the baseline model

4.2.1 Power of each test

We investigate behavior of the model under stress test for generation of function[3] and content words. Function words (e.g., *a*, *the*, *is*) have little lexical meaning in contrast to content words and thus we are curious whether response of the model to stress tests differs for generation of these two groups of words. Table 4.2 shows the percentage of function and content words generated by the baseline model. As expected, the majority of the generated tokens are function words.

Table 4.3 and 4.4 show the faithfulness of the model for generation of function and content words under different stress tests. `ZeroOutMax` test has been the most effective method for capturing unfaithful behavior as it has resulted in the lowest faithfulness. We also determine that `RandomPermute` is not as effective as the `Uniform` and `ZeroOutMax` methods. Our justification is that in the `RandomPermute` method, it is highly probable that the context vector is biased toward a random source hidden state. Such bias can lead to significant misleading noise in the context vector which can change the prediction of the model. However, there isn't such a bias in the `Uniform` or `ZeroOutMax` methods.

As evident from Table 4.3 and 4.4, the most strict test is when all stress tests are applied for capturing unfaithful behavior (*All* column).

[3]The reference for function words (we added new function words including the EOS token to this) can be found at: `semanticsimilarity.files.wordpress.com/2013/08/jim-oshea-fwlist-277.pdf`

Objective	Content Words				Function Words			
	ZOM	*Uniform*	*RandPerm*	*All*	*ZOM*	*Uniform*	*RandPerm*	*All*
$\mathcal{F}_{baseline}$	83%	90%	94%	78%	46%	48%	64%	33%
$\mathcal{F}_{zom}$	91%	93%	98%	86%	84%	87%	95%	74%
$\mathcal{F}_{uni}$	84%	98%	97%	83%	56%	98%	91%	54%
$\mathcal{F}_{perm}$	86%	95%	96%	83%	74%	97%	98%	71%
$\mathcal{F}_{all}$	**91%**	**99%**	**98%**	**89%**	**83%**	**98%**	**98%**	**82%**
$\mathcal{F}_{ent}$	78%	90%	94%	73%	46%	48%	64%	33%

Table 4.3: Faithfulness metric for the generated content and function words through different objectives in the **Czech-English** dataset. The columns are different tests included in the Eq.(3.1).

Objective	Content Words				Function Words			
	ZOM	*Uniform*	*RandPerm*	*All*	*ZOM*	*Uniform*	*RandPerm*	*All*
$\mathcal{F}_{baseline}$	81%	90%	93%	76%	45%	48%	64%	32%
$\mathcal{F}_{zom}$	91%	95%	98%	87%	87%	95%	97%	82%
$\mathcal{F}_{uni}$	81%	98%	91%	80%	60%	100%	95%	58%
$\mathcal{F}_{perm}$	85%	95%	97%	82%	74%	97%	98%	72%
$\mathcal{F}_{all}$	**91%**	**98%**	**98%**	**89%**	**87%**	**100%**	**99%**	**86%**
$\mathcal{F}_{ent}$	81%	90%	93%	76%	47%	47%	64%	33%

Table 4.4: Faithfulness metric for the generated content and function words through different objectives in the **German-English** dataset. The columns are different tests included in the Eq.(3.1).

4.2.2 Function words are more easily generated compared to content words

An important observation in Table 4.3 and 4.4 is that function words exhibit unfaithful behavior much more than content words. Faithfulness of the model for generation of content words is 78% and 76% for Czech-English and German-English respectively. However it is 33% and 32% for function words. The reason is that The production of function words rely more on the target context, in contrast to content words which rely more on the source context [62]. Accordingly, perturbation in the original attention weights likely has significantly more impact on diminishing content words compared to function words. This ties well with the main idea behind context gates in which the influence of source context and target context is controlled dynamically [62].

4.2.3 Highlighting top preserved tokens

To better understand the behavior of the model in the presence of stress test, we listed the top preserved tokens in the De-En dataset. Table 4.5 contains the top 20 content words sorted by the number of times they were preserved. It is interesting to note that for many of these frequent tokens, more than half of their total occurrences are preserved without focusing on their corresponding translation in the source sentence (e.g., “going”, “know”, “thing”, etc). In Table 4.6, we sort such tokens based

Token	# preserved	Coverage
going	310	70%
people	237	46%
know	219	62%
world	215	67%
like	189	47%
think	176	50%
way	162	68%
get	160	53%
thing	147	79%
things	142	56%
time	139	54%
see	137	51%
years	136	64%
make	126	49%
little	113	55%
just	109	29%
really	93	37%
bit	92	88%
said	89	59%
got	86	59%

Table 4.5: Top 20 content words preserved by the aggregate method sorted by the number of times they were preserved.

Token	Coverage	Total
bit	88%	105
course	87%	91
thank	83%	89
thing	79%	186
fact	78%	74
half	78%	27
own	75%	75
ones	73%	30
states	73%	30
difference	71%	21
going	70%	444
turns	69%	26
way	68%	237
able	67%	85
world	67%	323
doing	66%	103
planet	65%	37
years	64%	212
know	62%	353
united	62%	21

Table 4.6: Top 20 content words preserved by the aggregate method sorted by percentage of their total occurrences that are preserved (*coverage*).

Token	# preserved	Coverage
,	7329	85%
EOS	6364	94%
the	5210	82%
.	3947	60%
of	3003	87%
to	2923	86%
and	2639	67%
a	2187	65%
that	1936	69%
i	1737	76%
's	1732	95%
you	1501	72%
it	1497	72%
is	1496	88%
in	1364	64%
we	1246	64%
they	624	69%
"	620	81%
have	613	70%
be	582	91%
't	580	96%
're	542	86%
this	541	42%
so	531	57%
are	526	77%
was	514	66%
do	433	77%
about	417	65%
what	415	61%
can	400	54%

Table 4.7: Top 30 function words preserved by the aggregate method sorted by the number of times they were preserved.

Token	Coverage	Total
't	96%	602
's	95%	1819
EOS	94%	6748
be	91%	641
is	88%	1707
of	87%	3450
to	86%	3383
're	86%	631
,	85%	8582
'm	84%	311
been	82%	233
lot	82%	148
the	82%	6386
"	81%	770
are	77%	679
do	77%	565
i	76%	2290
who	73%	300
it	72%	2089
you	72%	2099
have	70%	876
up	70%	235
they	69%	904
that	69%	2812
well	67%	153
and	67%	3922
was	66%	774
were	65%	240
same	65%	154
a	65%	3369

Table 4.8: Top 30 function words preserved by the aggregate method sorted by coverage.

on their coverage, which is the percentage of their total occurrences that are not affected when a counterfactual attention is applied[4].

We repeat the same process for function words (Table 4.7 and Table 4.8) . As evident from Table 4.7, we have successfully yielded the same token in 94% of the occurrences of the EOS token but with a counterfactual attention. This can be explained by the previous findings suggesting special hidden units keep track of translation length [56]. As a result, the EOS token is generated upon

[4]We consider only the tokens that have appeared more than 20 times. The reason is that there are many preserved words that have appeared only once (coverage=1) and it is not clear if the coverage remains the same when frequency increases.

receiving signal from these units rather than using attention. This indicates that attention weights are highly unreliable for explaining the generation of EOS tokens. This is worth noting because early generation of the EOS token is often a major reason of the under-translation problem in NMT [30]. Thus, attention weights should not be used to debug early generation of EOS, and that some other underlying influence in the network [14] might be responsible for the model's decision in this case.

4.3 Analyzing the proposed methods

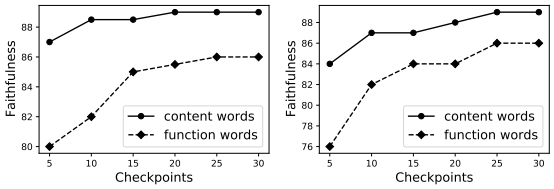
4.3.1 Impact on faithfulness

To measure the effectiveness of the proposed objectives, we choose the best model in terms of provided faithfulness but within the 0.5 BLEU score of the maximum achieved BLEU score in the validation set. The reason is that we prefer a model that is both accurate and with faithful attention-based explanations. Table 4.3 and 4.4 show the performance of the different faithfulness objective functions when generating content words and function words across different attention manipulation methods in the Czech-English (Cs-En) and German-English (De-En) datasets respectively. Results indicate that the proposed divergence-based objective has been effective in increasing the faithfulness metric. $\mathcal{F}_{all}$ is the most effective objective for increasing faithfulness when all stress tests are included in Eq. (3.1). When using $\mathcal{F}_{all}$, faithfulness of attention-based explanations for content words is increased 78% to 89%, while that of the function words is from 33% to 82%(see *All* column in Table 4.3). There are similar increases from 76% to 89% for content words and from 32% to 86% for function words in the De-En dataset. These results establish the effectiveness of our proposed objectives to increase the faithfulness metric. It is worth noting that increase in faithfulness of attention-based explanations for function words is much more than that of content words. This can be attributed to the fact the function words are mostly generated using the target-side information in the decoder [62, 46] and manipulating attention does not have much effect on generating them. However, our proposed faithfulness objective ($\mathcal{F}_{faith}$) seems to tighten the dependence of the decoder on the attention component. This results in much more increase in faithfulness for function words compared to such content words.[5] We also plot faithfulness over different checkpoints in Figure 4.1. It indicates that progress in faithfulness is much faster for function words compared to content words.

4.3.2 Effect of training with single adversary on passing other stress tests

An interesting observation in Table 4.3 and 4.4 is that training with an adversary has positive effects on the model for passing stress tests from other types of adversaries. As an example, in Table 4.3 the

[5] If this dependence is not desired, it is possible not to penalize function words in the faithfulness objective. However, relying on attention for generating function words can be helpful, not necessarily for interpretability but for dealing with long-range dependencies [63] and, as a result, better translations.



(a) Cs-En (b) De-En

Figure 4.1: Progress in faithfulness over different checkpoints. It increases much faster in function words compared to content words.

	De-En		**Cs-En**	
Tag	Baseline	Ours	Baseline	Ours
PUNC	0.19	**0.70**	0.28	**0.66**
PRON	0.42	**0.78**	0.35	**0.75**
VERB	0.47	**0.80**	0.50	**0.81**
ADP	0.30	**0.75**	0.40	**0.65**
DET	0.35	**0.74**	0.38	**0.70**
PRT	0.13	**0.63**	0.17	**0.50**
ADV	0.66	**0.80**	0.63	**0.79**
NOUN	0.63	**0.87**	0.64	**0.85**
ADJ	0.68	**0.87**	0.69	**0.85**
NUM	0.84	**0.86**	0.79	**0.86**
X	0.67	**0.78**	0.55	**0.80**

Table 4.9: Faithfulness metric within different part-of-speech (POS) tags.

column *Uniform* is the faithfulness metric when only Uniform test is employed in Eq. (3.1). When using this metric, we can observe that training a model with $\mathcal{F}_{perm}$ increased faithfulness from 90% to 95% for content words and from 48% to 97% for function words. We can see such effect in Table 4.4 as well. This observation indicates that training with each adversary can be beneficial for making model tolerant against other types of stress tests. It seems that training with each adversary strengthens the dependence of the decoder on the attention component which can be beneficial for passing other stress tests.

4.3.3 POS-tag analysis

In addition to categorizing tokens into function and content words, we also analyze the effect of our proposed objective within different universal part-of-speech (POS) tags [48] in Table 4.9. Our proposed objective has increased faithfulness in each POS tag and in our both datasets. Tokens with less lexical meaning are the ones affected the most as explained in Sec. 4.3.1. As expected, punctuations (PUNC) and particles (PRT) tags have benefited the most from increase in the faithfulness. Interestingly numbers (NUM tag) have the lowest increase in faithfulness. One reason might be that they already had a high initial faithfulness and this has made further increase less likely.

4.3.4 Regularization Effect

The model checkpoints used in Tables 4.3 and 4.4 were selected based on maximum increase in faithfulness without sacrificing accuracy. To investigate if the proposed objective can have a general positive side effect in terms of accuracy, we train three independent models using the $\mathcal{F}_{baseline}$ and $\mathcal{F}_{all}$ objectives. To make it fair for the baseline, we also add additional steps of training for the baseline model as well to isolate the benefit of adding the faithfulness objective.

Table 4.10 contains the average BLEU score of the trained models. It indicates that the model trained with $\mathcal{F}_{all}$, has +0.7 and +0.4 increase in BLEU score compared to the baseline for the Czech-English and German-English language pairs respectively.

	Objective	BLEU
Cs-En	$\mathcal{F}_{baseline}$	19.68
	$\mathcal{F}_{all}$	**20.4**
De-En	$\mathcal{F}_{baseline}$	24.85
	$\mathcal{F}_{all}$	**25.21**

Table 4.10: BLEU score of the baseline and the model trained with $\mathcal{F}_{all}$. Pairwise bootstrap resampling [27] resulted in a p-value < 0.01 which indicates the statistical significance of the observed difference.

Improved BLEU scores for the faithful model can be due to two reasons: 1) the faithfulness objective can be seen as a regularization term which prevents the model from relying too much on the target-side context and the implicit language model in the decoder, which results in increased contribution of attention on the decoder and reducing some bias in the model. 2) penalizing the model for the lack of connection between justification and prediction forces the model to learn better translations by forcing it to justify each output in a right answer for the right reason paradigm. Figure 4.2 shows some examples of how our proposed model can produce better translations.

4.3.5 Do the new models have sparser attention?

Table 4.11 shows the average entropy and average normalized entropy for the baseline, the proposed model ($\mathcal{F}_{all}$), and the model trained with attention entropy regularization respectively. Evidently, the proposed model has not increased sparsity. On the other hand attention entropy regularization has been very effective in making attention weights sparser. But Table 4.3 and 4.4 indicates that attention entropy regularization has not been effective in increasing faithfulness. This suggests that sharper attention weights only affect the context vector and do not contribute to increased dependence of the decoder on attention. The proposed model does not end up with sparse attention, and entropy regularization has been ineffective at increasing faithfulness although it does learn a sparse attention model.

src es ist alles hier es ist alles online
ref it 's all here it 's all on the web
base it 's all right it 's all online .
ours it 's all here it 's all online .

src die erste ist , dass wir uns nicht weiterentwickeln werden .
ref the first is that we will not evolve .
base the first is that we will not move forward .
ours the first is that we will not evolve .

src sie drängten wasser aus dem land heraus und hinaus in den fluss
ref they pushed water off the land and out into the river
base they kept running water from the land and out in the river
ours they pushed water out of the country and out in the river .

src anstatt hunderte von kilometern entfernt im norden
ref instead of hundreds of miles away in the north
base instead of hundreds of miles away from north america
ours instead of hundreds of miles away from north

Figure 4.2: These examples show some cases where the more faithful model trained using our faithfulness objective produces better translations compared to the baseline model. In each of these cases, perturbing the attention weights has no effect on the baseline model output. The faithful model is able to focus on the source side when needed in order to produce a more accurate translation.

	Model	*AvgEnt*	*AvgNormEnt*
Cs-En	$\mathcal{F}_{baseline}$	0.69	0.23
	$\mathcal{F}_{all}$	0.84	0.27
	$\mathcal{F}_{ent}$	**0.35**	**0.11**
De-En	$\mathcal{F}_{baseline}$	0.89	0.29
	$\mathcal{F}_{all}$	1.0	0.32
	$\mathcal{F}_{ent}$	**0.43**	**0.14**

Table 4.11: Average entropy and average normalized entropy of the baseline, the proposed model ($\mathcal{F}_{all}$), and the model trained with attention entropy regularization.

4.4 Summary

In this chapter we have demonstrated our findings regarding the faithfulness of a baseline NMT model and effect of our proposed method on it and also its relation to sparsity. Our findings show that attention is much less faithful for rationalizing prediction of function words compared to content words. Moreover we show that our proposed method can successfully improve faithfulness in NMT without sacrificing accuracy and in some cases even with improved accuracy. Moreover we find that neither our proposed model has more sparse attention, nor the model trained with attention entropy regularization for increased sparsity has less unfaithful attention.

Chapter 5

Conclusion

Using attention weights to justify a model's prediction is tempting and seems intuitive at the first glance. It is, however, not clear whether attention can be trusted for such purposes. To what extent is it trustworthy or faithful and reflect the true internal reasoning of the model? In this work we have proposed a method for quantifying faithfulness of NMT models. We have also investigated behavior of NMT under presence of different stress tests. To optimize faithfulness we have defined a novel objective function that rewards faithful behavior through probability divergence. We also show that the additional constraint in the training objective for NMT does not harm translation quality and in some cases we see some better translations presumably due to the regularization effect of our faithfulness objective. In future we intend to expand this work to language pairs where target language is not English. While this work is focused on NMT, our approach is more generally applicable to other neural models that exploit attention and where researchers are implicitly trusting attention heatmaps as a means of explanation of the model behaviour. Our faithfulness objective can be used for other NLP tasks such as text classification. We aim to investigate and improve faithfulness of attention-based explanations in more sophisticated attention models such as Transformers [63]. We can generalize our approach by designing explanatory modules in NMT through functionality separation (alignment, reordering, etc.) instead of relying only on attention. We also plan to investigate if faithful models can also be more useful for copy models and other applications of attention heatmaps in NMT.

Bibliography

[1] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049, 2018.

[2] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining predictions of non-linear classifiers in NLP. In Proceedings of the 1st Workshop on Representation Learning for NLP, pages 1–7, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1601. URL https://www.aclweb.org/anthology/W16-1601.

[3] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 159–168, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5221. URL https://www.aclweb.org/anthology/W17-5221.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[6] Joost Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2963–2977, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1284. URL https://www.aclweb.org/anthology/P19-1284.

[7] Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1–10, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1001.

[8] Jesús Calvillo and Matthew Crocker. Language production dynamics with recurrent neural networks. In Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing, pages 17–26, Melbourne, July 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-2803.

[9] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

[10] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 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, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-4828.

[11] Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. Incorporating structural alignment biases into an attentional neural translation model. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 876–885, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1102. URL https://www.aclweb.org/anthology/N16-1102.

[12] Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. Understanding and improving morphological learning in the neural machine translation decoder. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 142–151, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1015.

[13] Shuoyang Ding, Hainan Xu, and Philipp Koehn. Saliency-driven word alignment interpretation for neural machine translation. In Proceedings of the Fourth Conference on Machine Translation, pages 1–12, Florence, Italy, August 2019. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W19-5201.

[14] Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1150–1159, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1106. URL https://www.aclweb.org/anthology/P17-1106.

[15] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3719–3728, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1407. URL https://www.aclweb.org/anthology/D18-1407.

[16] Hamidreza Ghader and Christof Monz. What does attention in neural machine translation pay attention to? In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 30–39, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1004.

[17] Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4952–4957, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1537. URL https://www.aclweb.org/anthology/D18-1537.

[18] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3681–3688, 2019.

[19] Bernease Herman. The promise and peril of human evaluation for model interpretability. arXiv preprint arXiv:1711.07414, page 8, 2017.

[20] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, 2020.

[21] Alon Jacovi and Yoav Goldberg. Aligning faithful interpretations with their social attribution. arXiv preprint arXiv:2006.01067, 2020.

[22] Sarthak Jain and Byron C. Wallace. Attention is not explanation. 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), pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1357.

[23] Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4459–4473, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.409.

[24] Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. Representation of linguistic form and function in recurrent neural networks. Computational Linguistics, 43(4):761–780, 2017.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[26] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4012.

[27] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Proceedings of the 2004 conference on empirical methods in natural language processing, pages 388–395, 2004.

[28] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://www.aclweb.org/anthology/W17-3204.

[29] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-2045.

[30] Shaohui Kuang, Junhui Li, António Branco, Weihua Luo, and Deyi Xiong. Attention focusing for neural machine translation by bridging source and target embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1767–1776, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1164. URL https://www.aclweb.org/anthology/P18-1164.

[31] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1902.00006, 2019.

[32] Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. Interactive visualization and manipulation of attention-based neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 121–126, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2021. URL https://www.aclweb.org/anthology/D17-2021.

[33] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 107–117, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1011. URL https://www.aclweb.org/anthology/D16-1011.

[34] Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi. Evaluating explanation methods for neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 365–375, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.35.

[35] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 681–691, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1082. URL https://www.aclweb.org/anthology/N16-1082.

[36] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 681–691, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1082. URL https://www.aclweb.org/anthology/N16-1082.

[37] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. arXiv preprint arXiv:1612.08220, 2016.

[38] Zachary C Lipton. The mythos of model interpretability. Queue, 16(3):31–57, 2018.

[39] Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable NLP: A generative explanation framework for text classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5570–5581, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1560. URL https://www.aclweb.org/anthology/P19-1560.

[40] Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Neural machine translation with supervised attention. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3093–3102, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1291.

[41] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://www.aclweb.org/anthology/D15-1166.

[42] Chaitanya Malaviya, Pedro Ferreira, and André F. T. Martins. Sparse and constrained attention for neural machine translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 370–376, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2059. URL https://www.aclweb.org/anthology/P18-2059.

[43] Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In International Conference on Machine Learning, pages 1614–1623, 2016.

[44] Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. Supervised attentions for neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2283–2288, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1249. URL https://www.aclweb.org/anthology/D16-1249.

[45] Sina Mohseni and Eric D Ragan. A human-grounded evaluation benchmark for local explanations of machine learning. arXiv preprint arXiv:1801.05075, 2018.

[46] Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. Interrogating the explanatory power of attention in neural machine translation. In Proceedings of the 3rd Workshop on Neural Generation and Translation, pages 221–230, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5624. URL https://www.aclweb.org/anthology/D19-5624.

[47] W. James Murdoch, Peter J. Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=rkRwGg-0Z.

[48] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In Proceedings of the Eighth International Conference on Language Resources and Evaluation

(LREC'12), pages 2089–2096, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.

[49] Nina Poerner, Hinrich Schütze, and Benjamin Roth. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 340–350, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1032. URL https://www.aclweb.org/anthology/P18-1032.

[50] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. Learning to deceive with attention-based explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4782–4793, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.432.

[51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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, 2016.

[52] Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge, 2019.

[53] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pages 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://doi.org/10.24963/ijcai.2017/371.

[54] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems, 28(11):2660–2673, 2016.

[55] Sofia Serrano and Noah A. Smith. Is attention interpretable? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://www.aclweb.org/anthology/P19-1282.

[56] Xing Shi, Kevin Knight, and Deniz Yuret. Why neural translations are the right length. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2278–2282, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1248. URL https://www.aclweb.org/anthology/D16-1248.

[57] Felix Stahlberg, Danielle Saunders, and Bill Byrne. An operation sequence model for explainable neural machine translation. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 175–186, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5420. URL https://www.aclweb.org/anthology/W18-5420.

[58] Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. Obtaining faithful interpretations from compositional neural networks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5594–5608, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.495.

[59] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[60] Rico Tang, Gongbo andSennrich and Joakim Nivre. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 26–35, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6304.

[61] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 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, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1008. URL https://www.aclweb.org/anthology/P16-1008.

[62] Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. Context gates for neural machine translation. Transactions of the Association for Computational Linguistics, 5:87–99, 2017. doi: 10.1162/tacl_a_00048. URL https://www.aclweb.org/anthology/Q17-1007.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[64] Jesse Vig and Yonatan Belinkov. 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, pages 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://www.aclweb.org/anthology/W19-4808.

[65] Jesse Vig and Yonatan Belinkov. 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, pages 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-4808.

[66] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 606–615, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1058. URL https://www.aclweb.org/anthology/D16-1058.

[67] Sarah Wiegreffe and Yuval Pinter. 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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://www.aclweb.org/anthology/D19-1002.

[68] Jiajun Zhang, Yang Zhao, Haoran Li, and Chengqing Zong. Attention with sparsity regularization for neural machine translation and summarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(3):507–518, 2018.

[69] Ruiqi Zhong, Steven Shao, and Kathleen McKeown. Fine-grained sentiment analysis with faithful attention. arXiv preprint arXiv:1908.06870, 2019.