Evaluating Explanation Methods for Neural Machine Translation

Jierui Li[1] **Lemao Liu**[2*] **Huayang Li**[2] **Guanlin Li**[3] **Guoping Huang**[2] **Shuming Shi**[2]
[1]University of Electronic Science and Technology of China
[2]Tencent AI Lab, [3]Harbin Institute of Technology
{lijierui19,epsilonlee.green}@gmail.com,
{redmondliu,alanili,donkeyhuang,shumingshi}@tencent.com

Abstract

Recently many efforts have been devoted to interpreting the black-box NMT models, but little progress has been made on metrics to evaluate explanation methods. Word Alignment Error Rate can be used as such a metric that matches human understanding, however, it can not measure explanation methods on those target words that are not aligned to any source word. This paper thereby makes an initial attempt to evaluate explanation methods from an alternative viewpoint. To this end, it proposes a principled metric based on *fidelity* in regard to the predictive behavior of the NMT model. As the exact computation for this metric is intractable, we employ an efficient approach as its approximation. On six standard translation tasks, we quantitatively evaluate several explanation methods in terms of the proposed metric and we reveal some valuable findings for these explanation methods in our experiments.

1 Introduction

Neural machine translation (NMT) has witnessed great success during recent years (Sutskever et al., 2014; Bahdanau et al., 2014; Gehring et al., 2017; Vaswani et al., 2017). One of the main reasons is that neural networks possess the powerful ability to model sufficient context by entangling all source words and target words from translation history. The downside yet is its poor interpretability: it is unclear which specific words from the entangled context are crucial for NMT to make a translation decision. As interpretability is important for understanding and debugging the translation process and particularly to further improve NMT models, many efforts have been devoted to explanation methods for NMT (Ding et al., 2017; Alvarez-Melis and Jaakkola, 2017; Li et al., 2019; Ding et al., 2019; He et al., 2019). However, little progress has been made on evaluation metric to study how good these explanation methods are and which method is better than others for NMT.

Generally speaking, we recognize two orthogonal dimensions for evaluating the explanation methods: i) how much the pattern (such as source words) extracted by an explanation method matches *human understanding* on predicting a target word; or ii) how the pattern matches *predictive behavior* of the NMT model on predicting a target word. In terms of i), Word Alignment Error Rate (AER) can be used as a metric to evaluate an explanation method by measuring agreement between *human-annotated* word alignment and that derived from the explanation method. However, AER can not measure explanation methods on those target words that are not aligned to any source words according to human annotation.

In this paper, we thereby make an initial attempt to measure explanation methods for NMT according to the second dimension of interpretability, which covers all target words. The key to our approach can be highlighted as *fidelity*: when extracting the most relevant words with an explanation method, if those relevant words have the potential to construct an optimal proxy model that agrees well with the NMT model on making a translation decision, then this explanation method is good (§3). To this end, we formalize a principled evaluation metric as an optimization problem over the expected disagreement between the optimal proxy model and the NMT model(§3.1). Since it is intractable to exactly calculate the principled metric for a given explanation method, we propose an approximate metric to address the optimization problem. Specifically, inspired by statistical learning theory (Vapnik, 1999), we cast the optimization problem into a standard machine learning problem which is addressed in a two-step strat-

[*] This work was done during J.Li & G.Li's internship at Tencent AI Lab. L.Liu is the corresponding author.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 365–375
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

egy: firstly we follow empirical risk minimization to optimize the empirical risk; then we validate the optimized parameters on a held-out test dataset. Moreover, we construct different proxy model architectures by utilizing the most relevant words to make a translation decision, leading to variant approximate metric in implementation (§3.2).

We apply the approximate metric to evaluate four explanation methods including attention (Bahdanau et al., 2014; Vaswani et al., 2017), gradient norm (Li et al., 2016), weighted gradient (Ding et al., 2019) and prediction difference (Li et al., 2019). We conduct extensive experiments on three standard translation tasks for two popular translation models in terms of the proposed evaluation metric. Our experiments reveal valuable findings for these explanation methods: 1) The evaluation methods (gradient norm and prediction difference) are good to interpret the behavior of NMT; 2) The prediction difference performs better than other methods.

This paper makes the following contributions:

- It presents an attempt at evaluating the explanation methods for neural machine translation from a new viewpoint of fidelity.
- It proposes a principled metric for evaluation, and to put it into practice it derives a simple yet efficient approach to approximately calculate the metric.
- It quantitatively compares several different explanation methods and evaluates their effects in terms of the proposed metric.

2 NMT and Explanation Methods

2.1 NMT Models

Suppose $\boldsymbol{x} = \{x_1, \cdots, x_{|\boldsymbol{x}|}\}$ denotes a source sentence with length $|\boldsymbol{x}|$ and $\boldsymbol{y} = \{y_1, \cdots, y_{|\boldsymbol{y}|}\}$ is a target sentence. Most NMT literature models the following conditional probability $P(\boldsymbol{y} \mid \boldsymbol{x})$ in an encoder-decoder fashion:

$$\begin{aligned} P(\boldsymbol{y} \mid \boldsymbol{x}) &= \prod_t P(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x}) \\ &= \prod_t P(y_t \mid s_t), \end{aligned} \tag{1}$$

where $\boldsymbol{y}_{<t} = \{y_1, \cdots, y_{t-1}\}$ denotes a prefix of $\boldsymbol{y}$ with length $t-1$, and s_t is the decoding state vector of timestep t. In the encoding stage, the encoder of a NMT model transforms the source sentence $\boldsymbol{x}$ into a sequence of hidden vectors $\boldsymbol{h} = \{h_1, \cdots, h_{|\boldsymbol{x}|}\}$. In the decoding stage, the decoder module summarizes the hidden vectors $\boldsymbol{h}$ and the history decoding states $\boldsymbol{s}_{<t} = \{s_1, \cdots, s_{t-1}\}$ into the decoding state vector s_t. In this paper, we consider two popular NMT translation architectures, RNN-SEARCH (Bahdanau et al., 2014) and TRANSFORMER (Vaswani et al., 2017). RNN-SEARCH utilizes a bidirectional RNN to define $\boldsymbol{h}$ and it computes s_t by the attention function over $\boldsymbol{h}$, i.e.,

$$s_t = \mathrm{Attn}(s_{t-1}, \boldsymbol{h}), \tag{2}$$

where Attn is the attention function, which is defined as follows:

$$\begin{aligned} \mathrm{Attn}(q, \boldsymbol{v}) &= \sum_i \alpha(q, v_i) v_i, \\ \alpha(q, v_i) &= \frac{\exp\big(e(q, v_i)\big)}{\sum_j \exp\big(e(q, v_j)\big)}, \end{aligned} \tag{3}$$

where q and v_i are vectors, e is a similarity function over a pair of vectors and α is its normalized function.

Different from RNN-SEARCH, which relies on RNN, TRANSFORMER employs an attention network to define $\boldsymbol{h}$, and two additional attention networks to define s_t as follows: [1]

$$\begin{aligned} s_t &= \mathrm{Attn}(s_{t+\frac{1}{2}}, \boldsymbol{h}), \\ s_{t+\frac{1}{2}} &= \mathrm{Attn}(s_{t-1}, \boldsymbol{s}_{<t}). \end{aligned} \tag{4}$$

2.2 Explanation Methods

In this section, we describe several popular explanation methods that will be evaluated with our proposed metric. Suppose $c_t = \langle \boldsymbol{y}_{<t}, \boldsymbol{x} \rangle$ denotes the context at timestep t, w (or w') denotes either a source or a target word in the context c_t. According to Poerner et al. (2018), each explanation method for NMT could be regarded as a word relevance score function $\phi(w; y, c_t)$, where $\phi(w; y, c_t) > \phi(w'; y, c_t)$ indicates that w is more useful for the translation decision $P(y_t|c_t)$ than word w'.

Attention Since Bahdanau et al. (2014) propose the attention mechanism for NMT, it has been the most popular explanation method for NMT (Tu et al., 2016; Mi et al., 2016; Liu et al., 2016; Zenkel et al., 2019).

[1] Due to space limitation, we present the notations for a single layer NMT models, and for TRANSFORMER we only keep the attention (with a single head) block while skipping other blocks such as resNet and layer normalization. More details can be found in the references (Vaswani et al., 2017).

To interpret RNN-SEARCH and TRANSFORMER, we define different ϕ for them based on attention. For RNN-SEARCH, since attention is only defined on source side, $\phi(w; y, c_t)$ can be defined only for the source words:

$$\phi(x_i; y, c_t) = \alpha(s_{t-1}, h_i)$$

where α is the attention weight defined in Eq.(3), and s_{t-1} is the decoding state of RNN-SEARCH defined in Eq.(2). In contrast, TRANSFORMER defines the attention on both sides and thus $\phi(w; y, c_t)$ is not constrained to source words:

$$\phi(w; y, c_t) = \begin{cases} \alpha(s_{t+\frac{1}{2}}, h_i) & \text{if } w = x_i, \\ \alpha(s_{t-1}, s_j) & \text{if } w = y_j \text{ and } j < t, \end{cases}$$

where s_{t-1} and $s_{t+\frac{1}{2}}$ are defined in Eq.(4).

Gradient Different from attention that is restricted to a specific family of networks, the explanation methods based on gradient are more general. Suppose $g(w, y)$ denotes the gradient of $P(y \mid c_t)$ w.r.t to the variable w in c_t:

$$g(w, y) = \frac{\partial P(y \mid c_t)}{\partial w} \tag{5}$$

where ∂w denotes the gradient w.r.t the embedding of the word w, since a word itself is discrete and can not be taken gradient. Therefore, $g(w, y)$ returns a vector with the same shape as the embedding of w. In this paper, we implement two different gradient-based explanation methods and derive different definitions of $\phi(w; y, c_t)$ as follows.

- **Gradient Norm** (Li et al., 2016): The first definition of ϕ is the $\ell - 1$ norm of g:

$$\phi(w; y, c_t) = |g(w, y)|_{\ell-1}.$$

- **Weighted Gradient** (Ding et al., 2019): The second one is defined as the weighted sum of the embedding of w, with the return of g as the weight:

$$\phi(w; y, c_t) = g(w, y)^\top \cdot w.$$

It is worth noting that for each sentence $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, one has to independently calculate $\frac{\partial P(y|c_t)}{\partial w}$ for each timestep t. Therefore, one has to calculate $|\boldsymbol{y}|$ times of gradient for each sentence. In contrast, when training NMT, one only requires calculating sentence level gradient and it only calculates one gradient thanks to gradient accumulation in back propagation algorithm.

Prediction Difference Li et al. (2019) propose a prediction difference (PD) method, which defines the contribution of the word w by evaluating the change in the probability after removing w from c_t. Formally, $\phi(w; y, c_t)$ based on prediction difference is defined as follows:

$$\phi(w; y, c_t) = P(y \mid c_t) - P(y \mid c_t \backslash w)$$

where $P(y \mid c_t)$ is the NMT probability of y defined in Eq.(1), and $P(y \mid c_t \backslash w)$ denotes the NMT probability of y after excluding w from its context c_t. To achieve the effect of excluding w from c_t, it simply replaces the word embedding of w with zero vector before feeding it into the NMT model.

3 Evaluation Methodology

3.1 Principled Metric

The key to our metric is described as follow: to define an explanation method ϕ good enough in terms of our metric, the relevant words selected by ϕ from the context c_t should have the potential to construct an optimal model that exhibits similar behavior to the target model $P(y \mid c_t)$. To formalize this metric, we first specify some necessary notations.

Assume that $f(c_t)$ is the target word predicted by $P(y \mid c_t)$, i.e., $f(c_t) = \arg\max_y P(y \mid c_t)$. In addition, let $\mathcal{W}_\phi^k(c_t)$ be the top-k relevant words on the source side and target side of the context c_t:

$$\mathcal{W}_\phi^k(c_t) = \\ \text{top}_{w \in \boldsymbol{x}}^k \phi(w; f(c_t), c_t) \cup \text{top}_{w \in \boldsymbol{y}_{<t}}^k \phi(w; f(c_t), c_t)$$

where $\cup$ denotes the union of two sets, and $\text{top}_{w \in \boldsymbol{x}}^k \phi(w; f(c_t), c_t)$ returns words corresponding to the k largest ϕ values. [2]

In addition, suppose $Q(y \mid \mathcal{W}_\phi^k(c_t); \theta)$ ($Q(\theta)$ or Q for brevity) is a proxy model that makes a translation decision on top of $\mathcal{W}_\phi^k(c_t)$ rather than the entire context c_t like a standard NMT model. Formally, we define a principled metric as follows:

Definition 1 *The metric of ϕ is defined by*

$$\min_Q \min_\theta -\mathbb{E}_{c_t}\Big[\log Q\big(f(c_t) \mid \mathcal{W}_\phi^k(c_t); \theta\big)\Big] \tag{6}$$

[2] In fact, $\mathcal{W}_\phi^k(c_t) \rightarrow f(c_t)$ can be considered as generalized translation rules obtained by ϕ. In other words, the rules are extracted under teacher forcing decoding. In particular, if $k = 1$, this is similar to the statistical machine translation (SMT) with word level rules (Koehn, 2009), except that a generalized translation rule also involves a word from $\boldsymbol{y}_{<t}$ which simulates the role of language modeling in SMT.

where $\mathbb{E}_{c_t}[\cdot]$ denotes the expectation with respect to the data distribution of c_t, and Q is minimized over all possible proxy models.

The underlying idea of the above metric is to measure the expectation of the disagreement between an optimal proxy model Q constructed from ϕ and the NMT model P. Here the disagreement is measured by the minus log-likelihood of Q over the data $\langle \mathcal{W}_\phi^k(c_t), f(c_t)\rangle$ whose label $f(c_t)$ is generated from P. [3]

Definition of Fidelity The metric of ϕ actually defines fidelity by measuring how much the optimal proxy model defined on $\mathcal{W}_\phi^k(c_t)$ disagrees with $P(y \mid c_t)$. The mention of *fidelity* is widely used in model compression (Buciluǎ et al., 2006; Polino et al., 2018), model distillation (Hinton et al., 2015; Liu et al., 2018), and particularly in evaluating the explanation models for black-box neural networks (Lakkaraju et al., 2016; Bastani et al., 2017). These works focus on learning a specific model Q on which fidelity can be directly defined. However, we are interested in evaluating explanation methods ϕ where Q is a latent variable that we have to minimize. By doing this, fidelity in our metric is defined on ϕ as shown in Eq (6).

3.2 Approximation

Generally, it is intractable to exactly calculate the principled metric due to two main challenges. On one hand, the real data distribution of c_t is unknowable, making it impossible to exactly define the expectation with respect to an unknown distribution. On the other hand, the domain of a proxy model Q is not bounded, and it is difficult to minimize a model Q within an unbounded domain.

Empirical Risk Minimization Inspired by the statistical learning theory (Vapnik, 1999), we calculate the expected disagreement over c_t by a two-step strategy: we minimize the empirical risk to obtain an optimized θ for a given Q; and then we estimate the risk defined on a held-out test set by using the optimized θ. In this way, we cast the principled metric into a standard machine learning task.

For a given model architecture Q, to optimize θ, we first collect the training set as $\{\langle \mathcal{W}_\phi^k(c_t), f(c_t)\rangle\}$ for each sentence pair $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ at every time step t, where $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ is a sentence pair from a given bilingual corpus $\mathcal{D}_{\text{train}} = \{\langle \boldsymbol{x}^n, \boldsymbol{y}^n\rangle \mid n = 1, \cdots, N\}$. Then we optimize θ by the empirical risk minimization:

$$\min_\theta \sum_{\langle \boldsymbol{x}, \boldsymbol{y}\rangle \in \mathcal{D}_{\text{train}}} \sum_{c_t} -\log Q(f(c_t) \mid \mathcal{W}_\phi^k(c_t); \theta) \tag{7}$$

Proxy Model Selection In response to the second challenge of the unbounded domain, we define a surrogate distribution family $\mathcal{Q}$, and then approximately calculate Eq.(6) within $\mathcal{Q}$ instead:

$$\min_{Q \in \mathcal{Q}} \min_\theta -\mathbb{E}_{c_t}\Big[\log Q\big(f(c_t) \mid \mathcal{W}_\phi^k(c_t); \theta\big)\Big] \tag{8}$$

We consider three different proxy models including multi-layer feedforward network (FN), recurrent network (RN) and self-attention network (SA). In details, for different networks $\epsilon \in \{\text{FN}, \text{RN}, \text{SA}\}$, the proxy model Q^ϵ is defined as follows:

$$Q^\epsilon(y \mid \mathcal{W}_\phi^k(c_t)) = P(y \mid s_t^\epsilon)$$

where s_t^ϵ is the decoding state regarding different architecture ϵ. Specifically, for feedforward network, the decoding state is defined by

$$s_t^{\text{FN}} = \text{FNN}(\tilde{x}_1, \cdots, \tilde{x}_k, \tilde{y}_1, \cdots, \tilde{y}_k).$$

For $\epsilon \in \{\text{RN}, \text{SA}\}$, the decoding state s_t^ϵ is defined by

$$s_t^\epsilon = \text{Attn}\big(s_0, \{h_{\tilde{x}_1}, \cdots, h_{\tilde{x}_k}, h_{\tilde{y}_1} \cdots, h_{\tilde{y}_k}\}\big),$$

where $\tilde{x}$ and $\tilde{y}$ are source and target side words from $\mathcal{W}_\phi^k(c_t)$, s_0 is the query of init state, h is the position-aware representations of words, generated by the encoder of RN or SA as defined in Eq.(3) and Eq.(4). For RN, s_t^{RN} is the weight-sum vectors of a bidirectional LSTM over all selected top k source and target words; while for SA, s_t^{SA} is the weight-sum of vectors over the SA networks.

3.3 Evaluation Paradigm

Given a bilingual training set $\mathcal{D}_{\text{train}}$ and a bilingual test set $\mathcal{D}_{\text{test}}$, we evaluate an explanation method ϕ w.r.t the NMT model $P(y \mid c_t)$ by setting the proxy model family $\mathcal{Q}(\theta)$ to include three neural networks as defined before. Following the

[3] It is natural to extend our definition by using other similar disagreement measures such as the KL distance. Since the KL distance requires additional GPU memory to restore the distribution P in the implementation, we employ the minus log-likelihood for efficiency in our experiments.

Algorithm 1 Calculating the evaluation metric

```
Require: $\phi$, $\mathcal{Q}(\theta)$, $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$
Ensure: the metric score $m$ of $\phi$ over $\mathcal{D}_{\text{test}}$
$\mathcal{Q}^* = \{\}$
Collect $\langle f(c_t), \mathcal{W}_{\phi}^{k}(c_t) \rangle$ from $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$
    to obtain two sets $\mathcal{FW}_{\text{train}}$ and $\mathcal{FW}_{\text{test}}$
for $Q(\theta) \in \mathcal{Q}(\theta)$ do
    Optimize $\theta^*$ over $\mathcal{FW}_{\text{train}}$ w.r.t Eq.(7)
    Add $Q(\theta^*)$ into $\mathcal{Q}^*$
end for
for $Q^* \in \mathcal{Q}^*$ do
    $m_{Q^*} = 0$
    for $\langle f(c_t), \mathcal{W}_{\phi}^{k}(c_t) \rangle \in \mathcal{FW}_{\text{test}}$ do
        $m_{Q^*} \mathrel{+}= -\log Q^*(f(c_t) \mid \mathcal{W}_{\phi}^{k}(c_t))$
    end for
end for
Return $\min\limits_{Q^* \in \mathcal{Q}^*} \exp\left(\frac{m_{Q^*}}{|\mathcal{FW}_{\text{test}}|}\right)$
```

standard process of addressing a machine learning problem, Algorithm 1 summarizes the procedure to approximately calculate the metric of ϕ on the test dataset $\mathcal{D}_{\text{test}}$, which returns the preplexity (PPL) on $\mathcal{FW}_{\text{test}}$. [4]

In this paper, we try four different choices to specify the surrogate family, i.e., $\mathcal{Q} = \{Q^{\text{FN}}\}$, $\mathcal{Q} = \{Q^{\text{RN}}\}$, $\mathcal{Q} = \{Q^{\text{SA}}\}$, and $\mathcal{Q} = \{Q^{\text{FN}}, Q^{\text{RN}}, Q^{\text{SA}}\}$, leading to four instances of our metric respectively denoted as **FN**, **RN**, **SA** and **Comb**. In addition, as the **baseline** metric, we employ the well-trained NMT model P as the proxy model Q by masking out the input words that do not appear in the rule set $\mathcal{W}_{\phi}^{k}(c_t)$). For the baseline metric, it doesn't require to train $Q's$ parameter θ and tests on $\mathcal{D}_{\text{test}}$ only. Since P is trained with the entire context c_t whereas it is testified on $\mathcal{W}_{\phi}^{k}(c_t)$, this mismatch may lead to poor performance and is thus less trusted. This baseline metric extends the idea of Arras et al. (2016); Denil et al. (2014) from classification tasks to structured prediction tasks like machine translation which are highly dependent on context rather than just keywords.

4 Experiments

In this section, we conduct experiments to prove the effectiveness of our metric from two viewpoints: how good an explanation method is and which explanation method is better than others.

4.1 Settings

Datasets We carry out our experiments on three standard IWSLT translation tasks including IWSLT14 De⇒En (167k sentence pairs), IWSLT17 Zh⇒En (237k sentence pairs) and IWSLT17 Fr⇒En (229k sentence pairs). All these datasets are tokenized and applied BPE (Byte-Pair Encoding) following Ott et al. (2019). The target side vocabulary sizes of the three datasets are 8876, 11632, and 9844 respectively. In addition, we carry out extended experiments on three large-scale WMT translation tasks including WMT14 De⇒En (4.5m sentence pairs), WMT17 Zh⇒En (22m sentence pairs) and WMT14 Fr⇒En (40.8m sentence pairs), with vocabulary sizes 22568, 29832, 27168 respectively.

NMT Systems To examine the generality of our evaluation method, we conduct experiments on two NMT systems, i.e. RNN-SEARCH (denoted by **RNN**) and TRANSFORMER (denoted by **Trans.**), both of which are implemented with fairseq (Ott et al., 2019). For RNN, we adopt the 1-layer RNN with LSTM cells whose encoder (bi-directional) and decoder hidden units are 256 and 512 respectively. For TRANSFORMER on the IWSLT datasets, the number of layers and attention heads are 2 and 4 respectively. For both models, we set the embedding dimensions as 256. On WMT datasets, we simply use TRANSFORMER-BASE with 4 attention heads. The performances of our NMT models are comparable to those reported in recent literature (Tan et al., 2019).

Explanation Methods On both NMT systems, we implement four explanation methods, i.e. Attention (ATTN), gradient norm (NGRAD), weighted gradient (WGRAD), and prediction difference (PD) as mentioned in Section §2.

Our metric We implemented five instantiations of the proposed metric including FN, RN, SA, Comb, and Baseline (Base for brevity) as presented in section §3.3. To configurate them, we adopt the same settings from NMT systems to train SA and RN. FN is implemented with feeding the features of bag of words through a 3-layer fully connected network. As given in algorithm 1, the approximate fidelity is estimated through Q with the lowest PPL, therefore the best metric is

[4] Note that the negative log-likelihood in Eq. 6 is proportional to PPL and thus we use PPL as the metric value in this paper.

NMT	Metric	ATTN	PD	NGRAD	WGRAD
Trans	Base	196.9	54.3	193.4	13400
	FN	13.9	5.8	11.3	131.2
	RN	13.8	5.7	10.7	126.7
	SA	13.9	5.5	10.8	119.5
	Comb	13.8	5.5	10.7	119.5
RNN	Base	-	54.2	90.3	28587
	FN	-	6.7	8.3	170.8
	RN	-	6.5	7.8	163.2
	SA	-	6.5	8.1	154.9
	Comb	-	6.5	7.8	154.9

Table 1: The PPL comparison for the five metric instantiations on the IWSLT De⇒En dataset.

Method	Total	B_1	B_2	B_3	B_4	B_5
ATTN	1.97M	1.65M	298K	23.7K	1.54K	104
PD	1.62M	1.25M	328K	31.2K	2.11K	108
NGRAD	1.89M	1.54M	326K	27.6K	1.64K	83
WGRAD	2.62M	2.37M	278K	17.5K	0.86K	34

Table 2: Density of the extracted rules from TRANSFORMER on the IWSLT De⇒En . The density is measured by the total number of unique rules and the number of rules with certain frequency in each interval B_i: $B_1 = (0, 1]$, $B_2 = (1, 10]$, $B_3 = (10, 100]$, $B_4 = (100, 1000]$, and $B_4 = (1000, \infty)$.

that achieves the lowest PPL since it results in a closer approximation to the real fidelity.

4.2 Experiments on IWSLT tasks

In this subsection, we first conduct experiments and analysis on the IWSLT De⇒En task to configurate fidelity-based metric and then extend the experiments to other IWSLT tasks.

Comparison of metric instantiations We calculate PPL on the IWSLT De⇒En dataset for four metric instantiations (FN, RN, SA, Comb) and Baseline (Base) with $k = 1$ to extract the most relevant words. Table 1 summarizes the results for two translation systems (TRANSFORMER annotated as Trans and RNN-SEARCH annotated as RNN), respectively. Note that since there is no target-side attention in RNN-SEARCH, we can not extract the best relevant target word, so Table 1 does not include the results of ATTN method for RNN-SEARCH.

The baseline (Base) achieves undesirable PPL which indicates the relevant words identified by PD failed to make the same decision as the NMT system. The main reason is that the mismatch between training and testing leads to the issue as presented in section §3.3. On the contrary, the other four metric instantiations attain much lower PPL than the Baseline. In addition, the PPLs on PD, NGRAD, and ATTN are much better than those on WGRAD. This finding shows that all PD, NGRAD, and ATTN are good explanation methods except WGRAD in terms of fidelity.

Density of generalizable rules To understand possible reasons for why one explanation method is better under our metric, we make a naive conjecture: when it tries to reveal the patterns that the well-trained NMT has captured, it extracted more concentrated patterns. In other words, a generalized rule $\mathcal{W}_{\phi}^{k}(c_t) \rightarrow f(c_t)$ from one sentence pair can often be observed among other examples.

To measure the density of the extracted rules, we first divide all extracted rules into five bins according to their frequencies. Then we collect the number of rules in each bin as well as the total number of rules. Table 2 shows the statistics to measure the density of rules obtained from different evaluation methods. From this table, we can see that the density for PD is the highest among those for all explanation methods, because it contains fewer infrequent rules in B_1, whereas there are more frequent rules in other bins. This might be one possible reason that PD is better under our fidelity-based evaluation metric.

Stability of ranking order In Table 1 the ranking order is PD > NGRAD > ATTN > WGRAD regarding all five metric instantiations. Generally, a good metric should preserve the ranking order of explanation methods independent of the test dataset. Regarding this criterion of order-preserving property, we analyze the stability of different fidelity-based metric instantiations. To this end, we randomly sample one thousand test data with replacement whose sizes are variant from 1% to 100% and then calculate the rate whether the ranking order is preserved on these test datasets. The results in Table 3 indicate that FN, RN, SA, Comb are more stable than Base to the change of distribution of test sets.

According to Table 1 and Table 3, SA performs similar to the best metric Comb and it is faster than Comb or RN for training and testing, thereby, in the rest of experiments, we mainly employ SA to measure evaluation methods.

	Base	FN	SA	RN	Comb
1%	53.0%	97.1%	99.9%	99.8%	99.8%
5%	56.1%	100%	100%	100%	100%
20%	60.8%	100%	100%	100%	100%
50%	66.8%	100%	100%	100%	100%
100%	75.4%	100%	100%	100%	100%

Table 3: The rate (percentage) of sampled test dataset that have the same rankings as the test set on the IWSLT Zh⇒En dataset.

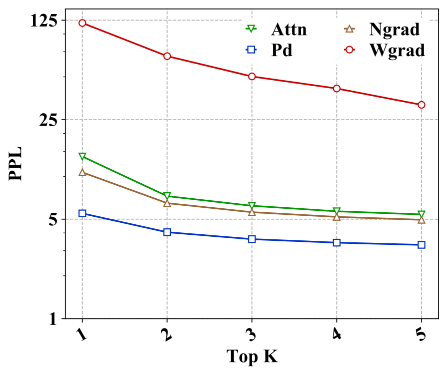


Figure 1: PPL for each explanation method on TRANSFORMER over the IWSLT De⇒En dataset with different k value.

Effects on different k In this experiment, we examine the effects of explanation methods on larger k with respect to SA. Figure 1 depicts the effects of k for TRANSFORMER on De⇒En task. One can clearly observe two findings: 1) the ranking order of explanation methods is invariant for different k. 2) as k is larger, the PPL is much better for each explanation method. 3) the PPL improvement for PD, ATTN, and NGRAD is less after $k > 2$, which further validates that they are powerful in explaining NMT using only a few words.

Testing on other scenarios In the previous experiments, our metric instantiations are trained and evaluated under the same scenario, where c_t used to extract relevant words is obtained from gold data and its label $f(c_t)$ is the prediction from NMT f, namely Teacher Forcing Decode. To examine the robustness of our metric, we apply the trained metric to two different scenarios: real decoding scenario (Real-Decode) where both c_t and its label $f(c_t)$ are from the NMT output; and golden data scenario (Golden-Data) where both c_t and its label are from golden test data. The results for both scenarios are shown in Table 5.

NMT	Methods	Zh⇒En		Fr⇒En	
		Base	**SA**	**Base**	**SA**
Trans	ATTN	897.1	30.8	359.6	12.1
	PD	215.1	10.8	55.3	4.6
	NGRAD	583.7	19	271.0	8.7
	WGRAD	24126	180.9	44287	155.4
RNN	ATTN	-	-	-	-
	PD	139.9	11.3	49.0	5.5
	NGRAD	263.0	13.2	85.8	6.7
	WGRAD	23068	243.1	50657	194.9

Table 4: The PPL comparison for two fidelity-based metric instantiations on two IWSLT datasets.

Methods	R-Dec	Golden	T-Dec
ATTN	11.5	57.1	13.8
PD	4.7	23.3	5.5
NGRAD	8.2	42.0	10.7
WGRAD	115.0	223.4	119.5

Table 5: Evaluating four explanation methods on 3 different scenarios Real-Decode (R-Dec), Golden-Data (Golden) and Teacher-Forcing Decode (T-Dec)) for TRANSFORMER over IWSLT De⇒En task.

From Table 5, we see that the ranking order for both scenarios is the same as before. To our surprise, the results in Real-Decode are even better than those in the matched Teacher Forcing Decode scenario. One possible reason is that the labels generated by a NMT system in the Real-Decode tend to be high-frequency words, which leads to better PPL. In contrast, our metric instantiation in the Golden-Data results in much higher PPL due to the mismatch between training and testing. The performance of experimenting training and testing in the same scenario like Golden-Data can be experimented in future works, however, it's not the focus of this paper.

4.3 Scalability on WMT tasks

Since our metric such as SA requires to extract generalized rules for each explanation method from the entire training dataset, it is computationally expensive for some explanation methods such as gradient methods to directly run on WMT tasks with large scale training data.

Effects on sample size We randomly sample some subsets over WMT Zh⇒En training data that includes 22 million sentence pairs to form several new training sets. The sample sizes of the new training sets are set up to 2 million and the results are illustrated in Figure 2. The following facts are revealed. Firstly, the ranking order of

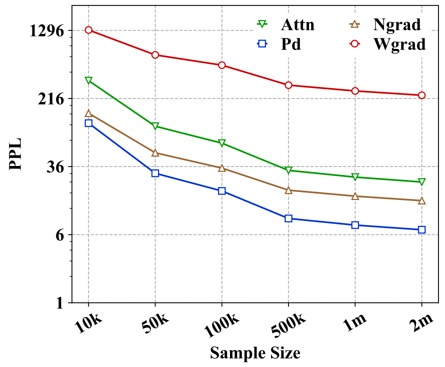


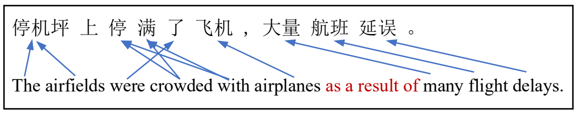
Figure 2: PPL for each explanation method on Transformer over WMT Zh⇒En task with different sample sizes.

Datasets	Methods	Base PPL	Base Rank	SA PPL	SA Rank
Zh⇒En	Attn	336.4	$\underline{2}$	27.3	$\underline{3}$
	Pd	165.3	1	7.7	1
	Ngrad	435.2	$\underline{3}$	16.5	$\underline{2}$
	Wgrad	1615.5	4	263.5	4
De⇒En	Attn	1862.3	$\underline{2}$	17.0	$\underline{3}$
	Pd	1118.2	1	5.4	1
	Ngrad	2827.7	$\underline{3}$	15.1	$\underline{2}$
	Wgrad	6678.1	4	197.4	4
Fr⇒En	Attn	4271.0	3	41.1	3
	Pd	1646.6	1	4.1	1
	Ngrad	2810.2	2	11.8	2
	Wgrad	6703.8	4	163.7	4

Table 6: The PPL and Ranking Order comparison between two fidelity-based metric instantiations (Base and SA) on three WMT datasets. " _ " denotes the mismatch of ranking order.

Datasets	Methods	SA PPL	SA Rank	Alignment AER	Alignment Rank
IWSLT Zh⇒En	Attn	30.8	3	55.0	3
	Pd	10.8	1	50.6	1
	Ngrad	19	2	52.9	2
	Wgrad	180.9	4	79.2	4
WMT Zh⇒En	Attn	27.3	$\underline{3}$	42.1	$\underline{2}$
	Pd	7.7	1	32.7	1
	Ngrad	16.5	$\underline{2}$	49.3	$\underline{3}$
	Wgrad	263.5	4	79.2	4
WMT De⇒En	Attn	17.0	3	48.7	3
	Pd	5.4	1	34.1	1
	Ngrad	15.1	2	48.1	2
	Wgrad	194.7	4	73.5	4

Table 7: Relation with word alignment. " _ " denotes the mismatch of ranking order.


停机坪 上 停 满 了 飞机 ， 大量 航班 延误 。

The airfields were crowded with airplanes as a result of many flight delays.

Figure 3: AER can not evaluate explanation methods on those target words "as a result of", which are not aligned to any word in the source sentence according to human annotation.

four explanation methods remains unchanged with respect to different sample sizes. Secondly, with the increase of the sample size, the metric score decreases slower and slower and there is no significant drop from sampling 2 million sentence pairs to sampling 1 million.

Results on WMT With the analysis of effects on various sample sizes, we choose a sample size of 1 million for the following scaling experiments. The PPL results for WMT De⇒En , Zh⇒En ,and Fr⇒En are listed in Table 6. We can see that the order Pd > Ngrad > Attn > Wgrad evaluated by SA still remains unchanged on these three datasets as before. One can observe that the ranking order under the baseline doesn't agree with SA on WMT De⇒En and Zh⇒En . Since the baseline yields in high PPL due to the mismatch we mentioned in section §3.3 ,in this case, we tend to trust the evaluation results from SA that achieves lower PPL leading to better fidelity.

4.4 Relation to Alignment Error Rate

Since the calculation of the Alignment Error Rate (AER) requires manually annotated test datasets with ground-truth word alignments, we select three different test datasets contained such alignments for experiments, namely, IWSLT Zh⇒En , NIST05 Zh⇒En [5] and Zenkel De⇒En (Zenkel et al., 2019). Note that unaligned target words account for 7.8%, 4.7%, and 9.2% on these three test sets respectively, which are skipped by AER for evaluating explanation methods. For example, in Figure 3, those target words 'as a result' cannot be covered by AER due to the impossibility of human annotation, but for a fidelity-based metric, they can be analyzed as well.

Table 7 demonstrates that our fidelity-based metric does not agree very well with AER on the WMT Zh⇒En task: Ngrad is better than Attn in terms of SA but the result is opposite in terms of AER. Since the evaluation criteria of SA and AER are different, it is reasonable that their evaluation results are different. This finding is in line with

[5] https://www.ldc.upenn.edu/collaborations/evaluations/nist

the standpoint by Jacovi and Goldberg (2020): SA is an objective metric that reflects fidelity of models while AER is a subject metric based on human evaluation. However, it is observed that the ranking by SA is consistent on all three tasks but that by AER is highly dependent on different tasks.

5 Related Work

In recent years, explaining deep neural models has been a growing interest in the deep learning community, aiming at more comprehensible and trustworthy neural models. In this section, we mainly discuss two dominating ways towards it. One way is to develop explanation methods to interpret a target black-box neural network (Bach et al., 2015; Zintgraf et al., 2017). For example, on classification tasks, Bach et al. (2015) propose layer-wise relevance propagation to visualize the relationship between a pair of neurons within networks, and Li et al. (2016) introduce a gradient-based approach to understanding the compositionality in neural networks for NLP. In particular, on structured prediction tasks, many research works design similar methods to understand NMT models (Ding et al., 2017; Alvarez-Melis and Jaakkola, 2017; Ding et al., 2019; He et al., 2019).

The other way is to construct an interpretable model for the target network and then indirectly interpret its behavior to understand the target network on classification tasks (Lei et al., 2016; Murdoch and Szlam, 2017; Arras et al., 2017; Wang et al., 2019). The interpretable model is defined on top of extracted rational evidence and learned by model distillation from the target network. To extract rational evidence from the entire inputs, one either leverages a particular explanation method (Lei et al., 2016; Wang et al., 2019) or an auxiliary evidence extraction model (Murdoch and Szlam, 2017; Arras et al., 2017). Although our work focuses on evaluating explanation methods and does not aim to construct an interpretable model, we draw inspiration from their ideas to design $Q \in \mathcal{Q}$ in Eq. (6) for our evaluation metric.

With the increasing efforts on designing new explanation methods, yet there are only a few works proposed to evaluate them. Mohseni and Ragan (2018) propose a paradigm to evaluate explanation methods for document classification that involves human judgment for evaluation. Poerner et al. (2018) conduct the first human-independent comprehensive evaluation of explanation methods for NLP tasks. However, their metrics are task-specific because they make some assumptions for a specific task. Our work proposes a principled metric to evaluate explanation methods for NMT and our evaluation paradigm is independent of any assumptions as well as humans. It is worth noting that Arras et al. (2016); Denil et al. (2014) directly measure the performance of the target model P on the extracted words without constructing Q to evaluate explanation methods for classification tasks. However, since translation is more complex than classification tasks, P trained on the entire context c_t typically makes a terrible prediction when testing on the compressed context $\mathcal{W}_{\phi}^{k}(c_t)$. As a result, the poor prediction performance makes it difficult to discriminate one explanation method from others, as observed in our internal experiments. Concurrently, Jacovi and Goldberg (2020) make a proposition to evaluate faithfulness of an explanation method separately from readability and plausibility (i.e., human-interpretability), which is similar to our definition of fidelity, but they do not formalize a metric or propose algorithms to measure it.

6 Conclusions

This paper has made an initial attempt to evaluate explanation methods from a new viewpoint. It has presented a principled metric based on fidelity in regard to the predictive behavior of the NMT model. Since it is intractable to exactly calculate the principled metric for a given explanation method, it thereby proposes an approximate approach to address the minimization problem. The proposed approach does not rely on human annotation and can be used to evaluate explanation methods on all target words. On six standard translation tasks, the metric quantitatively evaluates and compares four different explanation methods for two popular translation models. Experiments reveal that PD, NGRAD, and ATTN are all good explanation methods that are able to construct the NMT model's predictions with relatively low perplexity and PD shows the best fidelity among them.

Acknowledgments

We would like to thank all anonymous reviews for their valuable suggestions. This research was supported by Tencent AI Lab.

References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in nlp. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. " what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8):e0181142.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

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

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*.

Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541.

Misha Denil, Alban Demiraj, and Nando De Freitas. 2014. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of WMT*, page 1.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. 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.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 952–961, Hong Kong, China. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *ArXiv*, abs/2004.03685.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684. ACM.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. 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. 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.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. 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. The COLING 2016 Organizing Committee.

Yijia Liu, Wanxiang Che, Huaipeng Zhao, Bing Qin, and Ting Liu. 2018. Distilling knowledge for search-based structured prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1393–1402, Melbourne, Australia. Association for Computational Linguistics.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2283–2288, Austin, Texas. Association for Computational Linguistics.

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

W James Murdoch and Arthur Szlam. 2017. Automatic rule extraction from long short term memory networks. In *International Conference on Learning Representations*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. 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. Association for Computational Linguistics.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. 2018. Model compression via distillation and quantization. *CoRR*, abs/1802.05668.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

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

Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

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

Zhiguo Wang, Yue Zhang, Mo Yu, Wei Zhang, Lin Pan, Linfeng Song, Kun Xu, and Yousef El-Kurdi. 2019. Multi-granular text encoding for self-explaining categorization. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 41–45, Florence, Italy. Association for Computational Linguistics.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*.