Towards Understanding Neural Machine Translation with Word Importance

Shilin He[1,2] **Zhaopeng Tu**[3*] **Xing Wang**[3] **Longyue Wang**[3] **Michael R. Lyu**[1,2] **Shuming Shi**[3]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]Shenzhen Research Institute, The Chinese University of Hong Kong
[1,2]{slhe,lyu}@cse.cuhk.edu.hk
[3]Tencent AI Lab
[3]{zptu,brightxwang,vinnylywang,shumingshi}@tencent.com

Abstract

Although neural machine translation (NMT) has advanced the state-of-the-art on various language pairs, the interpretability of NMT remains unsatisfactory. In this work, we propose to address this gap by focusing on understanding the input-output behavior of NMT models. Specifically, we measure the word importance by attributing the NMT output to every input word through a gradient-based method. We validate the approach on a couple of perturbation operations, language pairs, and model architectures, demonstrating its superiority on identifying input words with higher influence on translation performance. Encouragingly, the calculated importance can serve as indicators of input words that are under-translated by NMT models. Furthermore, our analysis reveals that words of certain syntactic categories have higher importance while the categories vary across language pairs, which can inspire better design principles of NMT architectures for multi-lingual translation.

1 Introduction

Neural machine translation (NMT) has achieved the state-of-the-art results on a mass of language pairs with varying structural differences, such as English-French (Bahdanau et al., 2014; Vaswani et al., 2017) and Chinese-English (Hassan et al., 2018). However, so far not much is known about how and why NMT works, which pose great challenges for debugging NMT models and designing optimal architectures.

The understanding of NMT models has been approached primarily from two complementary perspectives. The first thread of work aims to understand the importance of representations by analyzing the linguistic information embedded in representation vectors (Shi et al., 2016; Belinkov et al., 2017) or hidden units (Bau et al., 2019; Ding et al., 2017). Another direction focuses on understanding the importance of input words by interpreting the input-output behavior of NMT models. Previous work (Alvarez-Melis and Jaakkola, 2017) treats NMT models as black-boxes and provides explanations that closely resemble the attention scores in NMT models. However, recent studies reveal that attention does not provide meaningful explanations since the relationship between attention scores and model output is unclear (Jain and Wallace, 2019).

In this paper, we focus on the second thread and try to open the black-box by exploiting the gradients in NMT generation, which aims to estimate the word importance better. Specifically, we employ the *integrated gradients* method (Sundararajan et al., 2017) to attribute the output to the input words with the integration of first-order derivatives. We justify the gradient-based approach via quantitative comparison with black-box methods on a couple of perturbation operations, several language pairs, and two representative model architectures, demonstrating its superiority on estimating word importance.

We analyze the linguistic behaviors of words with the importance and show its potential to improve NMT models. First, we leverage the word importance to identify input words that are under-translated by NMT models. Experimental results show that the gradient-based approach outperforms both the best black-box method and other comparative methods. Second, we analyze the linguistic roles of identified important words, and find that words of certain syntactic categories have higher importance while the categories vary across language. For example, nouns are more important for Chinese⇒English translation, while prepositions are more important for English-French and -Japanese translation. This finding can inspire bet-

* Zhaopeng Tu is the corresponding author. Work was mainly done when Shilin He was interning at Tencent AI Lab.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 953–962, Hong Kong, China, November 3–7, 2019.

ter design principles of NMT architectures for different language pairs. For instance, a better architecture for a given language pair should consider its own language characteristics.

Contributions Our main contributions are:

- Our study demonstrates the necessity and effectiveness of exploiting the intermediate gradients for estimating word importance.
- We find that word importance is useful for understanding NMT by identifying under-translated words.
- We provide empirical support for the design principle of NMT architectures: essential inductive bias (e.g., language characteristics) should be considered for model design.

2 Related Work

Interpreting Seq2Seq Models Interpretability of Seq2Seq models has recently been explored mainly from two perspectives: interpreting internal representations and understanding input-output behaviors. Most of the existing work focus on the former thread, which analyzes the linguistic information embeded in the learned representations (Shi et al., 2016; Belinkov et al., 2017; Yang et al., 2019) or the hidden units (Ding et al., 2017; Bau et al., 2019). Several researchers turn to expose systematic differences between human and NMT translations (Läubli et al., 2018; Schwarzenberg et al., 2019), indicating the linguistic properties worthy of investigating. However, the learned representations may depend on the model implementation, which potentially limit the applicability of these methods to a broader range of model architectures. Accordingly, we focus on understanding the input-output behaviors, and validate on different architectures to demonstrate the universality of our findings.

Concerning interpreting the input-output behavior, previous work generally treats Seq2Seq models as black-boxes (Li et al., 2016; Alvarez-Melis and Jaakkola, 2017). For example, Alvarez-Melis and Jaakkola (2017) measure the relevance between two input-output tokens by perturbing the input sequence. However, they do not exploit any intermediate information such as gradients, and the relevance score only resembles attention scores. Recently, Jain and Wallace (2019) show that attention scores are in weak correlation with the feature importance. Starting from this observation, we exploit the intermediate gradients to better estimate word importance, which consistently outperforms its attention counterpart across model architectures and language pairs.

Exploiting Gradients for Model Interpretation The intermediate gradients have proven to be useful in interpreting deep learning models, such as NLP models (Mudrakarta et al., 2018; Dhamdhere et al., 2019) and computer vision models (Selvaraju et al., 2017; Sundararajan et al., 2017). Among all gradient-based approaches, the integrated gradients (IG, Sundararajan et al., 2017) is appealing since it does not need any instrumentation of the architecture and can be computed easily by calling gradient operations. In this work, we employ the IG method to interpret NMT models and reveal several interesting findings, which can potentially help debug NMT models and design better architectures for specific language pairs.

3 Approach

3.1 Neural Machine Translation

In machine translation task, a NMT model F: $\mathbf{x} \rightarrow \mathbf{y}$ maximizes the probability of a target sequence $\mathbf{y} = \{y_1, ..., y_N\}$ given a source sentence $\mathbf{x} = \{x_1, ..., x_M\}$:

$$P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) = \prod_{n=1}^{N} P(y_n|\mathbf{y}_{<n}, \mathbf{x}; \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is the model parameter and $\mathbf{y}_{<n}$ is a partial translation. At each time step n, the model generates an output word of the highest probability based on the source sentence $\mathbf{x}$ and the partial translation $\mathbf{y}_{<n}$. The training objective is to minimize the negative log-likelihood loss on the training corpus. During the inference, beam search is employed to decode a more optimal translation. In this study, we investigate the contribution of each input word x_m to the translated sentence $\mathbf{y}$.

3.2 Word Importance

In this work, the notion of "word importance" is employed to quantify the contribution that a word in the input sentence makes to the NMT generations. We categorize the methods of word importance estimation into two types: *black-box* methods without the knowledge of the model and *white-box* methods that have access to the model

internal information (e.g., parameters and gradients). Previous studies mostly fall into the former type, and in this study, we investigate several representative black-box methods:

- *Content Words*: In linguistics, all words can be categorized as either content or content-free words. Content words consist mostly of nouns, verbs, and adjectives, which carry descriptive meanings of the sentence and thereby are often considered as important.

- *Frequent Words*: We rank the relative importance of input words according to their frequency in the training corpus. We do not consider the top 50 most frequent words since they are mostly punctuation and stop words.

- *Causal Model* (Alvarez-Melis and Jaakkola, 2017): Since the causal model is complicated to implement and its scores closely resemble attention scores in NMT models. In this study, we use *Attention* scores to simulate the causal model.

Our approach belongs to the white-box category by exploiting the intermediate gradients, which will be described in the next section.

3.3 Integrated Gradients

In this work, we resort to a gradient-based method, integrated gradients (Sundararajan et al., 2017) (IG), which was originally proposed to attribute the model predictions to input features. It exploits the handy model gradient information by integrating first-order derivatives. IG is implementation invariant and does not require neural models to be differentiable or smooth, thereby is suitable for complex neural networks like Transformer. In this work, we use IG to estimate the word importance in an input sentence precisely.

Formally, let $\mathbf{x} = (x_1, ..., x_M)$ be the input sentence and $\mathbf{x}'$ be a baseline input. F is a well-trained NMT model, and $F(\mathbf{x})_n$ is the model output (i.e., $P(y_n|\mathbf{y}_{<n}, \mathbf{x})$) at time step n. Integrated gradients is then defined as the integral of gradients along the straightline path from the baseline $\mathbf{x}'$ to the input $\mathbf{x}$. In detail, the contribution of the m^{th} word in $\mathbf{x}$ to the prediction of $F(\mathbf{x})_n$ is defined as follows.

$$IG^n_m(\mathbf{x}) = (\mathbf{x}_m - \mathbf{x}'_m) \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))_n}{\partial \mathbf{x}_m} \mathrm{d}\alpha$$

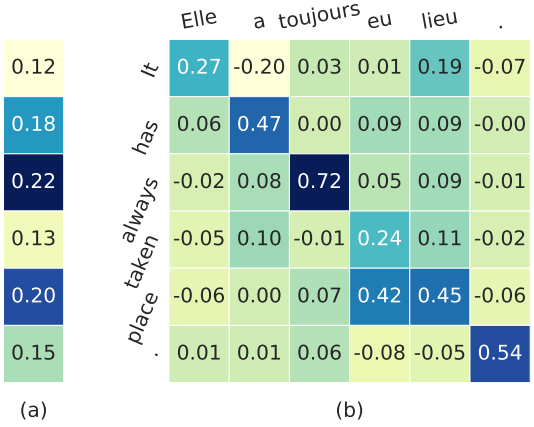


Figure 1: An example of (a) word importance and (b) contribution matrix calculated by *Attribution* (integrated gradients) on English⇒French translation task. Input in English: "It has always taken place ." Output in French: "Elle a toujours eu lieu ."

where $\frac{\partial F(\mathbf{x})_n}{\partial \mathbf{x}_m}$ is the gradient of $F(\mathbf{x})_n$ w.r.t. the embedding of the m^{th} word. In this paper, as suggested, the baseline input $\mathbf{x}'$ is set as a sequence of zero embeddings that has the same sequence length M. In this way, we can compute the contribution of a specific input word to a designated output word. Since the above formula is intractable for deep neural models, we approximate it by summing the gradients along a multi-step path from baseline $\mathbf{x}'$ to the input $\mathbf{x}$.

$$IG^n_m(\mathbf{x}) = \frac{(\mathbf{x}_m - \mathbf{x}'_m)}{S} \sum_{k=0}^{S} \frac{\partial F(\mathbf{x}' + \frac{k}{S}(\mathbf{x} - \mathbf{x}'))_n}{\partial \mathbf{x}_m}$$

where S denotes the number of steps that are uniformly distributed along the path. The IG will be more accurate if a larger S is used. In our preliminary experiments, we varied the steps and found 300 steps yielding fairly good performance.

Following the formula, we can calculate the contribution of every input word makes to every output word, forming a contribution matrix of size $M \times N$, where N is the output sentence length. Given the contribution matrix, we can obtain the *word importance* of each input word to the entire output sentence. To this end, for each input word, we first aggregate its contribution values to all output words by the *sum* operation, and then normalize all sums through the *Softmax* function. Figure 1 illustrates an example of the calculated word importance and the contribution matrix, where an English sentence is translated into a French sentence using the Transformer model. A negative

contribution value indicates that the input word has negative effects on the output word.

4 Experiment

Data To make the conclusion convincing, we first choose two large-scale datasets that are publicly available, i.e., Chinese-English and English-French. Since English, French, and Chinese all belong to the subject-verb-object (SVO) family, we choose another very different subject-object-verb (SOV) language, Japanese, which might bring some interesting linguistic behaviors in English-Japanese translation.

For Chinese-English task, we use WMT17 Chinese-English dataset that consists of 20.6M sentence pairs. For English-French task, we use WMT14 English-French dataset that comprises 35.5M sentence pairs. For English-Japanese task, we follow (Morishita et al., 2017) to use the first two sections of WAT17 English-Japanese dataset that consists of 1.9M sentence pairs. Following the standard NMT procedure, we adopt the standard byte pair encoding (BPE) (Sennrich et al., 2016) with 32K merge operations for all language pairs. We believe that these datasets are large enough to confirm the rationality and validity of our experimental analyses.

Implementation We choose the state-of-the-art Transformer (Vaswani et al., 2017) model and the conventional RNN-Search model (Bahdanau et al., 2014) as our test bed. We implement the *Attribution* method based on the Fairseq-py (Gehring et al., 2017) framework for the above models. All models are trained on the training corpus for 100k steps under the standard settings, which achieve comparable translation results. All the following experiments are conducted on the test dataset, and we estimate the input word importance using the model generated hypotheses.

In the following experiments, we compare IG (*Attribution*) with several black-box methods (i.e., *Content*, *Frequency*, *Attention*) as introduced in Section 3.2. In Section 4.1, to ensure that the translation performance decrease attributes to the selected words instead of the perturbation operations, we randomly select the same number of words to perturb (*Random*), which serves as a baseline. Since there is no ranking for content words, we randomly select a set of content words as important words. To avoid the potential bias introduced by randomness (i.e., *Random* and *Content*), we repeat the experiments for 10 times and report the averaged results. We calculate the *Attention* importance in a similar manner as the *Attribution*, except that the attention scores use a *max* operation due to the better performance.

Evaluation We evaluate the effectiveness of estimating word importance by the translation performance decrease. More specifically, unlike the usual way, we measure the decrease of translation performance when perturbing a set of important words that are of top-most word importance in a sentence. The more translation performance degrades, the more important the word is.

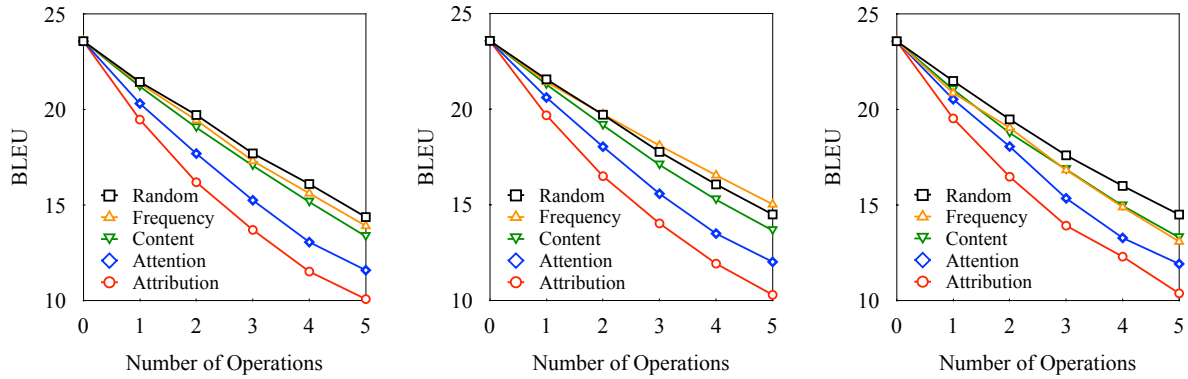
We use the standard BLEU score as the evaluation metric for translation performance. To make the conclusion more convincing, we conduct experiments on different types of synthetic perturbations (Section 4.1), as well as different NMT architectures and language pairs (Section 4.2). In addition, we compare with a supervised erasure method, which requires ground-truth translations for scoring word importance (Section 4.3).

4.1 Results on Different Perturbations

In this experiment, we investigate the effectiveness of word importance estimation methods under different synthetic perturbations. Since the perturbation on text is notoriously hard (Zhang et al., 2019) due to the semantic shifting problem, in this experiment, we investigate three types of perturbations to avoid the potential bias :

- *Deletion* perturbation removes the selected words from the input sentence, and it can be regarded as a specific instantiation of sentence compression (Cohn and Lapata, 2008).
- *Mask* perturbation replaces embedding vectors of the selected words with all-zero vectors (Arras et al., 2016), which is similar to *Deletion* perturbation except that it retains the placeholder.
- *Grammatical Replacement* perturbation replaces a word by another word of the same linguistic role (i.e., POS tags), yielding a sentence that is grammatically correct but semantically nonsensical (Chomsky and Lightfoot, 2002; Gulordava et al., 2018), such as "*colorless green ideas sleep furiously*".

Figure 2 illustrates the experimental results on Chinese⇒English translation with Transformer. It



(a) **Deletion** (b) **Mask** (c) **Grammatical Replacement**

Figure 2: Effect of three types of synthetic perturbations on Chinese⇒English translation using the Transformer.

shows that *Attribution* method consistently outperforms other methods against different perturbations on a various number of operations. Here the operation number denotes the number of perturbed words in a sentence. Specifically, we can make the following observations.

Important words are more influential on translation performance than the others. Under three different perturbations, perturbing words of top-most importance leads to lower BLEU scores than *Random* selected words. It confirms the existence of important words, which have greater impacts on translation performance. Furthermore, perturbing important words identified by *Attribution* outperforms the *Random* method by a large margin (more than 4.0 BLEU under 5 operations).

The gradient-based method is superior to comparative methods (e.g., Attention) in estimating word importance. Figure 2 shows that two *black-box* methods (i.e., *Content*, *Frequency*) perform only slightly better than the *Random* method. Specifically, the *Frequency* method demonstrates even worse performances under the *Mask* perturbation. Therefore, linguistic properties (such as POS tags) and the word frequency can only partially help identify the important words, but it is not as accurate as we thought. In the meanwhile, it is intriguing to explore what exact linguistic characteristics these important words reveal, which will be introduced in Section 5.

We also evaluate the *Attention* method, which bases on the encoder-decoder attention scores at the last layer of Transformer. Note that the *Attention* method is also used to simulate the best *black-box* method SOCRAT, and the results show that it is more effective than *black-box* methods and the *Random* baseline. Given the powerful *Attention* method, *Attribution* method still achieves best performances under all three perturbations. Furthermore, we find that the gap between *Attribution* and *Attention* is notably large (around 1.0+ BLEU difference). *Attention* method does not provide as accurate word importance as the *Attribution*, which exhibits the superiority of gradient-based methods and consists with the conclusion reported in the previous study (Jain and Wallace, 2019).

In addition, as shown in Figure 2, the perturbation effectiveness of *Deletion*, *Mask*, and *Grammatical Replacement* varies from strong to weak. In the following experiments, we choose *Mask* as the representative perturbation operation for its moderate perturbation performance, based on which we compare two most effective methods *Attribution* and *Attention*.

4.2 Results on Different NMT Architecture and Language Pairs

Different NMT Architecture We validate the effectiveness of the proposed approach using a different NMT architecture RNN-Search on the Chinese⇒English translation task. The results are shown in Figure 3(a). We observe that the *Attribution* method still outperforms both *Attention* method and *Random* method by a decent margin. By comparing to Transformer, the results also reveal that the RNN-Search model is less robust to these perturbations. To be specific, under the setting of five operations and *Attribution* method, Transformer shows a relative decrease of 55% on BLEU scores while the decline of RNN-Search model is 64%.

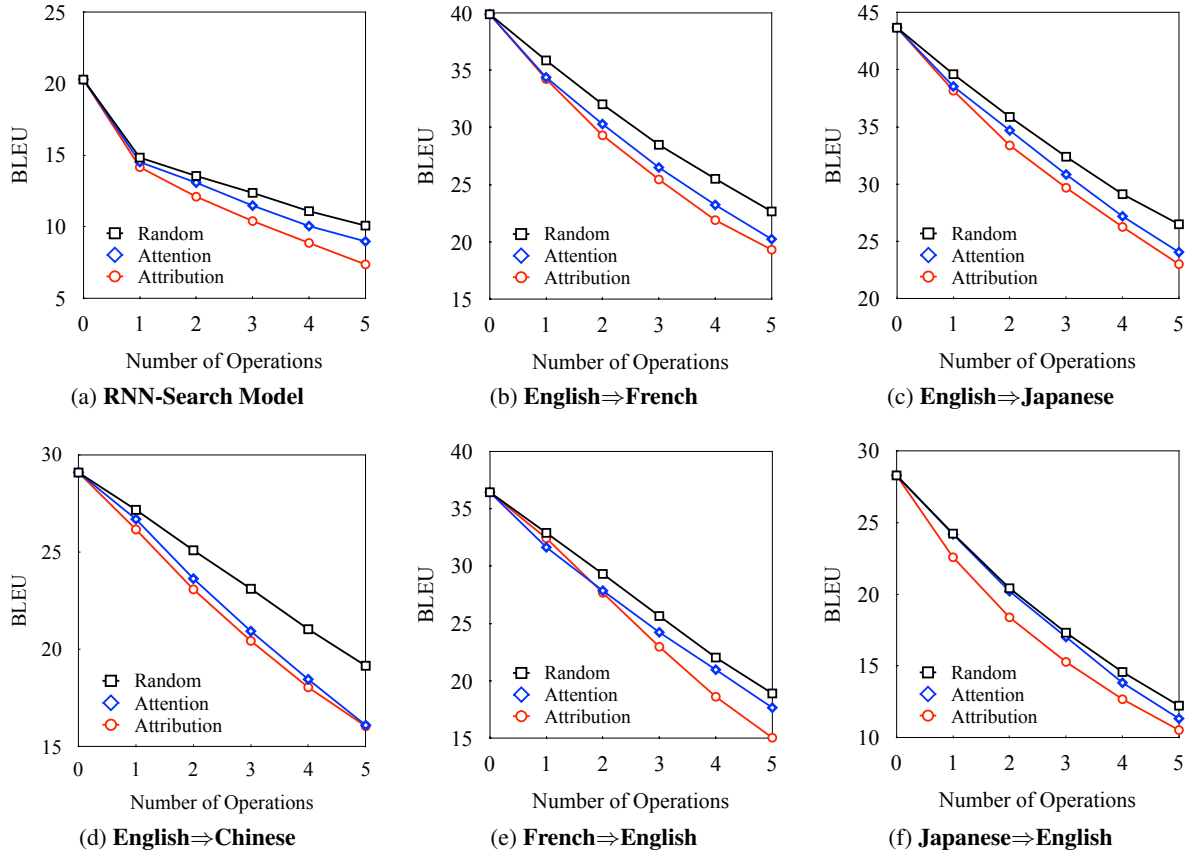


Figure 3: Effect of the *Mask* perturbation on (a) Chinese⇒English translation using the RNN-Search model, (b, c, d, e, f) other language pairs and directions using Transformer model.

Different Language Pairs and Directions We further conduct experiments on another two language pairs (i.e., English⇒French, English⇒Japanese in Figures 3(b, c)) as well as the reverse directions (Figures 3(d, e, f)) using Transformer under the *Mask* perturbation. In all the cases, *Attribution* shows the best performance while *Random* achieves the worst result. More specifically, *Attribution* method shows similar translation quality degradation on all three language-pairs, which declines to around the half of the original BLEU score with five operations.

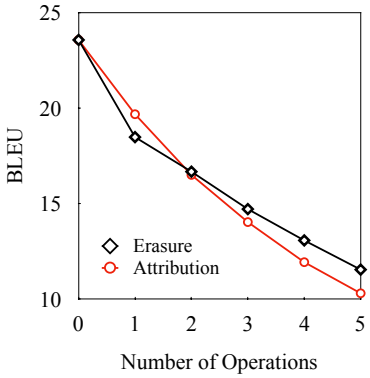


Figure 4: Effect of *Attribution* and *Erasure* methods on Chinese⇒English translation with *Mask* perturbation.

4.3 Comparison with Supervised Erasure

There exists another straightforward method, *Erasure* (Alvarez-Melis and Jaakkola, 2017; Arras et al., 2016; Zintgraf et al., 2017), which directly evaluates the word importance by measuring the translation performance degradation of each word. Specifically, it erases (i.e., *Mask*) one word from the input sentence each time and uses the BLEU score changes to denote the word importance (after normalization).

In Figure 4, we compare *Erasure* method with *Attribution* method under the *Mask* perturbation. The results show that *Attribution* method is less effective than *Erasure* method when only one word is perturbed. But it outperforms the *Erasure* method when perturbing 2 or more words. The results reveal that the importance calculated by erasing only one word cannot be generalized to multiple-words scenarios very well. Besides, the

Method	Top 5%	Top 10%	Top 15%
Attention	0.058	0.077	0.119
Erasure	0.154	0.170	0.192
Attribution	**0.248**	**0.316**	**0.342**

Table 1: F1 accuracy of detecting under-translation errors with the estimated word importance.

Erasure method is a supervised method which requires ground-truth references, and finding a better words combination is computation infeasible when erasing multiple words.

We close this section by pointing out that our gradient-based method consistently outperforms its black-box counterparts in various settings, demonstrating the effectiveness and universality of exploiting gradients for estimating word importance. In addition, our approach is on par with or even outperforms the supervised erasure method (on multiple-word perturbations). This is encouraging since our approach does not require any external resource and is fully unsupervised.

5 Analysis

In this section, we conduct analyses on two potential usages of word importance, which can help debug NMT models (Section 5.1) and design better architectures for specific languages (Section 5.2). Due to the space limitation, we only analyze the results of Chinese⇒English, English⇒French, and English⇒Japanese. We list the results on the reverse directions in Appendix, in which the general conclusions also hold.

5.1 Effect on Detecting Translation Errors

In this experiment, we propose to use the estimated word importance to detect the under-translated words by NMT models. Intuitively, under-translated input words should contribute little to the NMT outputs, yielding much smaller word importance. Given 500 Chinese⇒English sentence pairs translated by the Transformer model (BLEU 23.57), we ask ten human annotators to manually label the under-translated input words, and at least two annotators label each input-hypothesis pair. These annotators have at least six years of English study experience, whose native language is Chinese. Among these sentences, 178 sentences have under-translation errors with 553 under-translated words in total.

Table 1 lists the accuracy of detecting under-translation errors by comparing words of *least* importance and human-annotated under-translated words. As seen, our *Attribution* method consistently and significantly outperforms both *Erasure* and *Attention* approaches. By exploiting the word importance calculated by *Attribution* method, we can identify the under-translation errors automatically without the involvement of human interpreters. Although the accuracy is not high, it is worth noting that our under-translation method is very simple and straightforward. This is potentially useful for debugging NMT models, e.g., automatic post-editing with constraint decoding (Hokamp and Liu, 2017; Post and Vilar, 2018).

Type		Zh⇒En	En⇒Fr	En⇒Ja
POS Tags	Noun	**21.0%**	1.9%	0.7%
	Verb	0.3%	**25.0%**	0.3%
	Adj.	0.4%	9.3%	0.7%
	Prep.	1.3%	4.5%	**26.7%**
	Dete.	3.0%	5.7%	2.1%
	Punc.	3.5%	**18.3%**	**30.5%**
	Others	0.5%	1.2%	4.7%
Fertility	≥ 2	**50.2%**	**21.4%**	**21.7%**
	1	**15.4%**	7.0%	3.1%
	$(0, 1)$	2.5%	0.4%	3.0%
	0	0.0%	1.9%	3.8%
Syntactic	Low	1.6%	2.5%	1.2%
	Middle	0.3%	0.8%	1.4%
	High	0.0%	0.1%	0.1%

Table 2: Correlation between *Attribution* word importance with POS tags, Fertility, and Syntactic Depth. Fertility can be categorized into 4 types: one-to-many ("≥ 2"), one-to-one ("1"), many-to-one ("$(0, 1)$"), and null-aligned ("0"). Syntactic depth shows the depth of a word in the dependency tree. A lower tree depth indicates closer to the root node in the dependency tree, which might indicate a more important word.

5.2 Analysis on Linguistic Properties

In this section, we analyze the linguistic characteristics of important words identified by the attribution-based approach. Specifically, we investigate several representative sets of linguistic properties, including POS tags, and fertility, and depth in a syntactic parse tree. In these analyses, we multiply the word importance with the corresponding sentence length for fair comparison. We use a decision tree based regression model to calculate the correlation between the importance and linguistic properties.

Type		Chinese⇒English			English⇒French			English⇒Japanese		
		Count	Attri.	△	Count	Attri.	△	Count	Attri.	△
Content	Noun	0.383	0.407	*+6.27%*	0.341	0.355	*+4.11%*	0.365	0.336	*-7.95%*
	Verb	0.165	0.160	*-3.03%*	0.146	0.131	*-10.27%*	0.127	0.123	*-3.15%*
	Adj.	0.032	0.029	*-9.38%*	0.076	0.072	*-5.26%*	0.094	0.088	*-6.38%*
	Total	0.579	0.595	*+2.76%*	0.563	0.558	*-0.89%*	0.587	0.547	*-6.81%*
Content-Free	Prep.	0.056	0.051	*-8.93%*	0.120	0.132	*+10.00%*	0.129	0.151	*+17.05%*
	Dete.	0.043	0.043	*0.00%*	0.102	0.101	*-0.98%*	0.112	0.103	*-8.04%*
	Punc.	0.137	0.131	*-4.38%*	0.100	0.091	*-9.00%*	0.096	0.120	*+25.47%*
	Others	0.186	0.179	*-3.76%*	0.115	0.118	*+2.61%*	0.076	0.079	*+3.95%*
	Total	0.421	0.405	*-3.80%*	0.437	0.442	*+1.14%*	0.413	0.453	*+9.69%*

Table 3: Distribution of syntactic categories (e.g. content words vs. content-free words) based on word count ("Count") and *Attribution* importance ("Attri."). "△" denotes relative change over the count-based distribution.

Fertility	Chinese⇒English			English⇒French			English⇒Japanese		
	Count	Attri.	△	Count	Attri.	△	Count	Attri.	△
≥ 2	0.087	0.146	*+67.82%*	0.126	0.138	*+9.52%*	0.117	0.143	*+22.22%*
1	0.621	0.622	*+0.16%*	0.672	0.670	*-0.30%*	0.570	0.565	*-0.88%*
$(0, 1)$	0.115	0.081	*-29.57%*	0.116	0.113	*-2.59%*	0.059	0.055	*-6.78%*
0	0.176	0.150	*-14.77%*	0.086	0.079	*-8.14%*	0.254	0.237	*-6.69%*

Table 4: Distributions of word fertility and their relative change based on *Attribution* importance and word count.

Table 2 lists the correlations, where a higher value indicates a stronger correlation. We find that the syntactic information is almost independent of the word importance value. Instead, the word importance strongly correlates with the POS tags and fertility features, and these features in total contribute over 95%. Therefore, in the following analyses, we mainly focus on the POS tags (Table 3) and fertility properties (Table 4). For better illustration, we calculate the distribution over the linguistic property based on both the *Attribution* importance ("Attr.") and the word frequency ("Count") inside a sentence. The larger the relative increase between these two values, the more important the linguistic property is.

Certain syntactic categories have higher importance while the categories vary across language pairs. As shown in Table 3, *content* words are more important on Chinese⇒English but *content-free* words are more important on English⇒Japanese. On English⇒French, there is no notable increase or decrease of the distribution since English and French are in essence very similar. We also obtain some specific findings of great interest. For example, we find that noun is more important on Chinese⇒English translation, while preposition is more important on English⇒French translation. More interestingly, English⇒Japanese translation shows a substantial discrepancy in contrast to the other two language pairs. The results reveal that preposition and punctuation are very important in English⇒Japanese translation, which is counter-intuitive.

Punctuation in NMT is understudied since it carries little information and often does not affect the understanding of a sentence. However, we find that punctuation is important on English⇒Japanese translation, whose proportion increases dramatically. We conjecture that it is because the punctuation could affect the sense groups in a sentence, which further benefits the syntactic reordering in Japanese.

Words of high fertility are always important. We further compare the fertility distribution based on word importance and the word frequency on three language pairs. We hypothesize that a source word that corresponds to multiple target words should be more important since it contributes more to both sentence length and BLEU score.

Table 4 lists the results. Overall speaking, one-to-many fertility is consistently more important on all three language pairs, which confirms our hypothesis. On the contrary, null-aligned words receive much less attention, which shows a persis-

tently decrease on three language pairs. It is also reasonable since null-aligned input words contribute almost nothing to the translation outputs.

6 Discussion and Conclusion

We approach understanding NMT by investigating the word importance via a gradient-based method, which bridges the gap between word importance and translation performance. Empirical results show that the gradient-based method is superior to several black-box methods in estimating the word importance. Further analyses show that important words are of distinct syntactic categories on different language pairs, which might support the viewpoint that essential inductive bias should be introduced into the model design (Strubell et al., 2018). Our study also suggests the possibility of detecting the notorious under-translation problem via the gradient-based method.

This paper is an initiating step towards the general understanding of NMT models, which may bring some potential improvements, such as

- *Interactive MT and Constraint Decoding* (Foster et al., 1997; Hokamp and Liu, 2017): The model pays more attention to the detected unimportant words, which are possibly under-translated;
- *Adaptive Input Embedding* (Baevski and Auli, 2019): We can extend the adaptive softmax (Grave et al., 2017) to the input embedding of variable capacity – more important words are assigned with more capacity;
- *NMT Architecture Design*: The language-specific inductive bias (e.g., different behaviors on POS) should be incorporated into the model design.

We can also explore other applications of word importance to improve NMT models, such as more tailored training methods. In general, model interpretability can build trust in model predictions, help error diagnosis and facilitate model refinement. We expect our work could shed light on the NMT model understanding and benefit the model improvement.

There are many possible ways to implement the general idea of exploiting gradients for model interpretation. The aim of this paper is not to explore this whole space but simply to show that some fairly straightforward implementations work well. Our approach can benefit from advanced exploitation of the gradients or other useful intermediate information, which we leave to the future work.

Acknowledgement

Shilin He and Michael R. Lyu were supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14210717 of the General Research Fund), and Microsoft Research Asia (2018 Microsoft Research Asia Collaborative Research Award). We thank the anonymous reviewers for their insightful comments and suggestions.

References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *EMNLP*.

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*.

Alexei Baevski and Michael Auli. 2019. Adaptive input representations for neural language modeling. In *ICLR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *ICLR*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *ACL*.

Noam Chomsky and David W Lightfoot. 2002. *Syntactic structures*. Walter de Gruyter.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *COLING*.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. 2019. How important is a neuron? In *ICLR*.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *ACL*.

George Foster, Pierre Isabelle, and Pierre Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12(1/2):175–194.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

Édouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. 2017. Efficient softmax approximation for GPUs. In *ICML*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *NAACL*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. In *arXiv:1803.05567*.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *ACL*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *EMNLP*.

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

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2017. Ntt neural machine translation systems at wat 2017. In *WAT*.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *ACL*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *NAACL*.

Robert Schwarzenberg, David Harbecke, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. 2019. Train, sort, explain: Learning to diagnose translation models. In *NAACL*.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *EMNLP*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-Informed Self-Attention for Semantic Role Labeling. In *EMNLP*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *ICML*.

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 *NeurIPS*.

Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019. Assessing the ability of self-attention networks to learn word order. In *ACL*.

Wei Emma Zhang, Quan Z Sheng, and Ahoud Abdulrahmn F Alhazmi. 2019. Generating textual adversarial examples for deep learning models: A survey. In *arXiv preprint arXiv:1901.06796*.

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