A Primer in NMTology: What we have understood about NMT

Anonymous EMNLP submission

Abstract

Neural Machine Translation (NMT) has been through great revolutions in recent years. Accompanied with improvements in translation quality are works that attempted to understand the working mechanism of various aspects of the NMT framework. In our paper, we survey those efforts on unveiling the *black box* of the standard NMT framework. To begin with, we briefly introduce the three critical components of the holistic NNT framework; nextly, we deliver a clear *component-centric* categorization and clean summary of these specific works *guided* by *frequently-asked* questions (FAQs) that aim at making up *lack* of understanding; finally, we discuss several limitations, future directions and inspirations. We believe this paper could facilitate the community to weave a holistic and clear picture of our current understandings of the standard NMT framework and shed light on its future improvements and developments. Please check this website **https://nmtology.github.io/** for a visual guidance of the FAQs.

1 Introduction

Machine Translation is an extremely challenging task. Statistical Machine Translation (SMT), which models translation in a pipelined manner, was historically one of the popular approaches (Koehn et al., 2003; Chiang, 2005). In the SMT pipeline, each module plays a clear role and is parameterized by a relatively simple model, leading to easy interpretability. Recently, Neural Machine Translation (NMT) framework establishes new state-of-the-art performances (Barrault et al., 2019, 2020). The strengths of NMT come from its strong modeling power with *complex* deep encoder-decoder architecture and *holistic* end-to-end training, which lead to poor interpretability. Consequently, the poor interpretability prevents us from elegantly debugging the model, trusting its outputs, and particularly further improving performance (Ding et al., 2017).

This paper conducts a thorough survey on understanding components of the NMT framework, covering a hundred papers published in recent years. Our survey is *component-centric*, that is, we organize related papers in terms of every NMT component and highlight important questions *frequently-asked w.r.t.* that component. We want our readers to treat this paper as instructional FAQs about understanding the black-box of the NMT framework, so they can quickly zoom into certain question and find the corresponding papers to complement their lack of understanding. §2 briefly introduces every components of the NMT framework, while §3, §4, §5, §6 summarize works in terms of model architecture, training, inference and behavior. In §7, we discuss limitations and future directions.

Related surveys Lertvittayakumjorn and Toni (2021); Danilevsky et al. (2020); Luo et al. (2021) are more general surveys on principles for explainable NLP, as they mainly discuss general desiderata and possible explanation paradigms or frameworks. Sajjad et al. (2021) survey specific neuron-level interpretation methods for NLP models, while Belinkov and Glass (2019) and Sun et al. (2021) focus on surveying a broad range of general techniques and methods for interpreting NLP models on various tasks. The closest work in organization to us might be Rogers et al. (2020). They deliver a thorough survey on research questions, directions, and solutions around large pretrained models. In our paper, we organize research works guided by research questions that are more related to the interpretation and understanding side, so that researchers can gain in-depth insights in various components and learning phenomena of the NMT framework.

2 The NMT Framework

The NMT framework is proposed as a sequence-to-sequence transduction task (Sutskever et al., 2014). To make the framework clean to readers, we divide it roughly into three independent and indispensable modules: *i)* model architecture; *ii)* model training, and *iii)* inference mechanism.

Model architecture The NMT model is usually implemented with the encoder-decoder with attention architecture. Recurrent neural networks are used to parameterize the model (Bahdanau et al., 2014; Wu et al., 2016). Then convolutional and self-attention neural networks are proposed respectively (Gehring et al., 2017; Vaswani et al., 2017).

Model training Typical training uses maximum likelihood estimation (MLE) to minimize the negative log-likelihood: $\mathcal{L}(\theta) = -\log P_\theta(y|x) = -\sum_t \log P_\theta(y_t|x, y_{<t})$. Reinforcement Learning (RL) is also leveraged for optimizing the evaluation metric based loss: $\mathcal{L}(\theta) = -\mathbb{E}_{\hat{y}\sim P_\theta(y|x)} m(\hat{y}, y^*)$, where $m(\cdot, y^*)$ denotes certain evaluation metric, e.g., BLEU. Like RL, minimum risk training (MRT) is also used to optimize metric (Shen et al., 2016). Besides, many tricks such as learning rate schedule, normalization techniques, and label smoothing are also used. For Non-Autoregresstive neural machine Translation (NAT), knowledge distillation (KD) is used for performance boosting.

Model inference Beam search is used to find an approximate solution of the $\hat{y} = \arg_y \max P_\theta(y|x)$ problem. Due to several issues of the vanilla beam search, tricks like the length penalty are proposed.

3 Understanding Model Architecture

3.1 Understanding encoder/decoder

Q. Does encoder's representation entail linguistic knowledge? Most of the works on this topic use certain linguistic tasks to assess the power of the learned hidden representations. Early on, Shi et al. (2016) begin to answer whether string-based NMT models learn about source syntax. They test hidden states' ability to predict syntactic labels, e.g., voice, tense, smallest phrase constituent. Belinkov et al. (2017) deliver thorough analyses using the method of probing on what encoder learns about morphology knowledge of source languages. Main conclusions like the depth of the layer, the input representation (word, character), and the language types are important factors influencing the learned knowledge of morphology can be drawn from their analyses. Belinkov et al. (2020a) arrange together the analyses on the power of learned representation across various granularities of linguistic knowledge based on probing mainly for the encoder. They further add syntactic and semantic tasks . Bisazza and Tump (2018) also study morphology knowledge captured by embedding layer.

Q. Can encoder learn word sense disambiguation? Several works attempt to understand word sense disambiguation (WSD) ability of NMT models. Rios Gonzales et al. (2017) construct a contrastive dataset where references are accompanied with a rewritten one that has an incorrect translation of a source ambiguous word. They find that the model ranks 70% of such contrastive pairs correctly, indicating the model's strong WSD ability. Marvin and Koehn (2018) further investigate the hidden activations' WSD ability through visualization of hidden vector clusters. Tang et al. (2019a) take the encoder as a whole for WSD ability analyses under different model architectures via probing.

Q. Does decoder's representation entail linguistic knowledge? Belinkov et al. (2017) study the linguistic properties of decoder's representation compared to encoder's. They probe and find that decoder's representation falls back a lot in accuracy of predicting POS tags. In contrast, in their later work (Belinkov et al., 2020a), they find that decoder's representation is similar to or better than encoder's for morphological tag prediction. Instead, Li et al. (2019a) study the possibly learned coarse-to-fine characteristics of decoder's layer-wise representation with probes on hierarchical probing tasks.

Q. Can a single neuron entail linguistic knowledge? Instead of taking vector representation as a whole, Bau et al. (2019) leverage an unsupervised method to identify important neurons and use GMM to find neurons that controls linguistic features in prediction. Dalvi et al. (2019) also propose supervised methods to extract salient neurons and analyze their linguistic properties through probing.

Q. Can linguistic knowledge be preserved after pruning? Movva and Zhao (2020) study the representation of modules of the Transformer model while being pruned. They observe that pruning degrades semantic knowledge before affecting BLEU, and representation in higher layers changes most.

Q. Which component of NMT is more critical, encoder or decoder? Tang et al. (2019b) attempt to reveal the representational power of the encoder by removing it, so as a result, the encoder is just word and position embeddings. They find that the non-contextualized encoder representation largely degrades performance; however, the attention module complements this as a strong feature extractor. Kasai et al. (2020) study encoder/decoder with varied depths. They find that a sufficiently deep encoder with a single-layer decoder can achieve comparable performance with balanced layer depth.

3.2 Understanding attention

3.2.1 Cross-attention

Q. Does attention learn alignments? When attention was first introduced into NMT models (Bahdanau et al., 2014), it was believed as a word alignment module inside NMT. Liu et al. (2016); Mi et al. (2016); Li et al. (2018); Baan et al. (2019) try to improve NMT's translation quality by improving the alignment performance of its attention module. However, the attention module in NMT was far from qualified as a good word aligner compared with statistical word aligners (Koehn and Knowles, 2017; Li et al., 2019b). Although the AER of attention is dissatisfactory, Li et al. (2019b) did successfully induce decent alignments from NMT models by the method Prediction Difference (PD). Notably, Li et al. (2019b) empirically showed that, towards predictions instead of references, the performance of alignments induced by PD could surpass well-performed traditional statistical aligners. This result rekindled the confidence in inducing accurate alignments from the attention module. By improving training (Garg et al., 2019) and modeling (Alkhouli et al., 2018; Chen et al., 2020; Kobayashi et al., 2020) methods, the alignment performance of NMT's attention are constantly improved. In the situation where translation quality is not as important as alignment performance, attention can also be extremely helpful in building a well-performed word aligner (Zenkel et al., 2020).

Q. Do attention weights reflect NMT's reasoning? Since Bahdanau et al. (2014) introduced attention to NMT, attention weights were often claimed to explain the inner-working mechanism of neural models (Li et al., 2016). Jain and Wallace (2019) are the first to question attention's ability to provide transparency for model predictions by showing a weak correlation between intuitive feature importance measures and attention weights in text classification, question answering, and natural language inference tasks. However, Wiegreffe and Pinter (2019) argue that Jain and Wallace (2019) does not disprove the usefulness of attention for explainability by showing the attention weights cannot be easily hacked adversarially. Based on this observation, Moradi et al. (2020) provide a measure of the faithfulness of NMT and an adversarial regularization that can lead to more trustworthy attention heatmaps without reducing the translation quality. Current analyses are mainly focused on simpler single-head RNN based models. In the future, checking whether the current understanding holds on multi-head attention of Transformer could be an interesting direction.

3.2.2 Self-attention

Q. Is self-attention network better than RNN? The common suspicion is that self-attention can connect distant words via shorter network paths than RNNs to improve the ability to model long-range dependencies. However, this theoretical argument is not tested empirically. Tang et al. (2018) evaluate RNNs, CNNs, and multi-head attention networks (SAN) on two tasks: subject-verb agreement and word sense disambiguation to measure the ability to extract semantic features from the source text. Their experimental results show that the SAN performs distinctly better than RNNs and CNNs on word sense disambiguation. However, all of them are similar in modeling subject-verb agreement over long distances. Besides, SAN is ascribed to be weak at learning positional information of words for sequence modeling compared to the models with recurrence structure. Yang et al. (2019) show that although SAN trained on word reordering detection has difficulty learning positional information, SAN trained on machine translation learns better positional information than RNN.

Q. Is multi-head better than single-head? In Transformer, multi-head attention strengthens the expressive power of a model by extending a single head to multiple parallel heads. From a Bayesian perspective, An et al. (2020) understands why one needs multi-head attention by showing it is equivalent to using more samples to approximate an underlying posterior distribution. Snell et al. (2021) explain why attention obtained by MLE often correlates well with saliency and how attention can increase performance by improving its training dynamics rather than expressiveness. Raganato et al. (2020) deliver a finding that for the encoder's multi-head attention, fixing other heads' weight and only learning one head can achieve similar performance in high-resource translation tasks and even improve performance up to 3.5 BLEU points in low-resource scenarios. Behnke and Heafield (2020) propose simple heuristics for pruning attention heads at the early stage of training. It confirms that most attention heads are not confident in their decisions. Michel et al. (2019); Voita et al. (2019b); Liu et al. (2021) empirically show that multi-heads are redundant at test time but are greatly helpful in training. This opens up many opportunities for downsizing these humongous models for inference.

4 Understanding Training

4.1 Training data

Q. How does data noise affect NMT? Noise in bitext corpus impacts NMT a lot. Khayrallah and Koehn (2018) investigate the impact of various types of noise of the training data on the performance of the NMT model and an SMT model. By adding many controlled types of noise to the original high-quality data, they find that the NMT model is very susceptible to noise and can degrade up to 9 BLEU points, whereas the SMT model can even obtain 1 BLUE improvement. They build five types of noise and analyze how these noises can impact translation quality. They find copy noise, where the target is just the copy of the source, is most harmful. Ott et al. (2018) reemphasize the harmfulness of copy noise in training data. They also find that beam search puts too much probability mass over the whole search space due to data uncertainty, not concentrating on accurate and relevant translations.

Q. How does the src/tgt divergence affect NMT? Briakou and Carpuat (2021) study fine-grained semantic divergences in bitext. They propose three typical divergences, lexical substitution, phrase replacement, and subtree deletion. They study their effects on NMT and find subtree deletion degrade performance the most. In a semi-supervised setting, due to extra monolingual data, the textual domains of src/tgt might exhibit topic divergence. Shen et al. (2021) propose a metric to measure such mismatch phenomenon and study its effects, particularly with varying data scales and find it can severely degrade performance in a low-resource setting.

Q. Why does DA training help? Data Augmentation (DA) methods are effective in training NMT with few theory-oriented understanding. Li et al. (2019a) borrow empirical evidence that input sensitivity and prediction margin can measure generalization ability from the learning theory community and apply them to test intrinsic changes of the model before and after DA. DA methods generally lead to better insensitivity and a larger margin.

Q. What factors of BT data matter? Amongst all DA methods, Back-Translation (BT) is the most extensively adopted one in challenges and deployments to obtain state-of-the-art translation quality. Edunov et al. (2018a) conducts a large-scale analysis of practical BT training. They argue that *randomness* is an essential factor for improving performance, so they use sampling rather than beam search to obtain pseudo bitext. However, Caswell et al. (2019) argues that randomness might not be the reason for better practice in synthetic data generation in BT. They claim that the NMT model can automatically distinguish synthetic or real data and learn different attention patterns over them. So they propose tagged BT to improve standard BT. Following this work, Marie et al. (2020) further proves that tagged BT can prevent the NMT model from over-fitting to those machine-generated data. Besides, Graça et al. (2019) proposes a math interpretation of back-translation, which links BT to variational inference and motivates multi-turn BT.

4.2 Training loss

Q. What are the issues of NLL? *Negative log-likelihood* (NLL) loss is the default loss function to train NMT models with MLE. NLL is a token-level loss that is locally normalized and defined on ground-truth prefix. Such characteristics make NLL suffer from the following issues as discussed in Ranzato et al. (2015); Wiseman and Rush (2016): i) *exposure bias*: the model is never exposed to its own errors during training, and so the inferred histories at test-time do not resemble the gold training histories; ii) *train-test mismatch*: training uses a token-level loss, while at test-time, we target improving sequence-level evaluation metrics, such as BLEU; iii) *label bias*: the model score is locally normalized at the token level, whereas the search algorithm cares about the sequence level score. Edunov et al. (2018b) investigate other token-level loss choices such as margin-based losses and find they do not lead to significant improvement over NLL. Afterward, a large set of works have tried to propose methods based on RL to overcome the above three issues, though they seem to leave NLL unshakeable (Bengio et al., 2015; Shen et al., 2016; Wu et al., 2018; Zhang et al., 2019).

Q. Can RL-oriented loss be better than NLL? RL is used for solving pitfalls of NLL loss. However, large-scale experiments in Wu et al. (2016) do not find promising performance improvements. Later on, Wu et al. (2018) study effective training tricks that can stably improve RL over NLL, but analyses on why RL cannot reach our expectation is still lacking. More recently, Choshen et al. (2020) deliver a novel understanding of the limitations of RL-based training. They find a peaking effect statistics to clarify the poor exploration problem of RL training due to the model distribution, which renders reward for being less critical. Following their work, Kiegeland and Kreutzer (2021) provide several counter-evidences in terms of claims that regard model distribution to be more critical than reward in Choshen et al. (2020). They revisit tricks like variance reduction, explore-exploitation trade-off and find that peakiness cannot solely explain improvements, and successful exploration can also improve the likelihood of low-ranked tokens.

Q. How does KD help with NAT? Knowledge Distillation at sequence-level (KD) (Kim and Rush, 2016) is another loss used to train a student NMT model from the output distribution or prediction of a teacher model. In Non-Autoregressive machine Translation (NAT), KD is a crucial training technique to bring the NAT model's performance close to autoregressive ones (Gu et al., 2017). Zhou et al. (2020b) investigate the critical role of KD in non-autoregressive NMT training. They find that KD reduces the complexity of the training bitext corpora to alleviate the learning/optimization burden of the NAT model due to its less powerful modeling power. They also propose improved KD loss functions for improved training. Xu et al. (2021) further analyze the impacts of KD training over the intrinsic characteristics of the NAT model. By defining two measures, namely word ordering agree and lexical diversity, they empirically demonstrate that KD is actually reducing training data complexity in terms of word ordering and lexical choices.

4.3 Training tricks

Since Transformer has already become the de-facto architecture for NMT best practice, several works attempt to dig deeper into those tricks for making Transformer training really work.

Q. How does LN help? As for the trick of Layer Normalization (LN), Wang et al. (2019b) calculate the instability of gradient mathematically when putting LN layer after residual block (post-LN) and empirically prove the effectiveness of pre-LN for scaling up Transformer with deeper layers. Then, Xiong et al. (2020) take advantage of the mean field theory to prove that post-LN connection at initialization leads to a large gradient. They find that the warming-up stage is avoiding such a problem.

Q. How residual blocks cause training instability? Besides the position of LN, Liu et al. (2020) provide comprehensive analyses of what complicates Transformer training theoretically and empirically. Their analyses find that the residual blocks can also lead to the unbalanced gradient issue.

Q. How does label smoothing help? As for the typical trick label smoothing, Müller et al. (2019) find that label smoothing can help calibrate training instances. Gao et al. (2020) investigate its theoretical and empirical role. Theoretically, they find what objective label smoothing is optimized for and derive an analytical solution for visualization for picking a better probability mass hyper-parameter for smoothing (e.g., from usual 0.1 to 0.3).

5 Understanding Inference

5.1 Prediction explanation

Q. How to attribute NMT model's prediction? One effective way to interpret the NMT model's behavior is to understand why the model predicts specific tokens step-wise regarding input tokens. At the beginning of NMT, attention is leveraged to visualize output-input correlation (Bahdanau et al., 2014). Then, Alvarez-Melis and Jaakkola (2017) propose a perturbation-based method to collect correlation pairs from relating every target token to every source token so that the explanation is model-agnostic. They exemplify with a case study that model debugging could be conducted based on such attention-like visualizations. Ding et al. (2017) leverage the so-called layer-wise relevance propagation (LRP) to capture the correlation of any two nodes in the computation graph of the model. They further use this method to visualize the relationship between prediction and input. Several translation errors are analyzed using LRP visualization to show the power of this method. Treviso and Martins (2020) proposes sparse/selective attention as a better way than gradient and erasure methods that relate prediction to input features (sequence) in terms of a success rate of a communication game (in Sec. G of the paper). Abnar and Zuidema (2020) propose a new method for visualizing the flow of the information from each input token to the output. Their proposed methods correlate well with attention and gradient-based method. More broadly speaking, various kinds of so-called *attribution* methods in Sec. B.1 can all be adapted to explain step-wise prediction of NMT. Vafa et al. (2021) propose a combinatorial optimization formulation for finding a subset of input that correlates most with a given output token. Their experiments show that the proposed method is most faithful among other explanation methods.

Q. How to properly evaluate prediction attribution of NMT model? As mentioned in the previous question, various prediction attribution methods can be used to explain model prediction. However, in practice which method to choose? There seems to be no fixed answer currently since there are already several issues found with attributions. Regardless of these issues, several works have proposed methods to evaluate attributions from different perspectives. Li et al. (2020) propose to use the word-to-word correlation rules extracted by various attribution methods to train models close to the original NMT model and uses the closeness as a way to evaluate the attribution results. Treviso and Martins (2020)'s communication game can also be used as an evaluation method. Beyond NMT, several works propose methods and benchmarks for evaluating attribution (Hao, 2020; Arras et al., 2019; Ismail et al., 2020; Ding and Koehn, 2021).

5.2 Decoding explanation

Q. Do larger beams lead to better results? After widely adoption of NMT, Koehn and Knowles (2017) describe a common phenomenon of beam search decoding across various language pairs, that is, by increasing beam size, the BLEU score will rise up a little and then jump down quite a lot, compared to SMT. This represents the so-called *length bias* problem which has been investigated to show its correlation with **i)** decoding scoring function (Huang et al., 2018; Yang et al., 2018) and **ii)** beam size. Murray and Chiang (2018) further find that label bias is one factor of such a problem and propose a simple heuristic to alleviate it. Cohen and Beck (2019) deliver a more detailed analysis on beam search using the concept of search discrepancies, which is computed through the difference between the maximum log probability and log probability of the ground-truth token at every time-step under force decoding. They find that a larger beam size may cause larger and more discrepancies at the beginning of decoding, degrading performance.

Q. Is beam search good enough? Stahlberg and Byrne (2019) analyze the impacts of model/search errors on performance, based on exact inference for vanilla beam search. They find the model error is more responsible. Meister et al. (2020) cleverly frame beam search as exact solution to a different decoding objective to gain insights into why high probability under a model alone may not guarantee adequacy. Eikema and Aziz (2020) attempt to clarify the problem of maximum a posterior (MAP) based beam search. They find that translation distributions of the model do reproduce various statistics of the training data, but beam search strays from such statistics. They also propose to use Minimum Bayes Risk (MBR) decoding instead. Müller and Sennrich (2021) study the properties of MBR decoding. They find that MBR decoding still exhibits length and token frequency bias due to the bias of evaluation metrics, but MBR also increases robustness against copy noise and domain shift.

6 Understanding Model Behavior

Model behavior understanding is generally an effort to *characterize* and *analyze* certain property of the *model's predictions* in terms of certain aspect or factor in concern. Currently, we can divide those research works into three main categories according to their adopted analysis methodologies: a). **static analysis**: that tries to *directly* analyze properties of the model's predictions, e.g., fluency, grammaticality, word choice, the degree of literalness or creativity, etc.; b). **controlled analysis**: that tries to characterize the model's reaction to inputs *constructed* with certain properties in concern, e.g., compositionality, specific linguistic phenomenon, etc.; c). **dynamic analysis**: that attempts to do *interventions* and *manipulation* to the inputs or the model, which might help reveal weaknesses of the model when making predictions about these inputs, e.g., adversarial examples, syntactic/semantic variants, hallucinations, noise in training data, etc.

6.1 Static analysis

Q. Is NMT model's prediction linguistically natural? Toral and Sánchez-Cartagena (2017) conduct a comparative study between the predictions of NMT and SMT models in terms of fluency, reordering, sentence length among 9 language directions. It highlights the power of NMT models on generating more fluent, accurately reordered predictions. Later on, Martindale et al. (2019) also study the fluency-adequacy dilemma of neural models. Wei et al. (2018) investigate the grammaticality of NMT outputs. They leverage the so-called English Resource Grammar as a reference for comparison. They find that over 93% of the model translations are parseable, suggesting that the model learns to generate conforming to grammar; however, rare syntactic rules are seldom learned.

Q. Can NMT model generate long-tailed translation? Long-tailed translations can be predictions that contains low-frequent tokens, complex phrases, and advanced sentence structures. Raunak et al. (2020) characterize the hardness of NMT to predict long-tailed words and tokens through token-level and sentence-level metrics. Agrawal and Carpuat (2019) study text complexity of predictions and focus on controlling the outputs towards less complexity. Vanmassenhove et al. (2021) give a detailed and sufficient analysis on the richness of word choices and synonyms etc.. They also design several metrics to evaluate linguistic complexity. Long-tailed translations can also be indirect translations of phrases that are seldom in the common bitext corpus. Zhai et al. (2020) investigate whether NMT models are capable of producing non-literal translations. They propose methods to detect those non-literal translation phenomena in bitext.

Q. Can model's prediction be well calibrated? Calibration is a sound property of a learned model to predict the probability of the true correctness likelihood (Guo et al., 2017). Kumar and Sarawagi (2019) analyze the sources of surprising miscalibration in NMT. They find that the severe miscalibration of the EOS token and the suppression of attention uncertainty are two main reasons. Wang et al. (2020) further study the fine-grained calibration of the model predictions. They characterize miscalibrated tokens with linguistic features, such as questions about how part-of-speech, frequency, word position, word granularity affect calibration.

6.2 Controlled analysis

Q. Can NMT model handle inputs with different types of linguistic phenomenon? Inputs to an NMT model can be linguistically sophisticated. Burchardt et al. (2017) manually construct a test suite with different kinds of linguistic phenomenon of the source input sentences, for instances, multiword expressions, verb tense/aspect/mood, named entity, and terminology in German⇔English translation tasks. They compare the performance of the Google NMT system at that time with SMT and rule-based models on this test suite. They find that neural models handle multiword expressions much better than rule-based and SMT models, while rule-based ones can handle verb tense/aspect/mood structure the best, and SMT handles named entities the best. Similarly, Isabelle et al. (2017) construct a challenge set with yes/no questions for analyzing both phrase and neural translation models' capability to handle three categories of linguistic phenomenon in English⇒French task. They find NMT models are much better at tackling subject-verb agreement and perform well on handling both lexico-syntactic and syntactic divergences. They also identify some weaknesses of neural models; please refer to Table 3 in that paper for details.

Q. Can NMT model handle inputs compositionally? Raunak et al. (2019) measure two distinct traits of compositionality - *productivity* and *systematicity* - of the NMT model by comparing performance before and after sentence concatenation. Their experiments quantitatively attribute the poor performance to the weakness of the encoder's representational power. Li et al. (2021) build a benchmark for training and testing the model's compositional capability to tackle *compounds*, which are constructed through pre-defined atoms and syntactic rules. Dankers et al. (2021) evaluate the model's compositionality through the lens of the model's local/global processing of the input. Voita et al. (2019a) focus on a problem of NMT model trained on sentence-level, that is, while the model can accurately translate sentences A and B, but can not when A and B are concatenated in a broader context, which can be also regarded as a problem of compositionality in discourse translation. All the above works find that Transformer or more or other NMT models have poor compositionality.

6.3 Dynamic analysis

Q. Is NMT model robust to inputs? Adversarial examples are an essential direction for testing the NMT model's robustness where the adversarial inputs are created through input manipulation. Belinkov and Bisk (2018) is the first to investigate how realistic, natural adversarial input (e.g. character-level keyboard typing errors) can break the char-based translation model. Zhao et al. (2017) and Cheng et al. (2020) investigate model-based methods for generating adversarial examples; while Ebrahimi et al. (2018) focus on attacking the char-based NMT models. Besides adversarials, He et al. (2019) use different input word replacement strategies to identify important source words that guarantee the translation quality of the source input. They argue that the so-called *importance words* are crucial to guarantee fertility and should not be ignored. Fadaee and Monz (2020) study the so-called *volatility* of NMT models where the input is semantically and syntactically transformed while the prediction can have unexpected disastraous changes. They find RNN and Transformer display volatile behavior in 26% and 19% of sentence variations.

Q. When or why does NMT model hallucinate? Hallucination is a recently identified phenomenon in Lee et al. (2018). It is the problem of an NMT model that outputs irrelevant sentence predictions or textual spans with respect to certain *constructed* input. They analyze the attention patterns that distinguish hallucinated and normal predictions. Raunak et al. (2021) connect this phenomenon to long-tailed memorization effect of the model. Wang and Sennrich (2020) regard exposure bias as one factor of hallucination and find domain-shift amplifies its harmfulness. Zhou et al. (2020a) tackle the identification problem of hallucination of neural sequence model in general. They construct datasets for token-wise annotation of hallucination and explore some basic methods for detecting hallucinated tokens.

7 Limitations, Future and Conclusion

In this part, we summarize several current limitations of those aforementioned understandings, interpretations and findings, and propose a few future directions on the understanding course of NMT.

- Vacuousness of representation probing: probing measures the feature generalization ability of the NMT learned representations on certain concerned linguistic task, however, does doing well on that task really help the model with the translation task? Such direct correlation between probing task and translation is very vague as well. Elazar et al. (2021) attempt to resolve this issue through explicit removing certain linguistic knowledge in the learned representation of BERT to see its utility on the downstream classification tasks. So, how about using such analysis in more complex translation tasks (Ravichander et al., 2021).

- Usability of prediction attribution: § 5's first question discusses many methods for attributing predicted tokens to previous input tokens. Besides the evaluation issue of these methods, how to use such attributions to debug model, moreover, to improve user trust beyond sole alignment or to improve interactive translation (Santy et al., 2019) is not well explored.

- Insufficient understanding on learning dynamics: by exploring learning dynamics, theorists have found critical learning phases that determine final generalization (Achille et al., 2017; Hu et al., 2020; Jastrzebski et al., 2021). However, investigations of learning dynamics are largely neglected in NMT, except for Saphra and Lopez (2019); Zhu et al. (2020); Voita et al. (2021). We think gaining more insights in the learning dynamics of NMT model might help with better curriculum, data selection, instance reweighting, noise-based learning, etc..

- Lack of data-centric understanding: many of the current understandings leverage a model-centric analysis, i.e., only considering architectural inductive bias without knowing characteristics of the training data, however, the ultimate model behavior is largely determined by the training instances as well (Yona et al., 2021). In NLP, there have been works that using dataset attribution techniques like influence function (Koh and Liang, 2017) to find artifacts in the training set for text classification (Han et al., 2020). Thus how to adopt similar methods to the complex machine translation task should be studied. We think this direction may help researchers curate more compact and continuously-updated datasets for sample-efficient training and continual learning (Cao et al., 2021) of NMT.

As a conclusion, the understanding of the evolving NMT framework should be always on its way and, to find limitations of the current best practice, emerging topics with multilingual, continual and discourse NMT (Dabre et al., 2020; Garcia et al., 2021; Yin et al., 2021) require better understanding, theory-oriented and empirical analyses as well, so the FAQs here (`https://nmtology.github.io/`) might and should be revisited and updated in new scenarios. The authors believe that knowing the historic understandings could help the community pave the way towards the future.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Alessandro Achille, Matteo Rovere, and Stefano Soatto. 2017. Critical learning periods in deep neural networks. *arXiv preprint arXiv:1711.08856*.

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. 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 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. 2018. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Brussels, Belgium. Association for Computational Linguistics.

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.

Bang An, Jie Lyu, Zhenyi Wang, Chunyuan Li, Changwei Hu, Fei Tan, Ruiyi Zhang, Yifan Hu, and Changyou Chen. 2020. Repulsive attention: Rethinking multi-head attention as Bayesian inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 236–255, Online. Association for Computational Linguistics.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. 2019. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy. Association for Computational Linguistics.

Joris Baan, Jana Leible, Mitja Nikolaus, David Rau, Dennis Ulmer, Tim Baumgärtner, Dieuwke Hupkes, and Elia Bruni. 2019. On the realization of compositionality in neural networks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 127–137, Florence, Italy. Association for Computational Linguistics.

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*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, et al. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

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

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020a. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020b. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv preprint arXiv:1506.03099*.

Arianna Bisazza and Clara Tump. 2018. The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2871–2876, Brussels, Belgium. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7236–7249, Online. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Peter Jan-Thorsten, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural mt engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1):159.

Yue Cao, Hao-Ran Wei, Boxing Chen, and Xiaojun Wan. 2021. Continual learning for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3964–3974, Online. Association for Computational Linguistics.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

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

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3601–3608. AAAI Press.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pages 263–270.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. 2020. On the weaknesses of reinforcement learning for neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1290–1299. PMLR.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James R. Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6309–6317. AAAI Press.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2021. The paradox of the compositionality of natural language: a neural machine translation case study. *arXiv preprint arXiv:2108.05885*.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. *arXiv preprint arXiv:2104.05824*.

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

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, Vancouver, Canada. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018a. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018b. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364, New Orleans, Louisiana. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Marzieh Fadaee and Christof Monz. 2020. The unreasonable volatility of neural machine translation models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 88–96, Online. Association for Computational Linguistics.

Yingbo Gao, Weiyue Wang, Christian Herold, Zijian Yang, and Hermann Ney. 2020. Towards a better understanding of label smoothing in neural machine translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 212–223, Suzhou, China. Association for Computational Linguistics.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, Florence, Italy. Association for Computational Linguistics.

Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz

Górecki. 2019. GEval: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy. Association for Computational Linguistics.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.

Yiding Hao. 2020. Evaluating attribution methods using white-box LSTMs. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 300–313, Online. Association for Computational Linguistics.

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 953–962, Hong Kong, China. Association for Computational Linguistics.

Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. 2020. The surprising simplicity of the early-time learning dynamics of neural networks. *arXiv preprint arXiv:2006.14599*.

Liang Huang, Kai Zhao, and Mingbo Ma. 2018. When to finish? optimal beam search for neural text generation (modulo beam size). *arXiv preprint arXiv:1809.00069*.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. 2020. Benchmarking deep learning interpretability in time series predictions. In *Advances in Neural Information Processing Systems*, volume 33, pages 6441–6452. Curran Associates, Inc.

Sarthak Jain and Byron C. Wallace. 2019. 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. Association for Computational Linguistics.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. 2021. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pages 4772–4784. PMLR.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A Smith. 2020. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. *arXiv preprint arXiv:2006.10369*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Samuel Kiegeland and Julia Kreutzer. 2021. Revisiting the weaknesses of reinforcement learning for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1673–1681, Online. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. *OpenReview.net*.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *arXiv preprint arXiv:2104.15135*.

Guanlin Li, Lemao Liu, Xintong Li, Conghui Zhu, Tiejun Zhao, and Shuming Shi. 2019a. Understanding and Improving Hidden Representations for Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 466–477, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019b. 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.

Xintong Li, Lemao Liu, Zhaopeng Tu, Shuming Shi, and Max Meng. 2018. Target foresight based attention for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1380–1390, New Orleans, Louisiana. Association for Computational Linguistics.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4767–4780, Online. 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.

Liyuan Liu, Jialu Liu, and Jiawei Han. 2021. Multi-head or single-head? an empirical comparison for transformer training. *arXiv preprint arXiv:2106.09650*.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. 2020. Understanding the difficulty of training transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5747–5763, Online. Association for Computational Linguistics.

Siwen Luo, Hamish Ivison, Caren Han, and Josiah Poon. 2021. Local interpretations for explainable natural language processing: A survey. *arXiv preprint arXiv:2103.11072*.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 233–243, Dublin, Ireland. European Association for Machine Translation.

Rebecca Marvin and Philipp Koehn. 2018. Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 125–131, Boston, MA. Association for Machine Translation in the Americas.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. 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.

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

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2020. Training with adversaries to improve faithfulness of attention in neural machine translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 93–100, Suzhou, China. Association for Computational Linguistics.

Rajiv Movva and Jason Zhao. 2020. Dissecting lottery ticket transformers: Structural and behavioral study of sparse neural machine translation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 193–203, Online. Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum bayes risk decoding in neural machine translation. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *NeurIPS*.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. Fixed encoder self-attention patterns in transformer-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 556–568, Online. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. 2020. On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3088–3095, Online. Association for Computational Linguistics.

Vikas Raunak, V. Kumar, Florian Metze, and Jaimie Callan. 2019. On compositionality in neural machine translation. *ArXiv*, abs/1911.01497.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. *arXiv preprint arXiv:2104.06683*.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Hassan Sajjad, Narine Kokhlikyan, Fahim Dalvi, and Nadir Durrani. 2021. Fine-grained interpretation and causation analysis in deep NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 5–10, Online. Association for Computational Linguistics.

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive neural machine translation prediction. 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): System Demonstrations*, pages 103–108, Hong Kong, China. Association for Computational Linguistics.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. 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 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Jiajun Shen, Peng-Jen Chen, Matthew Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2021. The source-target domain mismatch problem in machine translation. In

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1519–1533, Online. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

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

Charlie Snell, Ruiqi Zhong, Dan Klein, and Jacob Steinhardt. 2021. Approximating how single head attention learns. *arXiv preprint arXiv:2103.07601*.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? 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 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2018. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363.

Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Qiu Han, Guoyin Wang, Eduard Hovy, and Jiwei Li. 2021. Interpreting deep learning models in natural language processing: A review. *arXiv preprint arXiv:2110.10470*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Madhumita Sushil, Simon Šuster, and Walter Daelemans. 2018. Rule induction for global explanation of trained models. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 82–97, Brussels, Belgium. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4263–4272, Brussels, Belgium. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019a. Encoders help you disambiguate word senses in neural machine translation. 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 1429–1435, Hong Kong, China. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019b. Understanding neural machine translation by simplification: The case of encoder-free models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1186–1193, Varna, Bulgaria. INCOMA Ltd.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Marcos Treviso and André F. T. Martins. 2020. The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 107–118, Online. Association for Computational Linguistics.

Keyon Vafa, Yuntian Deng, David M Blei, and Alexander M Rush. 2021. Rationales for sequential predictions. *arXiv preprint arXiv:2109.06387*.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT. *CoRR*, abs/2109.01396.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019a. VizSeq: a visual analysis toolkit for text generation tasks. 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): System Demonstrations*, pages 253–258, Hong Kong, China. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019b. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Johnny Wei, Khiem Pham, Brendan O'Connor, and Brian Dillon. 2018. Evaluating grammaticality in seq2seq models with a broad coverage HPSG grammar: A case study on machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 298–305, Brussels, Belgium. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

Y. Wu, M. Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Q. Gao, Klaus Macherey, J. Klingner, Apurva Shah, M. Johnson, X. Liu, Lukasz Kaiser, Stephan Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, George Kurian, Nishant Patil, W. Wang, C. Young, J. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. Corrado, Macduff Hughes, and J. Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR.

Weijia Xu, Shuming Ma, Dongdong Zhang, and Marine Carpuat. 2021. How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4392–4400, Online. Association for Computational Linguistics.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3635–3644, Florence, Italy. Association for Computational Linguistics.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. *arXiv preprint arXiv:1808.09582*.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801, Online. Association for Computational Linguistics.

Gal Yona, Amirata Ghorbani, and James Zou. 2021. Who's responsible? jointly quantifying the contribution of the learning algorithm and data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 1034–1041, New York, NY, USA. Association for Computing Machinery.

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

Yuming Zhai, Gabriel Illouz, and Anne Vilnat. 2020. Detecting non-literal translations by fine-tuning cross-lingual pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5944–5956, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. *arXiv preprint arXiv:1906.02448*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples. *CoRR*, abs/1710.11342.

Chunting Zhou, Jiatao Gu, Mona T. Diab, P. Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020a. Detecting hallucinated content in conditional neural sequence generation. *ArXiv*, abs/2011.02593.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020b. Understanding knowledge distillation in non-autoregressive machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Conghui Zhu, Guanlin Li, Lemao Liu, Tiejun Zhao, and Shuming Shi. 2020. Understanding learning dynamics for neural machine translation.

A Mindmap of FAQs

Fig. 1 demonstrates a screenshot of the mindmap on our website (**`https://nmtology.github.io/`**). Visitors can zoom in or zoom out the tree by the clicking inner nodes. And by clicking a specific question, you will be guided to a separate webpage that hosts the related papers under that question.

B Methodology

This section gives a focused introduction to several commonly used methodologies for understanding the NMT framework, which are commonly used in our surveyed papers. Please refer to Belinkov et al. (2020b); Belinkov and Glass (2019) for a general introduction to interpretation methodologies.

B.1 Attribution

Attribution is one of the local explanation methodologies for understanding and visualizing the decision of predictive models, i.e., classifiers (Carvalho et al., 2019). It relates every model prediction to a subset of input features that might be the cause of that prediction. A large number of attribution methods are proposed recently in vision and learning community (Simonyan et al., 2013; Bach et al., 2015; Montavon et al., 2017; Selvaraju et al., 2017; Sundararajan et al., 2017). In NMT, the prediction $\hat{y}$ could be seen as a sequence of classification steps. Given input x, predicted sequence $\hat{y}$, and the NMT model $\mathcal{M}$, an attribution method is defined as an algorithmic process $\mathcal{A}(x, \hat{y}, \mathcal{M})$, it outputs *relevant scores* over every token of x and $y_{<t}$ for each $\hat{y}_t$. Based on relevant scores, we can at least qualitatively know what $\hat{y}_t$ is probably aligned to.

Model-specific Methods Model-specific attributions can have access to the inner computation of the NMT model. Gradient-based attribution uses the activation of $\hat{y}_t$ for backward computation. It then uses the gradients on each embedding vector of every token in x and $\hat{y}_{<t}$ to compute its relevant score regarding $\hat{y}_t$ (Ding et al., 2019). Layer-wise Relevance Propagation (LRP) uses activation vectors and model weights to compute neuron relevance, and then back-propagates the relevance back to the input tokens (Ding et al., 2017).

Model-agnostic Methods Model-agnostic attributions regard the NMT model as a black-box. These methods usually calculate the relevant scores through manipulating model inputs (Alvarez-Melis and Jaakkola, 2017; Li et al., 2019b). For example, prediction difference (Li et al., 2019b) chooses a particular feature (token $x_{t'}$), and observe the probability difference of $\hat{y}_t$ before and after removing that feature, the larger the probability is, the more relevant between $\hat{y}_t$ and the removed one.

B.2 Probing

Probing is a method for investigating how much a component of the NMT model captures certain kind of knowledge. The main technique for probing is to train a classifier g which maps an intermediate representation $f(x)$ of the input x to certain property of interest z (Alain and Bengio, 2016;

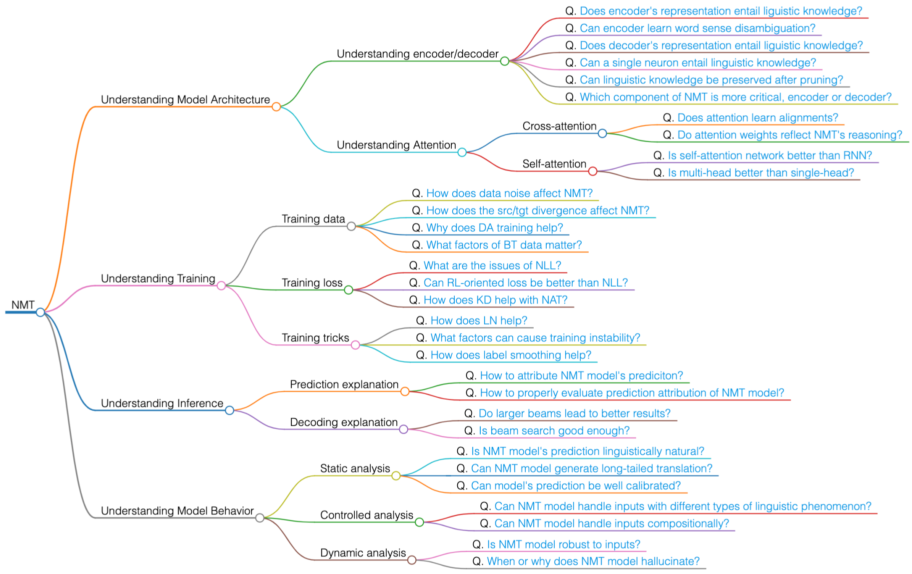

Figure 1: A screenshot of the mindmap of FAQs on our website **`https://nmtology.github.io/`**.

Conneau et al., 2018). This network component $f(\cdot)$ can be word embedding, sentence embedding, hidden state, attention weight, etc. The property z can be various linguistic features, such as part-of-speech tags, morphological information, or more complicated syntactic or semantic features. Then, the accuracy of $g(f(x))$ can reveal the quality of representations $f(x)$ with respect to the property z, so that different model components can be compared to each other. To show this accuracy is non-trivial, it can be compared to feeding random inputs into the classifier $g(\cdot)$. Meanwhile, comparing with state-of-the-art on that task can inform us how much is missing from the representation.

B.3 Others

In addition to attribution and probing which are most commonly used, several other methodologies are used in specific analysis scenarios (Belinkov and Glass, 2019). *i) Visualization* is always used accompanied with attribution to show the relationship between predicted and input tokens; it is also used to visualize clustering effects of learned representations through dimension reduction techniques (Alvarez-Melis and Jaakkola, 2017; Ding et al., 2017). *ii) Challenge set* is always used to investigate certain desirable characteristic of the model through data test suite construction (Isabelle et al., 2017). *i) Model extraction* is to extract use knowledge distillation to learn a transparent or interpretable surrogate NMT model (e.g. rules, syntactic trees) from the original one (Bastani et al., 2017; Sushil et al., 2018). Besides, several works also build toolkits for visualization and model debugging (Strobelt et al., 2018; Wang et al., 2019a; Graliński et al., 2019; Gauthier et al., 2020).