On the Inference Calibration of Neural Machine Translation

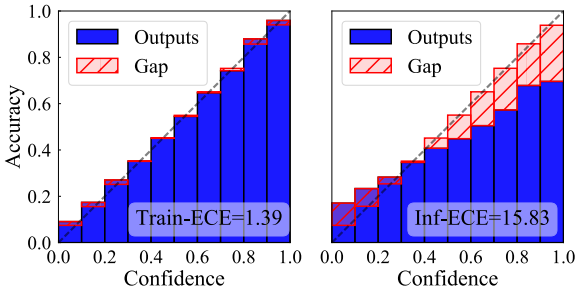
Shuo Wang[⋆∗] **Zhaopeng Tu**[⊤] **Shuming Shi**[⊤] **Yang Liu**[⋆†]
[⋆]Institute for Artificial Intelligence
Department of Computer Science and Technology, Tsinghua University
Beijing National Research Center for Information Science and Technology
[⊤]Tencent AI Lab
[†]Beijing Academy of Artificial Intelligence
Beijing Advanced Innovation Center for Language Resources
[⋆]{wangshuo.thu, liuyang.china}@gmail.com
[⊤]{zptu, shumingshi}@tencent.com

Abstract

Confidence calibration, which aims to make model predictions equal to the true correctness measures, is important for neural machine translation (NMT) because it is able to offer useful indicators of translation errors in the generated output. While prior studies have shown that NMT models trained with label smoothing are well-calibrated on the ground-truth training data, we find that miscalibration still remains a severe challenge for NMT during inference due to the discrepancy between training and inference. By carefully designing experiments on three language pairs, our work provides in-depth analyses of the correlation between calibration and translation performance as well as linguistic properties of miscalibration and reports a number of interesting findings that might help humans better analyze, understand and improve NMT models. Based on these observations, we further propose a new graduated label smoothing method that can improve both inference calibration and translation performance. [1]

1 Introduction

Calibration requires that the probability a model assigns to a prediction (i.e., *confidence*) equals to the correctness measure of the prediction (i.e., *accuracy*). Calibrated models are important in user-facing applications such as natural language processing (Nguyen and O'Connor, 2015) and speech recognition (Yu et al., 2011), in which one needs to assess the confidence of a prediction. For example, in computer-assisted translation, a calibrated machine translation model is able to tell a user when the model's predictions are likely to be incorrect, which is helpful for the user to correct errors.



(a) Training (b) Inference

Figure 1: Reliability diagrams in training and inference for the WMT14 En-De task. "Gap" denotes the difference between confidence and accuracy. Smaller gaps denotes better calibrated outputs. We find that the average gaps between confidence and accuracy are much larger in inference than in training (i.e., 15.83 > 1.39).

The study of calibration on classification tasks has a long history, from statistical machine learning (Platt et al., 1999; Niculescu-Mizil and Caruana, 2005) to deep learning (Guo et al., 2017). However, calibration on structured generation tasks such as neural machine translation (NMT) has not been well studied. Recently, Müller et al. (2019) and Kumar and Sarawagi (2019) studied the calibration of NMT in the training setting, and found that NMT trained with label smoothing (Szegedy et al., 2016) is well-calibrated. We believe that this setting would cover up a central problem of NMT, *the exposure bias* (Ranzato et al., 2015) – the training-inference discrepancy caused by teacher forcing in the training of auto-regressive models.

In response to this problem, this work focuses on the calibration of NMT in inference, which can better reflect the generative capacity of NMT models. To this end, we use translation error rate (TER) (Snover et al., 2006) to automatically annotate the correctness of generated tokens, which makes it feasible to evaluate calibration in infer-

[∗]Work was done when Shuo Wang was interning at Tencent AI Lab under the Rhino-Bird Elite Training Program.

[1]The source code is available at https://github.com/shuo-git/InfECE.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3070–3079
July 5 - 10, 2020.

ence. Experimental results on several datasets across language pairs show that even trained with label smoothing, NMT models still suffer from miscalibration errors in inference. Figure 1 shows an example. While modern neural networks on classification tasks have been found to be miscalibrated in the direction of *over-estimation* (i.e., confidence $>$ accuracy) (Guo et al., 2017), NMT models are also *under-estimated* (i.e., confidence $<$ accuracy) on low-confidence predictions. In addition, we found that miscalibrated predictions correlate well with the translation errors in inference. Specifically, the over-estimated predictions correlate more with over-translation and mis-translation errors, while the under-estimated predictions correlate more with under-translation errors. This demonstrates the necessity of studying inference calibration for NMT.

By investigating the linguistic properties of miscalibrated tokens in NMT outputs, we have several interesting findings:

- *Frequency*: Low-frequency tokens generally suffer from under-estimation. Moreover, low-frequency tokens contribute more to over-estimation than high-frequency tokens, especially on large-scale data.
- *Position*: Over-estimation does not have a bias on the position of generated tokens, while under-estimation occurs more in the left part of a generated sentence than in the right part.
- *Fertility*: Predicted tokens that align to more than one source token ("fertility$\geq$2") suffer more from under-estimation, while tokens with fertility $<$ 1 suffer from over-estimation.
- *Syntactic Roles*: Content tokens are more likely to suffer from miscalibration than content-free tokens. Specifically, verbs are more likely to suffer from over-estimation than under-estimation.
- *Word Granularity*: sub-words suffer more from both over-estimation and under-estimation, while full words are less likely to be miscalibrated.

Inspired by the finding that miscalibration on classification tasks is closely related to lack of regularization and increased model size (Guo et al., 2017), we revisit these techniques on the NMT (i.e., structured generation) task:

- *Regularization Techniques*: We investigate label smoothing and dropout (Hinton et al., 2012), which directly affect the confidence estimation. Both label smoothing and dropout improve the inference calibration by alleviating the over-estimation. Label smoothing is the key for well-calibration, which is essential for maintaining translation performance for inference in large search space. Inspired by this finding, we propose a novel *graduated label smoothing* approach, in which the smoothing penalty for high-confidence predictions is higher than that for low-confidence predictions. The graduated label smoothing can improve translation performance by alleviating inference miscalibration.
- *Model Size*: Increasing model size consistently improves translation performance at the cost of negatively affecting inference calibration. The problem can be alleviated by increasing the capacity of encoder only, which maintains the inference calibration and obtains a further improvement of translation performance in large search space.

To summarize, the main contributions of our work are listed as follows:

- We demonstrate the necessity of studying inference calibration for NMT, which can serve as useful indicators of translation errors.
- We reveal certain linguistic properties of miscalibrated predictions in NMT, which provides potentially useful information for the design of training procedures.
- We revisit recent advances in architectures and regularization techniques, and provide variants that can boost translation performance by improving inference calibration.

2 Related Work

Calibration on Classification Calibration on classification tasks has been studied for a long history in the statistics literature, including Platt scaling (Platt et al., 1999), isotonic regression (Niculescu-Mizil and Caruana, 2005) and many other methods for non-binary classification (Zadrozny and Elkan, 2002; Menon et al., 2012; Zhong and Kwok, 2013). For modern deep neural networks, Guo et al. (2017) demonstrated

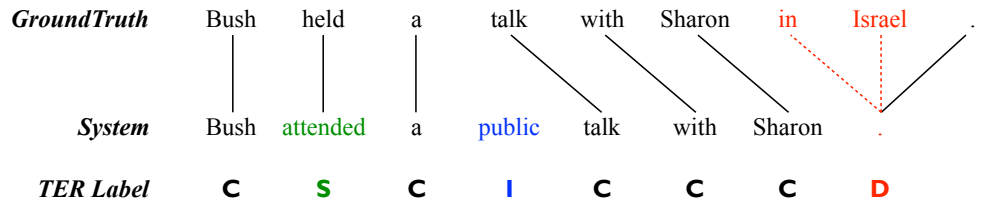


Figure 2: An example of TER labels. "C": correct, "S": substitution, corresponding to mis-translation, "I": insertion, corresponding to over-translation, "D": deletion, corresponding to under-translation. Dash line denotes mapping the label "D" from the ground-truth sequence to the generated sequence.

that recent advances in training and model architecture have strong effects on the calibration. Szegedy et al. (2016) propose the label smoothing technique which can effectively reduce the calibration error. Ding et al. (2019) extend label smoothing to adaptive label regularization.

Calibration on Structured Prediction Different from classification tasks, most natural language processing (NLP) tasks deal with complex structures (Kuleshov and Liang, 2015). Nguyen and O'Connor (2015) verified the finding of Niculescu-Mizil and Caruana (2005) in NLP tasks on log-linear structured models. For NMT, some works directed their attention to the uncertainty in prediction (Ott et al., 2018; Wang et al., 2019), Kumar and Sarawagi (2019) studied the calibration of several NMT models and found that the end of a sentence is severely miscalibrated. Müller et al. (2019) investigated the effect of label smoothing, finding that NMT models are well-calibrated in training. Different from previous works, we are interested in the calibration of NMT models in inference, given that the training and inference are discrepant for standard NMT models (Vaswani et al., 2017).

3 Definitions of Calibration

3.1 Neural Machine Translation

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

$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \prod_{t=1}^{T} P(y_t|\mathbf{y}_{<t}, \mathbf{x};\boldsymbol{\theta}), \quad (1)$$

where $\boldsymbol{\theta}$ is a set of model parameters and $\mathbf{y}_{<t}$ is a partial translation. At each time step, the model generates an output token of the highest probability based on the source sentence $\mathbf{x}$ and the partial translation $\mathbf{y}_{<t}$. The training objective is to minimize the negative log-likelihood loss on the training corpus.

Inference NMT models are trained on the ground-truth data distribution (*teaching forcing*), while in inference the models generate target tokens based on previous model predictions, which can be erroneous. The training-inference discrepancy caused by teacher forcing in maximum likelihood estimation training (Equation 1) is often referred to as *exposure bias* (Ranzato et al., 2015). In this work, we aim to investigate the calibration of NMT in inference, which we believe can better reflect the generation capacity of NMT models.

3.2 Calibration of NMT

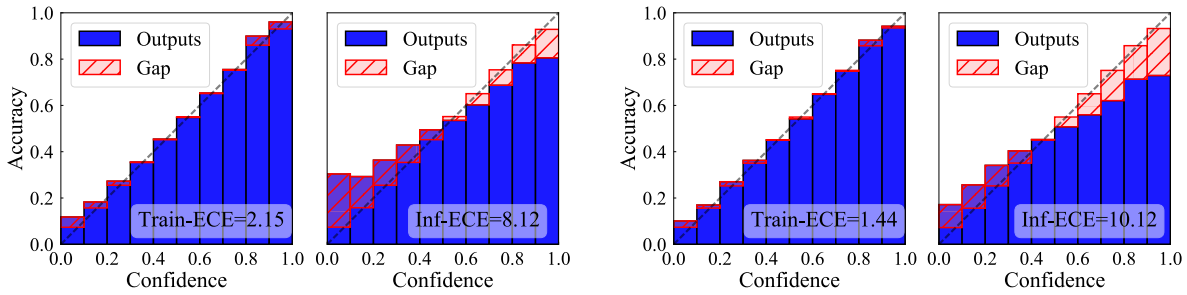
Calibration requires that the probability a model assigns to a prediction (i.e., *confidence*) equals to the true correctness measure of the prediction (i.e., *accuracy*). Modern neural networks have been found to be miscalibrated in the direction of over-estimation (Guo et al., 2017). In this study, we revisit the calibration problem in NMT. If an NMT model is well-calibrated, the gap between the confidence of the generated tokens and the accuracy of them will be small. [2]

Expected Calibration Error (ECE) ECE is a commonly-used metric to evaluate the miscalibration, which measures the difference in expectation between confidence and accuracy (Naeini et al., 2015). Specifically, ECE partitions predictions into M bins $\{B_1, \ldots, B_M\}$ according to their confidence and takes a weighted average of the bin's accuracy/confidence difference:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \Big| acc(B_m) - conf(B_m) \Big|, \quad (2)$$

where N is the number of prediction samples and $|B_m|$ is the number of samples in the m-th bin.

[2] For example, given 100 predictions, each with confidence 0.7. If the accuracy is also 0.7 (i.e., 70 of the 100 tokens are correct), then the NMT model is well calibrated.



(a) En-Jp (b) Zh-En

Figure 3: Reliability diagrams on (a) En-Jp and (b) Zh-En datasets. Left: training, right: inference.

ECE in Training and Inference In the case of considering just the topmost token in structured prediction tasks (e.g., machine translation), the prediction is $\hat{y} = \arg\max_{y \in \mathcal{V}} P(y)$ with $P(\hat{y})$ as *confidence*. The *accuracy* $C(\hat{y}) \in \{1, 0\}$ denotes whether the prediction $\hat{y}$ is correct.

In training, the correctness of the prediction $\hat{y}$ is calculated as whether $\hat{y}$ matches the ground-truth token y_n: $C(\hat{y}) \in \{1, 0\}$. However, in inference it is not straightforward to measure the accuracy of $\hat{y}$, since it requires to build an alignment between the generated tokens and the ground-truth tokens.

To this end, we turn to the metric of Translation Error Rate (TER) (Snover et al., 2006), which measures the number of edits required to change a model output into the ground-truth sequence. Specifically, it assigns a label $l \in \{\text{C, S, I}\}$ to each generated token. Figure 2 shows an example of TER labels of each generated token with respect to the reference. As a side product, TER annotations provide the information of translation errors. While TER only labels the mis-translation ("S") and over-translation ("I") errors, we describe a simple heuristic method to annotate the under-translation error by mapping the label "D" from the ground-truth sequence to the generated sequence.

4 Miscalibration in NMT

Data and Setup We carried out experiments on three different language pairs, including WAT17 English-Japanese (En-Jp), WMT14 English-German (En-De), and WMT17 Chinese-English (Zh-En). The training datasets consist of 1.9M, 4.5M, and 20.6M sentence pairs respectively. We employed Byte pair encoding (BPE) (Sennrich et al., 2016) with 32K merge operations for all the three language pairs. We used BLEU (Papineni et al., 2001) to evaluate the NMT models. We used the TER toolkit (Snover et al., 2006) to label whether the tokens in NMT outputs are correctly translated. Normalization was not used, and the maximum shift distance was set to 50.

The NMT model that we used in our experiments is Transformer (Vaswani et al., 2017). We used base model as default, which consists of a 6-layer encoder and a 6-layer decoder and the hidden size is 512. The model parameters are optimized by Adam (Kingma and Ba, 2015), with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. We used the same warm-up strategy for learning rate as Vaswani et al. (2017) with $\text{warmup\_steps} = 4,000$.

4.1 Observing Miscalibration

Reliability diagrams are a visual representation of model calibration, which plot accuracy as a function of confidence (Niculescu-Mizil and Caruana, 2005). Specifically, it partitions the output tokens into several bins according to their prediction confidence, and calculate the average confidence and accuracy of each bin. Figure 1 shows the reliability diagrams of both training and inference on En-De and Figure 3 shows those on En-Jp and Zh-En. Results are reported on the validation sets.

NMT still suffers from miscalibration. The difference between training and inference ECEs is that when estimating training ECE, NMT models are fed with ground-truth prefixes (Kumar and Sarawagi, 2019; Müller et al., 2019), while for inference ECE, NMT models are fed with previous model predictions. As seen, the training ECE is very small, indicating that NMT models are well-calibrated in training. This is consistent with the findings of Kumar and Sarawagi (2019); Müller et al. (2019). However, the inference ECE is much higher, suggesting that NMT models still suffer from miscalibration in inference.

Translation		Well-Cali.	Mis-Cali.
Correct	*En-Jp*	**0.53**	0.47
	En-De	**0.57**	0.43
	Zh-En	**0.60**	0.40
	All	**0.57**	0.43
Error	*En-Jp*	0.46	**0.54**
	En-De	0.43	**0.57**
	Zh-En	0.36	**0.63**
	All	0.42	**0.58**

Table 1: Cosine similarity between the calibration and the translation errors on the held-out data.

Type		Under-Est.	Over-Est.
Under-Tra.	En-Jp	**0.35**	0.22
	En-De	**0.28**	0.24
	Zh-En	0.31	0.31
	All	**0.32**	0.26
Over-Tra.	En-Jp	0.28	**0.32**
	En-De	0.20	**0.36**
	Zh-En	0.29	**0.35**
	All	0.26	**0.34**
Mis-Tra.	En-Jp	0.24	**0.36**
	En-De	0.17	**0.42**
	Zh-En	0.24	**0.40**
	All	0.21	**0.39**

Table 2: Cosine similarity between the miscalibration errors (under-estimation and over-estimation) and the translation errors (under-translation, mis-translation, and over-translation) on the held-out data.

NMT models are miscalibrated in directions of both over- and under-estimation. Modern neural networks have been found to be miscalibrated on classification tasks in the direction of over-estimation (Guo et al., 2017). In contrast, NMT models also suffer from under-estimation problems. The under-estimation problem is more serious on En-Jp than on Zh-En, which we attribute to the smaller size of the training data of the En-Jp task.

4.2 Correlation with Translation Errors

We investigated the calibration error of tokens with different TER labels. As the development set is small, to make the results more convincing, we sampled 100K sentences from the training set as a held-out set and retrained the NMT model on the remained training set excluding the held-out set. All results in this section is reported by the retrained model. We firstly compute the gap between the confidence and the accuracy of each token in each confidence bin on the held-out set. Tokens in bins whose gaps are less than a threshold are labeled as well-calibrated, otherwise they are labeled as miscalibrated. We use the inference ECE estimated on the development set as the threshold for each language pair respectively. Miscalibrated tokens can be divided into two categories: over-estimation and under-estimation.

As shown in Table 1, correct translations (i.e., "C") have higher correlations to well-calibrated predictions and erroneous translations (i.e., "S", "I", and "D") correlate more to miscalibrated predictions. This finding is more obvious when NMT models are trained on larger data (e.g., Zh-En).

Table 2 lists the correlation between different translation errors and different kinds of miscalibration. We find that over-estimated predictions are closely correlated with over-translation and mis-translation errors, while the under-estimated predictions correlate well with under-translation errors. This finding demonstrates the necessity of studying inference calibration for NMT.

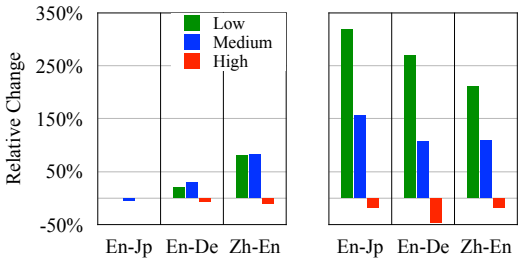
5 Linguistic Properties of Miscalibration

In this section, we investigate the linguistic properties of miscalibrated tokens in NMT outputs. We explore the following five types of properties: frequency, position, fertility, syntactic roles, and word granularity.

Frequency is generally related to miscalibration; position, fertility, and word granularity are three factors associated with structured prediction; syntactic roles or linguistic roles may vary across language pairs. The results in this section are reported on the held-out set by the retrained model.

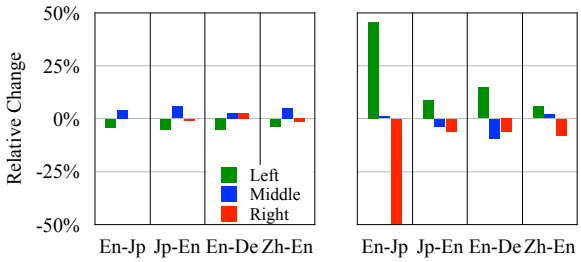
Relative Change We use the relative change of the proportion of a certain category of tokens to quantify to what extent they suffer from the under/over-estimation. For instance, in the Zh-En task, high-frequency tokens account for 87.6% on the whole held-out set, and among over-estimated tokens, high-frequency tokens account for 77.3%, thus for over-estimation the relative change of high-frequency tokens is (77.3-87.6)/87.6=-11.76% in Zh-En. Accordingly, the value of the red rectangle of Zh-En is -11.76% in Figure 4a.

Positive relative change denotes that a certain type of linguistic property accounts more in miscalibrated predictions than in all the predictions, suggesting this type of linguistic property suffers



(a) Over-Estimation (b) Under-Estimation

Figure 4: Effect of frequency on miscalibration.



(a) Over-Estimation (b) Under-Estimation

Figure 5: Effect of relative position on miscalibration.

from the miscalibration problem. Similarly, negative relative change suggests that a certainty type of linguistic property is less likely to be impaired by the miscalibration problem.

5.1 Frequency

We divide tokens into three categories based on their frequency, including *High*: the most 3,000 frequent tokens; *Medium*: the most 3,001-12,000 frequent tokens; *Low*: the other tokens.

Low-frequency tokens are miscalibrated in the direction of under-estimation. As shown in Figure 4, the relative changes of low- and medium-frequency tokens are much bigger than those of high-frequency tokens. The under-estimation in low- and medium-frequency tokens can be alleviated by increasing the size of training data (Figure 4b, data size: En-Jp $<$ En-De $<$ Zh-En).

Low-frequency tokens contribute more to over-estimation. As shown in Figure 4a, the relative changes of low- and medium-frequency tokens are positive while those of high-frequency tokens are negative, regarding over-estimation.

High-frequency tokens are less likely to be miscalibrated. We find the relative changes of high frequency tokens are negative across the three language pairs. The imbalance in token frequency plays an important role in the calibration of NMT.

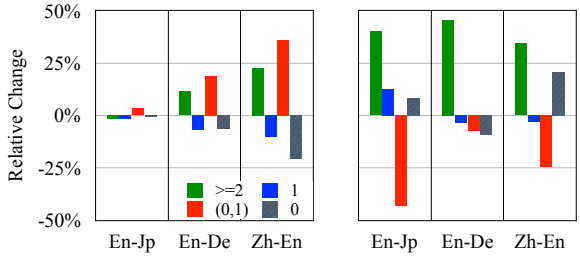
5.2 Position

In structured prediction, different positions may behave differently regarding miscalibration. Thus we divide all the tokens equally into three categories: *Left*: tokens on the left third; *Middle*: tokens on the middle third; *Right*: tokens on the right third. Figure 5 depicts the relative changes of these three positions. Since Japanese is a head-final language (Wu et al., 2018), we also include the results of Japanese-English ("Jp-En") for comparison.

Over-estimation does not have a bias on position. And this holds for both left-branching and right-branching languages. Increasing the size of training data is less likely to affect the over-estimation in different positions.

Under-estimation occurs more in the left part. This phenomenon is more obvious in left-branching languages (e.g., Japanese) than in right-branching languages (e.g., English and German), confirming that characteristics of a language play an important role in machine translation (Wu et al., 2018).

5.3 Fertility

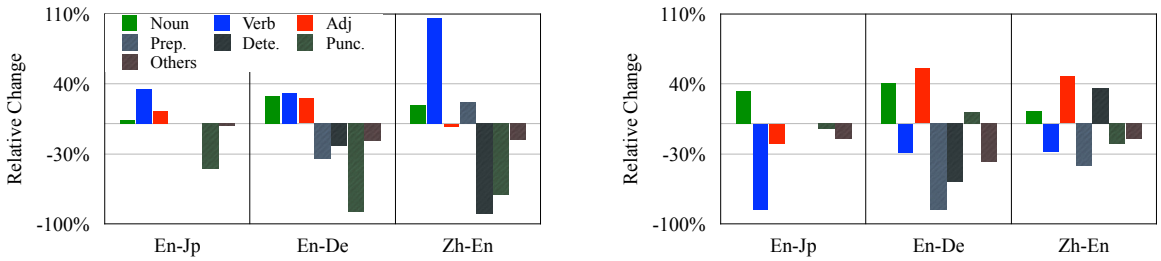


(a) Over-Estimation (b) Under-Estimation

Figure 6: Effect of fertility on miscalibration.

Fertility indicates how many source tokens a target token is aligned to, which is highly related to inference in NMT. We use `Fast Align` (Dyer et al., 2013) to extract bilingual alignment. We distinguish between four categories regarding fertility: "≥ 2": target tokens that are aligned to more than one source tokens; "1": target tokens that are aligned to a single source token; "$(0, 1)$": target tokens that are aligned to a single source token along with other target tokens; "0": target tokens that are not aligned to any source token. Figure 6 plots the results.

Tokens aligning to less than one source token suffer from over-estimation. The extent grows with



(a) Over-Estimation (b) Under-Estimation

Figure 7: Effect of POS tags on miscalibration.

the data size. In addition, these tokens ("(0, 1)") are less likely to suffer from under-estimation.

Tokens aligning to more than one source token suffer more from under-estimation. The relative change of fertility>=2 is much larger than that of the other types of fertility. Meanwhile, the null-aligned target tokens (fertility=0) also suffer from under-estimation problem instead of over-estimation problem on the large-scale Zh-En data.

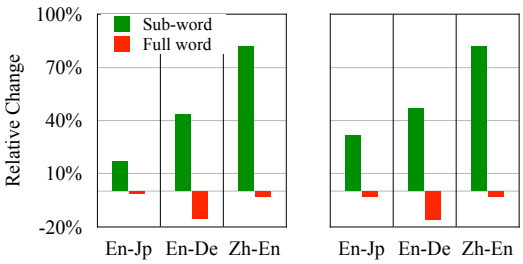
5.4 Syntactic Roles

In this experiment, we investigate the syntactic roles of miscalibrated tokens. [3] Words in English and German sentences are labeled by `Stanford POS tagger`[4], and Japanese sentences are labeled by `Kytea`[5]. We distinguish between the following POS tags: noun, verb, adjective, preposition, determiner, punctuation, and the others. Noun, verb, and adjective belong to content tokens. Preposition, determiner, punctuation and the others belong to content-free tokens.

Content tokens are more likely to suffer from miscalibration. From Figure 7 we find that the most relative changes of content tokens (i.e., "Noun", "Verb" and "Adj") are positive, while most of the relative changes of the content-free tokens (i.e., "Prep.", "Dete.", "Punc.", "Others") are negative. Among content tokens, the verbs ("Verb") face the over-estimation problem instead of the under-estimation problem. Surprisingly, the adjectives ("Adj") suffer from under-estimation problem on large data (e.g., En-De and Zh-En).

5.5 Word Granularity

BPE segmentation is the preliminary step for current NMT systems, which may segment some words into sub-words. To explore the effect of word granularity on the miscalibration of NMT models, we divide the tokens after BPE segmentation into two categories: *Sub-Words* that are divided into word fragments by BPE (e.g., with "@@"), and *Full Words* that are not divided by BPE. Figure 8 depicts the results.



(a) Over-Estimation (b) Under-Estimation

Figure 8: Effect of word granularity on miscalibration.

Sub-words suffer more from miscalibration, while full words are less likely to be miscalibrated. The relative changes of sub-words are all positive for both over- and under-estimation, while those of full words are all negative. Sennrich et al. (2016) showed that BPE addresses the open-vocabulary translation by encoding rare and unknown words as sequences of sub-word units. Our results confirm their claim: the behaviors of sub-words and full words correlate well with those of low- and high-frequency tokens respectively.

6 Revisiting Advances in Architecture and Regularization

Guo et al. (2017) have revealed that the miscalibration on classification tasks is closely related to lack of regularization and increased model size. In this section we check whether the conclusion holds on the inference of NMT models, which belong to a family of structured generation.

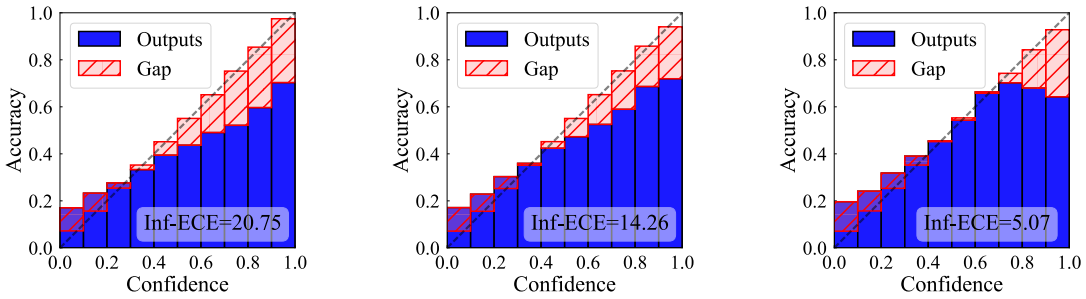
[3] If a token is a sub-word segmented by BPE, the token shares the syntactic role of the full word that it belongs to.

[4] https://nlp.stanford.edu/software/tagger.shtml

[5] http://www.phontron.com/kytea/

Label Smoothing	Dropout	Beam Size = 10				Beam Size = 100			
		BLEU	*ECE*	*Over.*	*Under.*	*BLEU*	*ECE*	*Over.*	*Under.*
×	×	23.03	25.49	58.3%	9.6%	22.90	26.46	59.4%	9.3%
✓	×	24.51	14.99	42.3%	17.3%	24.58	15.97	42.8%	16.9%
×	✓	27.52	20.75	52.3%	10.1%	26.93	22.57	53.6%	9.8%
✓	✓	27.65	14.26	39.7%	14.1%	27.68	14.75	40.1%	14.2%
GRADUATED	✓	**27.76**	**5.07**	29.1%	31.6%	**28.07**	**5.23**	29.5%	31.4%

Table 3: Results of label smoothing and dropout on the En-De task. "Over." and "Under." denote over-estimation and under-estimation, respectively.



(a) None (b) Vanilla (c) Graduated

Figure 9: Reliability diagrams of different label smoothing strategies: (a) no label smoothing; (b) vanilla label smoothing; (c) graduated label smoothing. The results are reported on the WMT14 En-De translation task.

One criticism of NMT inference is that the translation performance inversely decreases with the increase of search space (Tu et al., 2017). Quite recently, Kumar and Sarawagi (2019) claimed that this problem can be attributed to miscalibration. Accordingly, we also report results on large beam size and find that reducing miscalibration can improve the NMT performance in large beam size.

6.1 Regularization Techniques

We revisit two important regularization techniques that directly affect confidence estimation:

- *Label Smoothing* (Szegedy et al., 2016): distributing a certain percentage of confidence from the ground truth label to other labels uniformly in training.
- *Dropout* (Hinton et al., 2012): randomly omitting a certain percentage of the neural networks on each training case, which has been shown effective to prevent the over-fitting problem for large neural networks.

For comparison, we disable label smoothing or dropout to retrain the model on the whole training set. The results are shown in Table 3. We find that label smoothing improves the performance by greatly reducing the over-estimation, at the cost of increasing the percentage of under-estimation error. Dropout alleviates the over-estimation problem, and does not aggravate under-estimation. Although label smoothing only marginally improves performance on top of dropout, it is essential for maintaining the translation performance in larger search space (i.e., Beam Size = 100).

As seen from Table 3, reducing ECE can only lead to marginal BLEU gains. We attribute this phenomenon to the fact that ECE is another metric to evaluate NMT models, which is potentially complementary to BLEU. Accordingly, ECE is not necessarily strictly negatively related to BLEU.

Graduated Label Smoothing Inspired by this finding, we propose a novel *graduated label smoothing* approach, in which the smoothing penalty for high-confidence predictions is bigger than that for low-confidence predictions. We firstly use the model trained by vanilla label smoothing to estimate the confidence of each token in the training set, then we set the smoothing penalty to 0.3 for tokens with confidence above 0.7, 0.0 for tokens with confidence below 0.3, and 0.1 for the remaining tokens.

As shown in Table 3, the graduated label smoothing can improve translation performance by alle-

Enc.	Dec.	Para.	Beam Size = 10				Beam Size = 100			
			BLEU	*ECE*	*Over.*	*Under.*	*BLEU*	*ECE*	*Over.*	*Under.*
BASE	BASE	88M	27.65	14.26	39.7%	14.1%	27.68	14.75	40.1%	14.2%
DEEP	DEEP	220M	28.86	14.99	40.3%	14.1%	28.64	15.55	41.8%	14.0%
DEEP	BASE	145M	**29.09**	**14.28**	39.6%	14.1%	**29.29**	**14.53**	39.6%	14.2%
WIDE	WIDE	264M	28.66	16.09	42.3%	12.6%	28.42	17.22	43.2%	12.5%
WIDE	BASE	160M	**28.97**	**14.83**	39.7%	13.6%	**29.09**	**15.06**	39.8%	13.7%

Table 4: Effect of model size by enlarging encoder ("Enc.") and decoder ("Dec.") on the En-De dataset.

viating inference miscalibration, and the improvement is more significant in large beam size. Figure 9 shows the reliability diagrams of different label smoothing strategies. The graduated label smoothing can effectively calibrate the predictions with $0.4 \leq$ confidence ≤ 0.8, while is less effective for low- (i.e., < 0.4) and high-confidence (i.e., > 0.8) predictions. We believe that the design of more advanced techniques to solve this problem is a worthwhile future direction of research.

6.2 Increased Model Size

The model size of NMT models has increased significantly recently (Bahdanau et al., 2015; Vaswani et al., 2017; Wang et al., 2019). We evaluated the inference calibration of models with different sizes. We increase model size in the following two ways:

- *Deeper model*: both the encoder and the decoder are deepened to 24 layers;
- *Wider model*: the hidden size of the encoder and the decoder is widened to 1024.

The BLEU score and inference ECE of different models are shown in Table 4.

Increasing model size negatively affects inference calibration. We find that increasing both the encoder and the decoder increases the inference calibration error despite increasing the BLEU, confirming the finding of Guo et al. (2017) that increased model size is closely related to model miscalibration. This leads to a performance drop in a larger search space (i.e., Beam Size = 100).

Only enlarging the encoder improves translation quality while maintaining inference calibration. As the decoder is more directly related to the generation, it is more likely to result in miscalibration. In order to maintain the performance improvement and do not aggravate over-estimation, we propose to only increase the size of encoder and keep the decoder unchanged. Results in Table 4 indicate that only enlarging the encoder can achieve better performance with fewer parameters compared to enlarging both the encoder and the decoder. In a larger search space (i.e., Beam Size = 100), models with high inference ECE will generate worse translations while models with low inference ECE can achieve improved translation performance.

7 Conclusion

Although NMT models are well-calibrated in training, we observe that they still suffer from miscalibration during inference because of the discrepancy between training and inference. Through a series of in-depth analyses, we report several interesting findings which may help to analyze, understand and improve NMT models. We revisit recent advances and find that label smoothing and dropout play key roles in calibrating modern NMT models. We further propose graduated label smoothing that can reduce the inference calibration error effectively. Finally, we find that increasing model size can negatively affect the calibration of NMT models and this can be alleviated by only enlarging the encoder. As well-calibrated confidence estimation is more likely to establish trustworthiness with users, we plan to apply our work to interactive machine translation scenarios in the future.

Acknowledgments

We thank all anonymous reviewers for their valuable comments and suggestions for this work. This work was supported by the National Key R&D Program of China (No. 2017YFB0202204), National Natural Science Foundation of China (No. 61925601, No. 61761166008, No. 61772302), Beijing Advanced Innovation Center for Language Resources (No. TYR17002), and the NExT++ project supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@Singapore Funding Initiative.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Shu-Tao Xia. 2019. Adaptive Regularization of Labels. *arXiv*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *NAACL*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated structured prediction. In *NeurIPS*.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of Encoder Decoder Models for Neural Machine Translation. In *ICLR Debugging Machine Learning Models Workshop*.

Aditya Krishna Menon, Xiaoqian J Jiang, Shankar Vembu, Charles Elkan, and Lucila Ohno-Machado. 2012. Predicting accurate probabilities with a ranking loss. In *ICML*, volume 2012, page 703. NIH Public Access.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When Does Label Smoothing Help? In *NeurIPS*.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *EMNLP*.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *ICML*, pages 625–632.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *ICML*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

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

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *AAAI*.

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*.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *ACL*.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *EMNLP*.

Lijun Wu, Xu Tan, Di He, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Beyond error propagation in neural machine translation: Characteristics of language also matter. In *EMNLP*.

Dong Yu, Jinyu Li, and Li Deng. 2011. Calibration of confidence measures in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2461–2473.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699. ACM.

Wenliang Zhong and James T Kwok. 2013. Accurate probability calibration for multiple classifiers. In *Twenty-Third International Joint Conference on Artificial Intelligence*.