

Competency-Aware Neural Machine Translation: Can Machine Translation Know its Own Translation Quality?

Pei Zhang*, Baosong Yang, Haoran Wei, Dayiheng Liu, Kai Fan, Luo Si and Jun Xie
Alibaba Group Inc.

{xiaoyi.zp, yangbaosong.ybs, funan.whr, qingjing.xj}@alibaba-inc.com

Abstract

Neural machine translation (NMT) is often criticized for failures that happen without awareness. The lack of competency awareness makes NMT untrustworthy. This is in sharp contrast to human translators who give feedback or conduct further investigations whenever they are in doubt about predictions. To fill this gap, we propose a novel competency-aware NMT by extending conventional NMT with a self-estimator, offering abilities to translate a source sentence and estimate its competency. The self-estimator encodes the information of the decoding procedure and then examines whether it can reconstruct the original semantics of the source sentence. Experimental results on four translation tasks demonstrate that the proposed method not only carries out translation tasks intact but also delivers outstanding performance on quality estimation. Without depending on any reference or annotated data typically required by state-of-the-art metric and quality estimation methods, our model yields an even higher correlation with human quality judgments than a variety of aforementioned methods, such as BLEURT, COMET, and BERTScore. Quantitative and qualitative analyses show better robustness of competency awareness in our model.¹

1 Introduction

With the exploitation of large amounts of training data, neural machine translation (NMT) models (Bahdanau et al., 2015; Vaswani et al., 2017) have shown promising quality. However, NMT often makes mistakes, especially on outliers such as out-of-domain samples, input noises, and low-frequency words, on which models are potentially not trained or evaluated (Koehn and Knowles, 2017; Khayrallah and Koehn, 2018; Müller et al., 2020). Unlike human translators who are generally aware

Method	Input \rightarrow Output	Human-Assist
MT	(<i>src</i>) \rightarrow (<i>trans</i>)	N/A
Metric	(<i>src, ref, trans</i>) \rightarrow (<i>quality</i>)	✓
QE	(<i>src, trans</i>) \rightarrow (<i>quality</i>)	✓
CANMT	(<i>src</i>) \rightarrow (<i>trans, quality</i>)	×

Table 1: Comparison among Machine Translation (MT), Automatic Translation Evaluation (Metric), Quality Estimation (QE), and our Competency-Aware Neural Machine Translation (CANMT). ‘Input \rightarrow Output’ indicates the input and output for each task, where *src*, *ref*, *trans*, and *quality* represent source, reference, translation, and quality score respectively. ‘Human-Assist’ denotes whether quality evaluation depends on human-annotated references or samples with quality scores for training.

of their expertise and weaknesses, NMT lacks competency awareness, which limits the trustworthiness when it is widely adopted. Consequently, it is necessary to equip NMT models with good awareness of their translation competency, i.e. the ability to give a translation along with the quality score of the translation. With good awareness of its competency, NMT can then take conservative actions (such as human intervention) on low-quality translations, and request fewer manual inspections for high-quality ones.

In order to inform users of the competency of MT, a traditional alternative is to employ extra quality evaluation tools² that are independent of the MT model, including automatic evaluation metrics (Metric; Papineni et al., 2002; Rei et al., 2020; Wan et al., 2021) and quality estimation (QE; Kim and Lee, 2016; Fan et al., 2019). As shown in Table 1, given a source sentence and machine translation, metrics evaluate quality score relying on reference translations, thus being unsuitable for tackling this issue. Unlike metrics that require references, QE estimates the quality of translation given only the source sentence (Kim and Lee, 2016). Nevertheless, well-performing QE mod-

*Corresponding author.

¹Code and test sets are available at: <https://github.com/xiaoyi0814/CANMT>.

²For simplification, we call these methods as extra-estimation tools in the subsequent of this paper.

els are typically supervised methods trained on a large amount of expert-annotated data (Specia et al., 2020, 2021). Besides, recent studies have pointed out that the extra-estimation tools like metrics and QE potentially drop accuracy when switching to out-of-domain (Freitag et al., 2021).

In this paper, we propose a novel method, **Competency-Aware Neural Machine Translation (CANMT)**, which has both translation and self-estimation capabilities without any ground-truth translation or expert-scored data for the competency assessment. Based on the assumption that “a good translation should be able to reconstruct the meaning of source sentence from it”, we extend the standard translation model with a self-estimator, which adopts a reconstruction strategy, and use the semantic gap between the original source sentence and its reconstruction to achieve competency awareness. Our self-estimator is a part of the NMT model and exploits the continuous representations as input. The intuition we believe is that continuous representations are more informative about the decoding procedure of NMT, thus better reflecting the quality of translation candidates than the discrete output. Moreover, we update the original NMT decoder to a two-stream decoder, in which translation and reconstruction information flows are separated to avoid the interference of the source sentence during reconstruction.

We conduct experiments on four widely used translation tasks: CWMT17 Chinese-to-English (Zh→En), ASPEC Japanese-to-English (Ja→En), WMT14 French-to-English (Fr→En) and WMT14 English-to-German (En→De). Results show that while keeping the same translation performance as Transformer (Vaswani et al., 2017) baseline, CANMT exhibits promising capability on self-estimation of quality, which outperforms the state-of-the-art unsupervised metrics and QE methods such as BERTScore (Zhang et al., 2020), RTT (Moon et al., 2020) and confidence-based QE (Fomicheva et al., 2020). It is encouraging to see that our method even surpasses those existing supervised extra-estimation tools like BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020) which depend on ground-truth translations or learn from expert-scored data. Extensive analyses demonstrate the robustness of CANMT that effectively tackles problems of out-of-domain, quality drift, and miscalibration in self-estimation. Complementary effects are also observed between

CANMT and extra-estimation tools.

2 Competency-Aware Neural Machine Translation (CANMT)

Given an arbitrary source sentence X , the goal of CANMT model is to generate a target translation Y and assign a self-estimated competency score $Q_{X \rightarrow Y}$ for the generated translation Y without any reference translation or expert-annotated data, which can be represented as:

$$Y, Q_{X \rightarrow Y} = \text{CANMT}(X). \quad (1)$$

We propose to exploit the reconstruction strategy and use the semantic gap between original source sentence and its reconstruction to enable competency awareness. The main challenge of this approach is how to capture as many quality issues of the predicted translation as possible in reconstructing, such as semantic or syntactic errors and disfluency. To overcome this challenge, we utilize the continuous representation from the NMT decoder, as it contains more information about the decoding procedure than a discrete translation and thus is easier to reconstruct the source sentence.

2.1 Model Architecture

The architecture of CANMT is illustrated in Figure 1. Our model consists of three components: **encoder**, **two-stream decoder**, and **self-estimator**, which are all built with Transformer blocks (Vaswani et al., 2017). The source sentence X is first encoded by the encoder, then passed to the two-stream decoder. One stream jointly attends to source representations to generate a target translation Y . The other feeds the information about the decoding procedure to the self-estimator, which then reconstructs the source sentence and estimates quality using cross-entropy loss. In the following, we specify the structure of each component.

Encoder The encoder of CANMT is the same as the standard Transformer. It maps the token sequence of input sentence $X = (x_1, \dots, x_l)$ (with length of l) to a contextual representation C_X :

$$C_X = \text{Encoder}([e(x_1), \dots, e(x_l)]; \theta_{Enc}), \quad (2)$$

where $e(\cdot)$ is the source embedding lookup table and θ_{Enc} denotes the encoder parameters. Then C_X is fed into the decoder to generate translation.

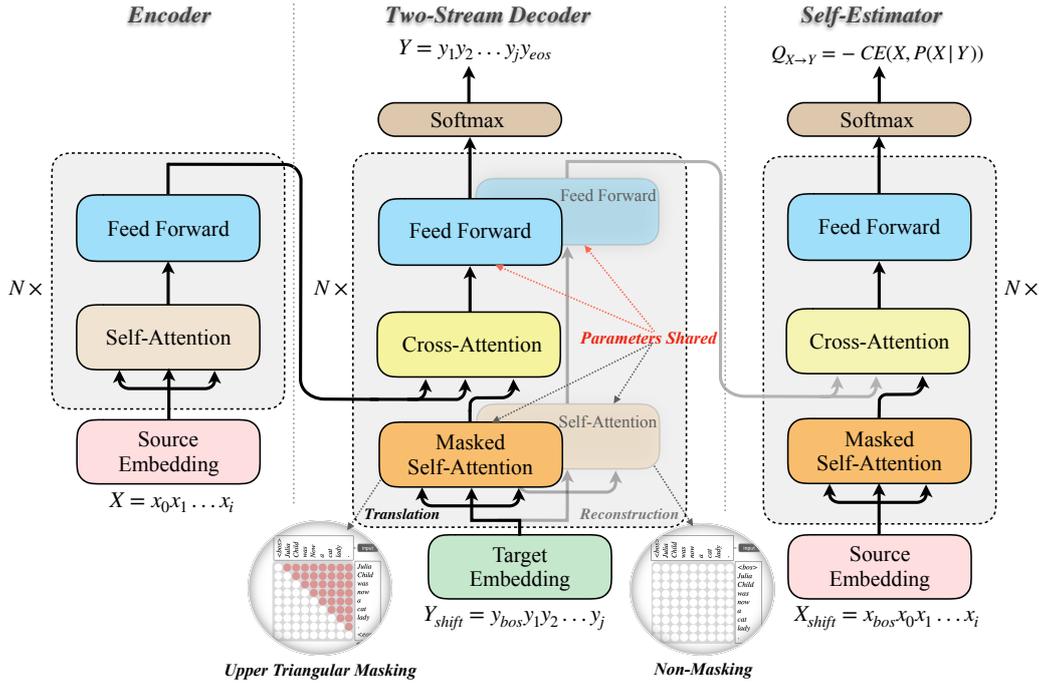


Figure 1: Illustration of CANMT, where the encoder-decoder NMT model is extended with a self-estimator based on reconstruction. We introduce a two-stream mechanism in the decoder for both translation and reconstruction. A source sentence X is first encoded, and then joint-attended by *Translation* stream of the decoder to generate translation Y . The continuous representation from *Reconstruction* stream of the decoder is further fed into the self-estimator to reconstruct X . Then competency is estimated by the cross-entropy (CE) loss of reconstruction. The difference in self-attention masks for the two-stream decoder is elaborated in the circular zoom areas.

Two-Stream Decoder Since we propose to use the information of translation procedure for source-side reconstruction, however, the cross-attention in regular decoder introduces source-side information which makes the learning of reconstruction ineffective. Therefore, we adopt a controllable two-stream mechanism in the decoder. Given a target translation $Y = (y_1, \dots, y_m)$ (with length of m),

- *Translation* stream is in charge of forward translation, which serves a similar role in the vanilla Transformer. This stream can have access to the sequential information $y_{<t}$ of each target word y_t and the source context.
- *Reconstruction* stream is in charge of the source-side reconstruction. This stream instead captures the full contextual information $y_{1:m}$ but does not include the source context to avoid information leaking.

We share the parameters in the two-stream mechanism to retain more information about the translation decoding procedure for reconstruction.

Computationally, we denote the representation of t -th layer in *Translation* and *Reconstruction*

streams as h and g , respectively. The first layer representations of the two streams are both set as the precedent word embedding, i.e. $h_t^0 = g_t^0 = e(y_t)$. For each layer $n = 1, 2, \dots, N$, the representations of two streams are calculated as follows:³

- for *Translation* stream

$$\begin{aligned} \hat{h}_t^{(n)} &= \text{MHA}_{\text{slf}}^{(n)}(Q=h_t^{(n-1)}, KV=h_{<t}^{(n-1)}) \\ h_t^{(n)} &= \text{FFN}^{(n)}(\text{MHA}_{\text{ctx}}^{(n)}(Q=\hat{h}_t^{(n)}, KV=C_X)) \end{aligned} \quad (3)$$

- for *Reconstruction* stream

$$\begin{aligned} \hat{g}_t^{(n)} &= \text{MHA}^{(n)}(Q=g_t^{(n-1)}, KV=g_{1:m}^{(n-1)}) \\ g_t^{(n)} &= \text{FFN}^{(n)}(\hat{g}_t^{(n)}) \end{aligned} \quad (4)$$

where $\text{MHA}_{\text{slf}}^{(n)}$, $\text{MHA}_{\text{ctx}}^{(n)}$ and $\text{FFN}^{(n)}$ are self-attention, cross-attention and position-wise feed-forward sub-layers in the n -th decoder layer. As the causal attention masking (normally upper triangular masking) in the self-attention of the Transformer decoder would limit the contextual modeling ability and undermine the reconstruction performance, we discard this mask in *Reconstruction* stream to enable the full access of context. A specific example

³Here, we omit the computation of LayerNorm and residual connections for simplification.

illustrating the difference in masking mechanism for two streams is presented in the circular zoom areas in Figure 1.

The output of *Translation* stream is used to generate the final translation. We use a learned linear transformation and the softmax function to convert $h_i^{(N)}$ to the probability $P(y_i|y_{<i}, X)$ of the i -th token, and obtain translation Y by maximizing the product of these probabilities:

$$Y = \operatorname{argmax} \prod_{t=1}^m P(y_t|y_{<t}, X). \quad (5)$$

Specifically, the translation inference is performed by the commonly used beam search. The output of *Reconstruction* stream, which is denoted as $C_Y = [g_1^{(N)}, \dots, g_m^{(N)}]$, is fed into the self-estimator to obtain the self-estimation score of Y .

Self-Estimator The structure of the self-estimator is the same as that of the regular decoder. In the inference time, the hidden states of *Reconstruction* stream of decoder C_Y are jointly attended by self-estimator to reconstruct the source sentence, and the output probability $P(X|Y) = \prod_{i=1}^l P(x_i|x_{<i}, Y)$ is computed by forced decoding with X .

The quality score of translation can be deduced by the negative Cross-Entropy (CE) loss between the output probability of self-estimator and X , which can be mathematically represented as:

$$Q_{X \rightarrow Y} = -\operatorname{CE}(X, P(X|Y)) \\ = \frac{1}{l} \sum_{i=1}^l \log P(x_i|x_{<i}, Y). \quad (6)$$

For any given example, the higher the cross-entropy is, the poorer the quality of the corresponding translation, which reflects the inability in this case.

2.2 Training

The training of CANMT relies entirely on a bilingual corpus and aims at joint learning of forward translation and backward reconstruction. Let $\mathcal{D} = \{(X^n, Y^n)\}_{n=1}^{|\mathcal{D}|}$ denote the training corpus with $|\mathcal{D}|$ paired sentences, where (X^n, Y^n) is a pair of sentences in source and target languages. The final training objective is to minimize the following cross-entropy loss function:

$$\mathcal{L} = \mathbb{E}_{(X^n, Y^n) \sim \mathcal{D}} [-\log P(Y^n|X^n; \theta_{Enc}, \theta_{Dec}) \\ - \log P(X^n|Y^n; \theta_{Dec}, \theta_{Est})], \quad (7)$$

where θ_{Enc} , θ_{Dec} , θ_{Est} indicate the parameters of encoder, decoder and self-estimator. The two negative log-likelihoods respectively represent the training objectives of forward translation and backward reconstruction. Significantly, the input to decoder during training is ground truth, while the input of that during inference is the model translation.

3 Experiments

For the competency awareness of CANMT, the evaluation of performance consists of two aspects: translation and self-estimation. We use BLEU to validate translation performance. As for self-estimation, it considers whether an MT model is aware of its own competency, thus public metric and QE tests are unsuitable for us as the translations are from other models. Therefore, we generate translations by CANMT and evaluate the Pearson correlations between the self-estimation scores and human judgments.

3.1 Experimental Setting

Dataset We train and evaluate models on four standard machine translation datasets: CWMT17 Zh→En, WMT14 Fr→En, WMT14 En→De of news domain and ASPEC (Nakazawa et al., 2016) Ja→En of the scientific domain. We segment Chinese and Japanese sentences with Jieba⁴ and KyTea⁵ respectively, while the segmentation for the other languages is processed with Moses⁶ tokenizer. The byte-pair encoding (Sennrich et al., 2016) for Zh→En, Fr→En, En→De, and Ja→En are trained respectively with 32k/40k/32k/16k merge operations and source-target vocabularies are shared except for Zh→En. We follow the standard setup of valid and test sets for each task. Statistics and details are presented in Appendix A.

Implementation Details The proposed CANMT can be exploited in both RNN and Transformer models. We implement it based on the Transformer-based model with Fairseq (Ott et al., 2019). Our encoder and decoder follow the hyper-parameter setting of Base Transformer (Vaswani et al., 2017), as well as the self-estimator, which has the same setting as the decoder. All the experiments are applied with the share-embedding setting except for Zh→En and trained on 8 NVIDIA Tesla V100 GPUs with a batch size of 4096 tokens. We use

⁴<https://github.com/fxsjy/jieba>

⁵<http://www.phontron.com/kytea/>

⁶<https://github.com/alvations/sacremoses>

Tasks	Methods	Ref.	Zh→En	Fr→En	Ja→En	En→De	Average
MT	Transformer		23.1	36.1	28.9	26.9	28.8
	CANMT (Ours)		23.5	35.8	29.0	26.8	28.8
Supervised Methods							
	COMET-QE (Rei et al., 2020)	×	0.51	0.54	0.13	0.57	0.44
	BLEURT (Sellam et al., 2020)	✓	0.46	0.49	0.28	0.31	0.39
	COMET (Rei et al., 2020)	✓	0.49	0.50	0.35	0.58	0.48
Unsupervised Methods							
Eval.	RTT-SentBLEU (Moon et al., 2020)	×	0.22	0.13	0.20	0.28	0.21
	RTT-BERTScore (Moon et al., 2020)	×	0.21	0.24	0.30	0.49	0.31
	TP	×	0.35	0.41	0.41	0.41	0.40
	D-TP($K = 30$) (Fomicheva et al., 2020)	×	0.55	0.46	0.44	0.50	0.49
	SentBLEU (Papineni et al., 2002)	✓	0.08	0.15	0.09	0.23	0.14
	BERTScore (Zhang et al., 2020)	✓	0.37	0.42	0.31	0.41	0.38
	CANMT (Ours)	×	<u>0.61</u>	<u>0.50</u>	<u>0.52</u>	<u>0.52</u>	<u>0.54</u>

Table 2: Results on four translation tasks. The performance of translation (MT) is measured by BLEU, while the performance of quality evaluation (Eval.) is assessed by Pearson correlations with human judgments. “Ref.” indicates whether reference translations are needed for Eval. “Average” denotes the average performance of four tasks. The best results of all the methods are marked in bold, and that of all the unsupervised methods are underlined. CANMT offers better quality evaluation ability on its translations than its unsupervised and supervised counterparts, without drops in translation quality.

beam search with a beam size of 8 and length penalty of 1.0 for Zh→En, while for the other translation tasks we set beam size to 4 and length penalty to 0.6. We perform early stopping on valid sets, average the last 5 checkpoints and report case-sensitive SacreBLEU (Post, 2018) on test sets.

Baselines For translation performance, we compare CANMT with Transformer, while for self-estimation we compare with widely used metric and QE methods:

- **SentBLEU** denotes the sentence-level BLEU that computed by SacreBLEU (Post, 2018).
- **BERTScore** (Zhang et al., 2020) calculates the cosine similarity between the sentence embeddings of hypothesis and reference translations based on BERT (Devlin et al., 2019).
- **BLEURT** (Sellam et al., 2020) and **COMET** (Rei et al., 2020) are two supervised regression-based metrics which are built based on pre-trained language models and then fine-tuned on human-scored data.
- **COMET-QE** (Rei et al., 2020) is the reference-less version of COMET.
- **RTT-SentBLEU** and **RTT-BERTScore** (Moon et al., 2020) respectively report the sentence-level BLEU and BERTScore between a source sentence and a reconstruction outputted by an extra backward MT as introduced in Section 5.2.⁷

⁷We used Google Translate as the backward translation system because it outperforms models trained on the WMT

Methods	News	Subtitles	Laws	Ted	Medical
Supervised Methods					
COMET-QE	0.51	0.41	0.36	0.39	0.51
BLEURT	0.46	0.28	0.60	0.40	0.45
COMET	0.49	0.35	0.57	0.47	0.48
Unsupervised Methods					
TP	0.35	0.45	0.24	0.48	0.28
D-TP	0.55	0.58	0.41	0.54	0.48
SentBLEU	0.08	0.18	0.41	0.22	0.09
BERTScore	0.37	0.43	0.46	0.44	0.37
CANMT(Ours)	0.61	0.58	0.66	0.58	0.65

Table 3: Pearson correlation on out-of-domain sets of Zh→En translation direction. The best results of all the methods are marked in bold. CANMT performs better robustness on four unlearned domains.

- **Translation Probability (TP)** (Fomicheva et al., 2020): the sentence-level translation probability normalized by length.
- **Monte Carlo Dropout Translation Probability (D-TP)** (Fomicheva et al., 2020): the expectation of TP scores obtained by running $K=30$ stochastic forward passes through MT model perturbed by Monte Carlo Dropout.

Human Evaluation We randomly select 200 sentences for each test set and translate them into the target language using CANMT models. Each sentence is scored by three professional translators presented with pairs of a source sentence and a translated hypothesis, following the scoring criteria

News training corpus in previous work.

ranging from 1 to 5⁸ (larger is better) in previous work (Weng et al., 2020). We use the average value given by three experts as the final human judgment.

3.2 Main Results

The system-level BLEU results shown in the upper part of Table 2 verify that our method holds the same translation performance as Transformer, which is in line with our expectations. Therefore, we focus on the evaluation of self-estimation in the following. The lower part of Table 2 shows the Pearson correlations between the predicted scores and human judgments. In general, CANMT performs pretty well and stably across tasks.

Comparison to the supervised methods. Overall, our approach surpasses state-of-the-art methods like COMET, COMET-QE, and BLEURT. Concretely, CANMT has advantages of 0.1 and 0.17 Pearson scores over the best supervised method on Zh→En and Ja→En and yields considerable results on Fr→En and En→De. In the following, we will further verify the robustness of out-of-domain.

Comparison to the unsupervised methods. CANMT outperforms all unsupervised methods by a large margin. Reference-based metrics like BLEU and BERTScore rely on the given reference, but there are many possible correct translations for a given source. They are good at distinguishing “good” translations that have a high degree of overlap with the given reference, but poor at judging translations that differ largely (Freitag et al., 2020). The improvements over RTT justify our motivation that the continuous representations from the decoder capture more information about the decoding than discrete translation, thus facilitating to quality estimation. On the other hand, computing cross-entropy loss with forced decoding way further avoids the error accumulation of beam search in RTT during the reconstruction procedure. Confidence-based QE method, i.e. D-TP, also gets good results. Nevertheless, confidence-based methods mainly utilize the prediction distribution of NMT regardless of the fidelity of the source sentence and the latent states in the model, resulting in miscalibration between accuracy and confidence (Guo et al., 2017; Wang et al., 2020). More analyses are conducted in Section 4.2.

Robustness under Out-of-Domain To measure the reliability of CANMT under out-of-domain con-

ditions, we further validate our method on four out-of-domain Zh→En test sets. Neither our model nor the other methods are trained or fine-tuned on data of these domains. We collect the test sets from UM-Corpus, IWSLT, and Medline abstracts of the WMT21 biomedical task, and randomly sample around 200 sentences for each domain test set. Results in Table 3 show that our method can outperform other baselines, and achieve stable performance with smaller variances across all domain tests. As proven by Freitag et al. (2021), extra-estimation tools like COMET and BLEURT are proficient at domains on which they have trained or fine-tuned, but probably deficient at unseen domains. On the contrary, our method performs better across out-of-domain tests due to its competency awareness. Even if the domain has not been learned by CANMT, it still knows its incapability.

4 Analysis

In this section, we conduct analyses on competency awareness to explain why CANMT performs better.

4.1 Robustness under Quality Drift

In order to verify whether CANMT behaves stably under different levels of translation quality, we draw on the quality-biased evaluation method in previous work of text generation evaluation (Guan and Huang, 2020). We create four quality-biased test sets from Zh→En task by sampling human scores from four quality levels, each ranging (1, 2], (2, 3], (3, 4], (4, 5] to represent quality from poor to good. The distribution (on the left) and correlation (on the right) for each quality-biased test-set are illustrated in Figure 2a (news test set) and Figure 2b (the hybrid multi-domain test set). The Pearson scores of our method are stable across all quality-biased tests. The main reason lies in the fact that the self-estimator of CANMT is trained jointly with NMT and also takes advantage of more decoding information from the MT model, which makes it aware of its own expertise and weaknesses.

4.2 Analysis on Miscalibration

It is known that neural networks suffer from the miscalibration problem between accuracy and confidence (Guo et al., 2017; Wang et al., 2020). Over-confidence can lead to predicting high confidence for poor translations, while under-confidence leads to low confidence for good translations. Here, we evaluate whether CANMT alleviates the over-

⁸Please refer to Appendix C for the scoring standard.

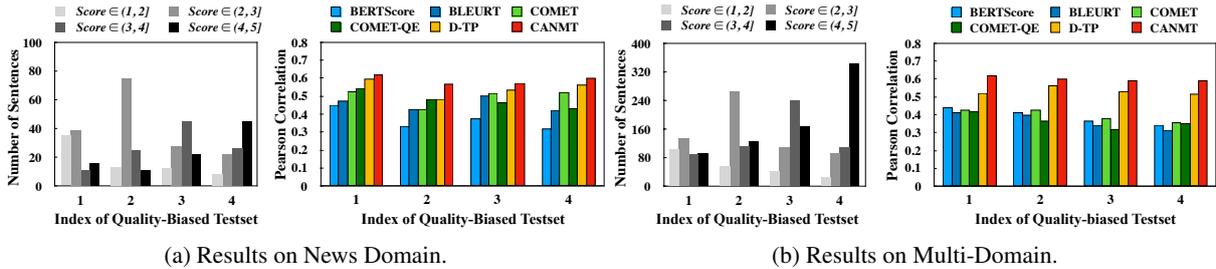


Figure 2: Robustness on different quality-biased test sets. Four quality-biased tests (quality level improves from 1 to 4) are built according to the translation quality judged by linguists (left in each sub-figure). We show the Pearson correlation on different tests (right in each sub-figure). CANMT can better handle the problem of quality drift.

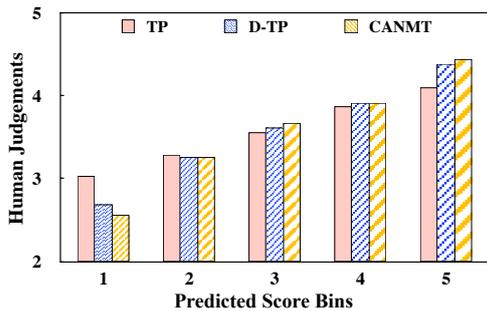


Figure 3: Analysis on miscalibration. The x-axis represents five bins that divide translation samples according to predicted scores in ascending order. The y-axis indicates the average human score in each bin.

and under-confidence problem on Zh→En multi-domain tests. For each method, we split sentences into five bins according to the predicted scores (ranging from low quality to high) and compute the average human scores. We plot these values of CANMT and confidence-based QE methods (TP and D-TP) in Figure 3.

As seen, comparing with CANMT, TP and D-TP achieve much higher average human scores in bin 1 (with the lowest predicted scores) but much lower ones in bin 5, suggesting that these methods face more serious over- and under-confidence problems than ours. Our approach performs better because of using more informative features from NMT models and an effective reconstruction strategy that considers the fidelity of the source sentence.

4.3 Complementary Effects with Extra-Estimation

Our method has the advantage of being more robust and better aware of its own competency, while extra-estimation is able to capture the supervised knowledge of the labeled data. One interesting question is whether they can be complementary. To verify it, we simply sum up the z-normalized

scores predicted by any two different methods for each sentence. We plot all the results in Figure 4.

Firstly, we can see that almost all the combinations of CANMT and extra-estimation tools (metrics or QE) outperform both its corresponding two single methods. The best combination of CANMT and extra-estimation can outperform the best single method with 0.06/0.06/0.02/0.11 improvements on the four tasks. This supports the complementary effects between extra-estimation and our method.

Secondly, we can find that on all the tasks, the best-performing combination of each task always consists of one extra-estimation method and our method, which also outperforms the combination of the top two single methods. For example, in En→De task, the combination of CANMT (0.52) and COMET-QE (0.57) outperforms the combination of COMET (0.58) and COMET-QE (0.57) with 0.04, although the latter two methods are stronger single methods. This indicates the higher complementary effects between extra-estimation and CANMT are not simply explained by the ensemble of two strong methods. It is worth further research on how to integrate supervised information of human-scored data into competency awareness.

5 Related Work

Our work is related to studies in the translation evaluation community. As it currently stands, machine translation is generally evaluated by extra-estimation tools, including reference-based methods (metrics) and reference-less methods (QE).

5.1 Metric

Existing metrics can be roughly categorized into two types: unsupervised metrics like BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009) CHRf (Popovic, 2015), and BERTScore (Zhang et al., 2020), and super-

	BLEURT	COMET	COMET-QE	CANMT												
BLEURT	0.46	0.50	0.54	<u>0.64</u>	0.49	0.51	0.57	<u>0.60</u>	0.29	0.33	0.25	<u>0.52</u>	0.31	0.52	0.54	<u>0.52</u>
COMET		0.49	0.55	<u>0.66</u>		0.50	0.57	<u>0.59</u>		0.35	0.29	<u>0.54</u>		0.58	0.64	<u>0.65</u>
COMET-QE			0.51	<u>0.67</u>			0.54	<u>0.60</u>			0.13	<u>0.46</u>			0.57	<u>0.68</u>
CANMT				<u>0.61</u>				<u>0.50</u>				<u>0.52</u>				<u>0.52</u>
	(a) Zh→En				(b) Fr→En				(c) Ja→En				(d) En→De			

Figure 4: Complementary between different methods, where values in the cells of diagonal are the Pearson correlations of a single method, those of upper triangular matrix are the combination of two methods. Best values are underlined. As seen, ours and extra-estimation methods are, to some extent, complementary to each other.

vised metrics like ESIM (Mathur et al., 2019), BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020) and UniTE (Wan et al., 2022). The evaluation ability of a metric relies on the given ground-truth references, which, however, are unavailable in a real-world application scenario. Moreover, despite achieving strong performance, most of the state-of-the-art metrics are resource-heavy, relying on large-scale pre-trained language models and a significant amount of in-domain labeled data for training. The findings of Freitag et al. (2021) highlight that all state-of-the-art metrics exhibit performance drops when switching to out-of-domain test sets. In contrast to these methods, CANMT exploits the internal information from the MT model to enable self-estimation of competency. Via decoupling the dependence on the human translation or expert-annotated data, our method is marginally affected by domain drift.

5.2 Quality Estimation

Different from Metrics, QE methods estimate translation quality without ground-truth references. There exist traditional feature-based approaches (Specia et al., 2009) and neural network based methods (Kepler et al., 2019; Fonseca et al., 2019; Specia et al., 2020) that exploit large scale language models and supervised on human-annotated data. Moon et al. (2020) proposes an unsupervised method (RTT) that reconstructs source sentences by another backward MT and then estimates translation quality in terms of the lexical or semantic similarity between the original source sentence and its reconstruction through BLEU or BERTScore. Wang et al. (2019) and Fomicheva et al. (2020) propose confidence-based unsupervised QE methods that employ the model confidence to estimate quality with Monte Carlo Dropout. Nevertheless, these methods merely consider the predicted distribution, regardless of the

fidelity with respect to source sentences and the internal information of NMT. This makes confidence-based method exist certain miscalibration (over- and under-confidence) between translation accuracy and confidence (Guo et al., 2017; Wang et al., 2020; Wan et al., 2020). Besides, the Monte Carlo sampling requires repeatedly inference decoding and is thus heavily time-consuming.

5.3 Reconstruction

Reconstruction strategy has been widely used in machine translation (He et al., 2016; Liu et al., 2016; Cheng et al., 2016; Tu et al., 2017; Yee et al., 2019) to improve translation performance. However, we employ reconstruction to equip machine translation with self-estimation. Unlike previous studies that simply use an individual backward NMT for reconstruction, we make the self-estimator a part of the NMT model and exploit the continuous representations as input from the NMT decoder for reconstruction, which performs better in reflecting the quality of the translation.

6 Conclusion

In this paper, we pay attention to the competency awareness of NMT. The major contributions of our work are three-fold: 1) We believe that competency awareness should be an essential capability of NMT and requires careful study. We hope our study can attract more interest in this problem; 2) We propose CANMT, which is able to output both the competency and the translation for a given source sentence. It novelly leverages the internal information of itself to estimate the competency, showing superiority compared to other approaches that merely consider the outputs or the prediction distribution of NMT; and 3) Empirical results exhibit the strong performance and robustness of CANMT on quality estimation, which, to the best of our knowledge, is the first report that self-estimation method sur-

passes supervised ones across translation directions and domains.

For future work, it is a promising direction to employ competency awareness in reinforcement learning for NMT. The self-estimated competency scores correlate better with human judgments than traditional reinforcement learning rewards such as sentence-level BLEU (Wu et al., 2018), we can use it to find weaknesses in NMT and take appropriate optimization.

Limitations

The limitations of CANMT are mainly twofold. Firstly, the self-estimation performance of CANMT can only be evaluated by human translators at present, and there is a lack of an automated evaluation method, which may limit its further exploration. Secondly, as we focus on the competency awareness of MT, we only verify the effectiveness of the model in estimating its own translation quality, and it is unknown for estimating the quality of translations generated by other NMT models.

Ethics Statement

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors. Informed consent was obtained from all individual participants included in the study.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- Yong Cheng, Wei Xu, Zhongjun He, W. He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. abs/1606.04596.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
- Kai Fan, Bo Li, Fengming Zhou, and Jiayi Wang. 2019. "bilingual expert" can find translation errors. In *AAAI*.
- M. Fomicheva, Shuo Sun, Lisa Yankovskaya, F. Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*.
- Erick Rocha Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *WMT*.
- Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *EMNLP*, abs/2004.06063.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Alan Stewart, George Foster, Alon Lavie, and Ondrej Bojar. 2021. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In *WMT*.
- Jian Guan and Minlie Huang. 2020. Union: An un-referenced metric for evaluating open-ended story generation. In *EMNLP*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.
- Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *NIPS*.
- Fábio Natanael Kepler, Jonay Trénous, Marcos Vinícius Treviso, Miguel Vera, and André F. T. Martins. 2019. Openkiwi: An open source framework for quality estimation. *ArXiv*.
- Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *NMT@ACL*.
- Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *NAACL*.
- Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *NMT@ACL*.
- Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*.
- Lemao Liu, Masao Utiyama, Andrew M. Finch, and Eiichiro Sumita. 2016. Agreement on target-bidirectional neural machine translation. In *NAACL*.
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *ACL*.
- Jihyung Moon, Hyunchang Cho, and Eunjeong Lucy Park. 2020. Revisiting round-trip translation for quality estimation. In *EAMT*.

- Mathias Müller, Annette Rios Gonzales, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *AMTA*.
- Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. *Aspec*: Asian scientific paper excerpt corpus. In *LREC*.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- Maja Popovic. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *WMT@EMNLP*.
- Matt Post. 2018. A call for clarity in reporting bleu scores. In *WMT*.
- Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *EMNLP*.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *ACL*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909.
- Lucia Specia, F. Blain, M. Fomicheva, Erick Rocha Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the wmt 2020 shared task on quality estimation. In *WMT*.
- Lucia Specia, F. Blain, M. Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *WMT*.
- Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *EAMT*.
- Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *AAAI*.
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- Yu Wan, Dayiheng Liu, Baosong Yang, Tianchi Bi, Haibo Zhang, Boxing Chen, Weihua Luo, Derek F. Wong, and Lidia S. Chao. 2021. Robleurt submission for wmt2021 metrics task. *ArXiv*, abs/2204.13352.
- Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek F. Wong, and Lidia S. Chao. 2022. Unite: Unified translation evaluation. In *ACL*.
- Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. Self-paced learning for neural machine translation. In *EMNLP*.
- Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. *ArXiv*.
- Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *ACL*.
- Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. In *EMNLP*.
- Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. *ArXiv*, abs/1808.08866.
- Kyra Yee, Nathan Ng, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. abs/1908.05731.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *ArXiv*.

A Dataset and Processing Strategies

Corpus	Language	Domain	Data Size		
			train	dev	test
ASPEC	Ja→En	Science	2M	1790	1812
CWMT17	Zh→En	News	9.2M	2002	2001
WMT14	Fr→En	News	35.7M	26854	3003
WMT14	En→De	News	4.5M	3000	3003

Table 4: Number of parallel sentences in each dataset. “M” stands for “million”.

The corpus size and processing details are shown in Table 4. The byte-pair encoding (Sennrich et al., 2016) for Zh→En, Fr→En, En→De, and Ja→En are trained respectively with 32k/40k/32k/16k merge operations and source-target vocabularies are shared except for Zh→En.

B Baseline Details

We use the official code for each quality estimation tool. For BERTScore, the officially recommended F1 score is adopted. BERTScore-en based on the pre-trained model RoBERTa-large is used for to-English translation tasks, while BERTScore-multi based on the pre-trained model BERT-base is used for En→De. The evaluated metrics BLEURT and COMET are pre-trained on BERT-base and

XLM-RoBERTa-large respectively, and then fine-tuned on WMT15-18 Metric rating datasets(for BLEURT) and WMT17-19 Metric Direct Assessment datasets(for COMET).

C Standard of Human Evaluation

The human evaluation of translation quality mainly focuses on two aspects: accuracy and fluency. Details about human scoring for translation quality are elaborated in Table 5. Each sentence is independently annotated by three qualified translators from professional language service providers. We train the annotators before annotations. Any evaluation in which the variance of scores is greater than 0.5 will be rechecked to ensure consistency. We use the average score given by three experts as the final human judgment. The human annotation results are released in the supplemental material.

D Computational Cost

According to inference, the translation cost is the same as Transformer, while the self-estimation is equal to the forced decoding of the Transformer Decoder. For example, the QE inference costs (in seconds) for the 200-sentence Fr→En test set of each method are as follows: CANMT 5 < BLEURT 15 < BERTScore 31 < COMET-QE 42 < COMET 43 < D-TP 70. The computation of training can be represented as much as twice of a standard Transformer.

Score	Accuracy	Fluency
5/Excellent	The translation (basically) faithfully reflects the meaning of the original text.	Accurate grammar and words. Fluent and idiomatic.
4/Good		The grammar and words are accurate. The translation is relatively fluent with only a few errors that do not affect understanding.
3/Acceptable		The grammar is basically accurate, but there are some problems, such as inaccurate words, improper collocation, and insufficient fluency, which can be fixed at a small cost.
2/Poor		The syntax is basically accurate, but there are obvious semantic losses or errors.
1/Very bad		Difficult to read or understand. The translation is irrelevant to the semantics of the original text.

Table 5: Standard of Human Evaluation.