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

Pei Zhang*, Baosong Yang, HaoranWei, Dayiheng Liu, Kai Fan, Luo Si and Jun Xie

DAMO ALIBABA DAMO ACADEMY

Motivation

Untrustworthy of Neural machine translation (NMT): NMT is often criticized for failures that happen without awareness, especially when it is widely adopted.

Human competency awareness: Human translators give feedback or conduct further investigations whenever they are in doubt about predictions.

Can machine translation know its own translation quality? We propose a novel competency-aware NMT (CANMT) by extending conventional NMT with a self-estimator, offering abilities to translate a source sentence and estimate its competency (i.e. translation quality).

Differences with previsou methods

Method	Input → Output	Human-Assist
MT	*(src)* → *(trans)*	N/A
Metric	*(src, ref, trans)* → *(quality)*	✓
QE	*(src, trans)* → *(quality)*	✓
CANMT	*(src)* → *(trans, quality)*	×

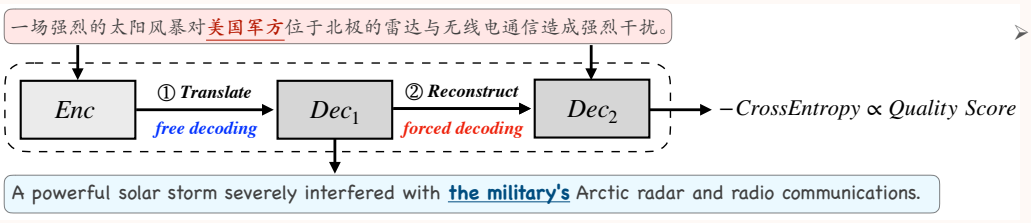
Metric and QE
- Independent from MT model.
- Rely on human-annotated references or quality scores for training.

CANMT
- **Translate and self-estimate in one model.**
- Requires no references or quality scores for training and is trained entirely on a bilingual corpus.

Method

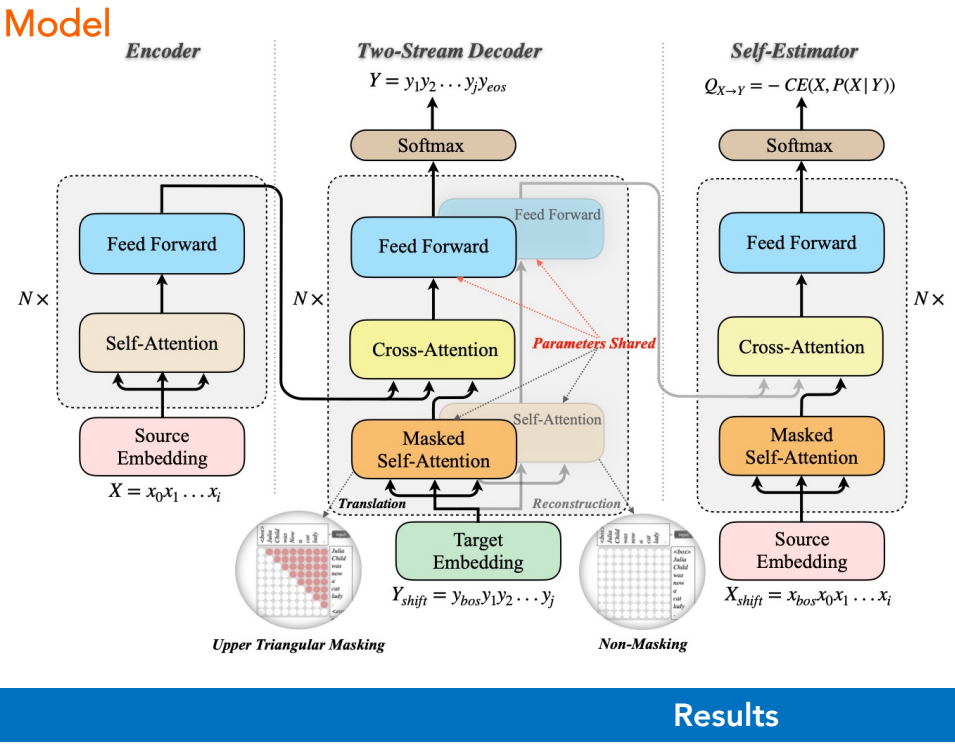
Assumption: A good translation should be able to reconstruct the meaning of the source sentence from it.

➢ We use the semantic gap between the original source sentence and its reconstruction to enable competency awareness.



➢ To better reflect the quality issues of predicted translations in reconstructing, we utilize the continuous representation from the NMT decoder, as it contains more information about the decoding procedure than a discrete translation.

Model



Two-Stream Decoder

Translation Stream generates the target translation.

Reconstruction Stream captures the target decoding information and is in charge of the source-side reconstruction.

A source sentence X is first encoded, and then joint-attended by the Translation stream of the decoder to generate translation Y. The continuous representations from the Reconstruction stream of the decoder are further fed into the self-estimator to reconstruct X. Then translation competency is estimated by the cross-entropy (CE) loss of reconstruction.

Results

CANMT offers better quality evaluation ability on its translations than the unsupervised and supervised counterparts, without drops in translation quality. Translation (MT task in BLEU) and self-estimation (Eval. Task in Pearson Correlation).

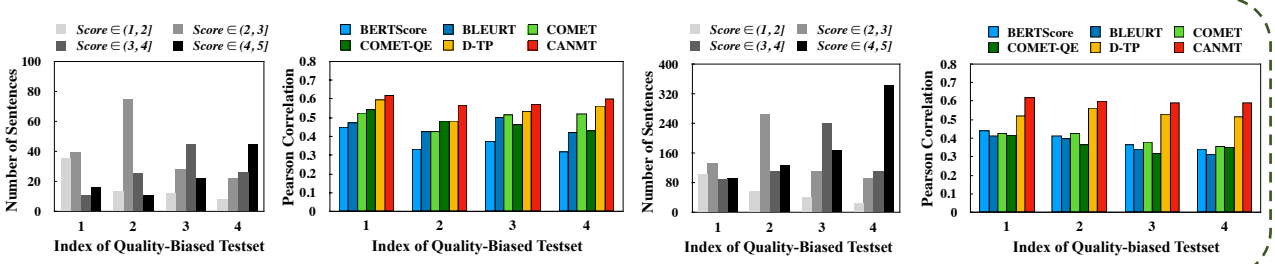
Tasks	Methods	Ref.	Zh→En	Fr→En	Ja→En	En→De	Average
MT	Transformer		23.1	36.1	28.9	26.9	28.8
	CANMT (Ours)		23.5	35.8	29.0	26.8	28.8
Eval.	Supervised Methods						
	COMET-QE (Rei et al., 2020)	×	0.51	**0.54**	0.13	0.57	0.44
	BLEURT (Sellam et al., 2020)	✓	0.46	0.49	0.28	0.31	0.39
	COMET (Rei et al., 2020)	✓	0.49	0.50	0.35	**0.58**	0.48
	Unsupervised Methods						
	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)	×	**0.61**	0.50	**0.52**	0.52	**0.54**

Better robustness on four unlearned domains.

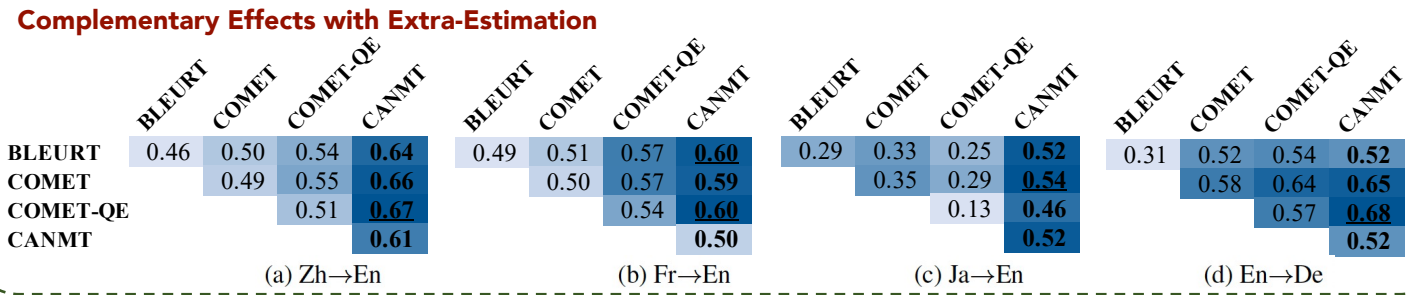
Methods	News	Subtitles	Laws	Ted	Medical
Supervised Methods					
COMET-QE	0.51	0.41	0.36	0.39	0.51
BLEURT	0.46	0.28	0.60	0.40	0.45
COMET	0.49	0.35	0.57	0.47	0.48
Unsupervised Methods					
TP	0.35	0.45	0.24	0.48	0.28
D-TP	0.55	**0.58**	0.41	0.54	0.48
SentBLEU	0.08	0.18	0.41	0.22	0.09
BERTScore	0.37	0.43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

More robust under quality drift

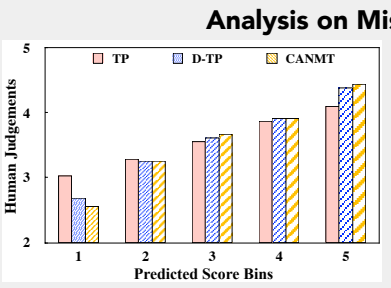
Left (News)
Right (Multi-domain)



Complementary Effects with Extra-Estimation



Analysis on Miscalibration



CANMT can alleviate the over- and under-confidence problems because of using more informative features from NMT models and the effective reconstruction strategy that considers the fidelity of the source sentence.

Code and human evaluation data are publicly available at:

https://github.com/xiaoyi0814/CANMT

Conclusion

Takeaways:
- Competency awareness is an essential capability of NMT (AI models).
- We propose CANMT, which can output both the competency and the translation for a given source sentence. It novelly adopts a reconstruction strategy and leverages its internal information of decoding to estimate competency.
- CANMT has strong performance and robustness in quality estimation, which is the first report that the self-estimation method surpasses supervised ones across translation directions and domains.

Future work:
- Reinforcement learning with the self-estimated competency score as a reward.
- Reranking candidates.

References

[1] Zhang, Tianyi et al. "BERTScore: Evaluating Text Generation with BERT." (ICLR2020)
[2] Sellam, Thibault et al. "BLEURT: Learning Robust Metrics for Text Generation." (ACL2020)
[3] Rei, Ricardo et al. "COMET: A Neural Framework for MT Evaluation." (EMNLP2020)
[4] Fomicheva, M. et al. "Unsupervised Quality Estimation for Neural Machine Translation." (ACL2020)