

LETTER

Exploiting Multi-Level Data Uncertainty for Japanese-Chinese Neural Machine Translation

Ze Zhong LI^{†a)}, Jianjun MA^{††}, *Nonmembers*, and Fuji REN^{†††b)}, *Fellow*

SUMMARY The performance of Neural Machine Translation (NMT) heavily depends on the severity of data uncertainty existing in the training examples. In terms of its causes, data uncertainty can be categorized into intrinsic and extrinsic uncertainty, both of which can increase the learning difficulty in NMT, and lead to degradation in translation performance. To cope with this challenge, we propose a simple yet effective method to estimate the data uncertainty and incorporate it into the adaptive training of NMT, which can mitigate the hurt brought by the uncertain data. Our method consists of two modules: 1) we propose a mBERT based model for estimating token-level and sentence-level data uncertainties jointly, which is trained with multi-task learning; 2) we propose heterogeneous ways of incorporating the derived multi-level uncertainties into the NMT training by using soft-masked embedding and weighted loss. Extensive experiments on Japanese↔Chinese translation show that our proposed methods substantially outperform the strong baselines in terms of BLEU scores, and verify the effectiveness of modeling data uncertainty in NMT.

key words: data uncertainty, adaptive training, Japanese-Chinese NMT

1. Introduction

NMT is generally trained on parallel data in an end-to-end manner, whose performance heavily depends on the degree of data uncertainty in the training examples. A typical uncertainty is noise, which refers to the manual mistakes and semantic divergences occurring in the procedure of data collection. Prior work [1] has found that NMT is more vulnerable to the noise than statistical machine translation. Besides the noise (which can be categorized as extrinsic uncertainty), another type of data uncertainty is referred to as intrinsic uncertainty, which is caused by the one-to-many nature of the translation task, i.e. there exists multiple ways to express the same meaning [2]. Obviously, both of these uncertainties can increase the learning difficulty in NMT, and lead to degradation in translation performance. Therefore, it's a critical challenge for NMT to deal with the impact of data uncertainty.

There already exist several approaches to address the data uncertainty issue. The most direct approach is to calcu-

late a confidence score for each training example and then filter the uncertain examples with low scores [3]. Apparently, it's inappropriate to abandon the entire training examples when the uncertainty only occurs on very few tokens, thus a more sophisticated and fine-grained process is necessary. Gu et al. [4] propose a token-level adaptive training for NMT, in which the most frequent tokens and extremely rare tokens are assigned smaller weights in objectives, that is, these tokens are treated as suspected noise. However, Xu et al. [5] argue that only using token frequency is insufficient, so they introduce bilingual mutual information (BMI). Similarly, Zhou et al. [6] propose cross entropy and perplexity to measure uncertainty. In essence, all these approaches use heuristic statistical metrics, which fail to make much of the contextual information and may lead to inaccurate estimation for data uncertainty. Briakou et al. [7] use mBERT [8] to estimate the fine-grained semantic divergences, which takes full advantage of the amazing contextual capabilities in mBERT, and provides a more principled way for uncertainty estimation. However, the estimation relies on external tools such as WordNet and POS tagger, which limits its application beyond English.

To address these problems, we propose a simple yet effective framework to estimate the multi-level data uncertainties and incorporate them into NMT training with a heterogeneous manner. The framework has two characteristics: 1) we propose a mBERT-based model for data uncertainty estimation using multi-task learning, which can obtain the token-level and sentence-level uncertainties simultaneously; meanwhile, we create synthetic data without human intervention or requiring external language-specific tools, which makes it more universal across various languages. 2) we propose heterogeneous ways of incorporating data uncertainties into NMT via an adaptive training approach, which can take full advantage of the multi-level data uncertainties.

We experiment Japanese↔Chinese translation tasks on the JPO Patent Corpus. Extensive experiments show that our proposed methods substantially outperform the strong baselines in terms of BLEU scores, and verify the effectiveness of modeling multi-level data uncertainties for NMT.

2. Background

During training, NMT is usually optimized on each example (x, y) by minimizing a cross entropy loss:

Manuscript received June 24, 2024.

Manuscript revised October 11, 2024.

Manuscript publicized November 19, 2024.

[†]Faculty of Computer Science, Zhejiang University of Water Resources and Electric Power, China.

^{††}Faculty of Foreign Languages, Dalian University of Technology, China.

^{†††}College of Computer Science and Engineering, and also with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, China.

a) E-mail: lizezhonglaile@163.com

b) E-mail: renfuji@uestc.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2024EDL8059

$$\mathcal{L}(x, y) = - \sum_{j=1}^{|y|} \log p(y_j | x, y_{<j}, \theta) \quad (1)$$

Since the training is performed on the entire training dataset $\mathcal{D}$, which is composed of multiple examples, the total loss is written as a summation of Eq. (1):

$$\mathcal{L}_{total} = \sum_{(x, y) \in \mathcal{D}} \mathcal{L}(x, y) \quad (2)$$

3. Methods

Here we first introduce the data uncertainty estimation in Sect. 3.1, followed by the adaptive training in Sect. 3.2.

3.1 Multi-Level Data Uncertainty Estimation

Synthetic data generation: Currently there is no data annotated with information about which token is uncertain or not, and the cost is high to annotate supervised data manually. To generate such data automatically, we first extract a seed dataset in which each example is a sentence pair with perfect parallelism; then, we get synthetic examples by adding perturbations to the seed examples, and obtain token labels and sentence-level scores without human intervention.

The seed dataset is extracted by selecting top- N examples from the full training data ranked by the cross-lingual sentence similarity. To compute this similarity, we utilize the source and target sentence representations generated by mBERT. Note that, we use two tricks for computing the similarity accurately: 1) we continue to pretrain the mBERT using masked language model (MLM) and translation language model (TLM) objectives on the full training data, as proposed in [9], which can alleviate the issue of curse of multilinguality and strengthen the bilingual representation alignment; 2) instead of directly using the output embedding of “[CLS]” in mBERT as the sentence representation, we adopt the language-agnostic sentence representation proposed in [10], which can avoid the bias generated by the language-specific information in the sentence encoders.

For each seed example (x, y) , we generate two synthetic examples (x', y) and (x, y') in the same way, where x' (y') is the “corrupted” version of the original sentence x (y). Take generating (x', y) for example. We select a random number k of tokens, and perform the following token substitution operation for each of them: with probability 0.8, substitute it with another token randomly chosen from its nearest 10 neighbours in the word embedding space; with probability 0.2, substitute it with an arbitrary token randomly chosen from the entire vocabulary. The motivation behind this corruption scheme is that, we expect the substitution with its similar tokens can mimic the intrinsic uncertainty and the arbitrary token substitution can mimic the extrinsic uncertainty. Next, we obtain token labels and sentence scores for the two synthetic examples using simple heuristics: the positions of substituted tokens x'_i are labeled with 1, otherwise 0; the sentence score is assigned with a real value of

$1 - (1 - k/L)^\alpha$, like a “soft” label for the sentence uncertainty, where L is the length of the corrupted sentence, and α is a scaling factor. To this end, we complete the generation of synthetic data, which is comprised of $2 * N$ synthetic examples, alongside annotated token labels and sentence scores.

Modeling: Inspired by the recent progress in quality estimation for machine translation [11], we propose a mBERT based model to measure the token- and sentence-level data uncertainties jointly. In concrete, we adopt mBERT as the backbone model and fine-tune it on the above synthetic data. We frame token-level uncertainty estimation as a sequential binary classification task, and sentence-level uncertainty estimation as a regression task. We build the classifier and regressor on the output of mBERT. For the token-level uncertainty classifier, the input is the embedding sequence of word pieces in the concatenation of “[CLS], x, [SEP], y”, and the output is a sequence of token labels. For each token’s label l_i , we compute the uncertainty probability using information from both the word embedding and contextualized representation:

$$p_i = P(l_i = 1) = \text{sigmoid}(W_0 f(H_i, E_i)) \quad (3)$$

$$f(H_i, E_i) = \text{ReLU}(W_1 H_i + W_2 E_i + b) \quad (4)$$

$$H_i = \text{pooling}(h_{i,1}^l, h_{i,2}^l, \dots, h_{|i, w_i|}^l) \quad (5)$$

$$E_i = \text{pooling}(e_{i,1}, e_{i,2}, \dots, e_{|i, w_i|}) \quad (6)$$

where W_0, W_1, W_2, b are trainable parameters; $e_{i,k}$ is embedding for the k -th word piece in token w_i , which is fed into mBERT and outputs its contextualized representation $h_{i,k}^l$ in the l -th layer. Since mBERT works on word pieces rather than tokens (here w_i is divided into $|w_i|$ pieces after applying WordPiece [12]), we convert the word piece representations into token-level representations E_i and H_i by performing a pooling operation on $e_{i,1}, e_{i,2}, \dots, e_{|i, w_i|}$ and $h_{i,1}^l, h_{i,2}^l, \dots, h_{|i, w_i|}^l$ respectively.

As for the sentence-level uncertainty, the contextualized representation of “[CLS]” (noted as h_{cls}^l) is fed into a feed-forward network to approximate the sentence score, that is,

$$q(x, y) = \text{sigmoid}(U h_{cls}^l + b_1) \quad (7)$$

where U and b_1 are learnable parameters.

Training: For the token-level training, we use cross entropy as the objective loss. For the sentence-level, we use square loss as the objective. Since the data uncertainty estimation for token- and sentence-level are highly related, we adopt a multi-task learning framework, i.e. the final objective is a linear summation of the above two losses.

We initialize the model with the fine-tuned mBERT in 3.1.1. During training on the synthetic data, we froze all the parameters in mBERT, and only fine-tune the parameters in the feed-forward network for classification and regression. Following [13], for each training example, we feed the concatenation of its source and target sentence twice in different orders instead of resetting the positions of target sentences.

Prediction: For prediction, we also input each example twice with different concatenation orders, and averaging their

predictions as the final prediction, which contributes to a more robust estimation.

3.2 Uncertainty Based Adaptive Training

We propose an adaptive training approach for NMT, which can mitigate the hurt of uncertainty to translation performance. Since various data uncertainties impact different components of NMT, we design heterogeneous ways of incorporating them in the training process.

Source Token Uncertainty: For the source token uncertainty, to mitigate its impact on the encoder, we borrow the idea of soft-masking mechanism which is originally designed for spelling-error correction [14]. Formally, suppose x_i is the i -th token in the source sentence, and $\mathbf{x}_i$ is its word embedding (in bold), we input the following soft-masked embedding into the NMT encoder instead of the original embedding $\mathbf{x}_i$:

$$\mathbf{x}'_i = (1 - p_i) \cdot \mathbf{x}_i + p_i \cdot \mathbf{e}_{mask} \quad (8)$$

where p_i is the uncertainty probability of token x_i , $\mathbf{e}_{mask}$ is the mask embedding. Intuitively, when the probability p_i is higher (i.e. being more uncertain), the final input will mask the original input embedding to a larger extent, thus alleviating the hurt of uncertain tokens to learning for NMT.

It's worth noting that there are no such uncertainty probabilities at inference time, so all the input source tokens are assigned with $p_i = 0$, that is, we treat all source tokens being perfect without uncertainties for inference. However, this trick would cause discrepancy between training and inference for NMT, and deteriorate the translation performance. To amend this discrepancy, we only mask the most uncertain tokens (i.e. tokens satisfying p_i bigger than a threshold σ) in training, and set other tokens without masking mechanism.

Target Token Uncertainty: Following [4], [5], we integrate the target token uncertainty into the adaptive training objective by modifying Eq. (1):

$$\mathcal{L}(x, y) = - \sum_{j=1}^{|y|} (1 - p_j) \cdot \log p(y_j | x, y_{<j}, \theta) \quad (9)$$

where p_j is the target token uncertainty for y_j . This formula reflects the intuition that tokens with more uncertainties are assigned with lower weights and have weaker influence on the loss. Instead of the weights based on simple statistical metrics like token frequency and mutual information, we use the estimated token uncertainty as weights, which are derived from mBERT in an supervised way. Due to the powerful capability of contextualization and cross-lingual representation, we expect a more robust estimation for data uncertainty which would boost the final translation performance.

Sentence Uncertainty: Similarly, it's easy to incorporate the sentence-level uncertainty into the adaptive training objective by modifying Eq. (2):

$$\mathcal{L}_{total} = \sum_{(x, y) \in \mathcal{D}} (1 - q(x, y)) \cdot \mathcal{L}(x, y) \quad (10)$$

Table 1 Results of uncertainty estimation

Pooling Method	Precision	Recall	F1	RMSE
first word piece	85.10	64.47	73.36	0.110
avg	82.96	66.03	73.53	0.106
max	85.82	72.63	78.67	0.104
sum	81.46	68.55	74.45	0.109

Serving as a weight (inversely proportional to the degree of uncertainty) assigned to each instance, it is expected to be beneficial for mitigating the impact of learning parameters on sentence pairs with less parallelism.

4. Experiments

4.1 Data and Setup

We carry out the Japanese $\leftrightarrow$ Chinese experiments on JPO corpus, which contains 1M, 2K and 10204 Japanese-Chinese sentences for training, validation and test set respectively. The test set is further grouped into four subsets, i.e. test-n1 (2K), test-n2 (3K), test-n3 (204) and test-n4 (5K). For the data pre-processing, we use MeCab and Jieba to perform segmentation for them respectively. We filter the training examples more than 120 words either in the Japanese or Chinese sentence, and finally 973K examples left. We conduct a joint BPE encoding with a vocabulary of 32K tokens.

For uncertainty estimation, $N = 10K$, that is, we extract 10K seed examples from the parallel data and generate 20K synthetic examples. We split the synthetic examples into training, validation and test sets with the ratio of 0.8:0.1:0.1. k is an integer uniformly sampled from the scope of $[0, 5]$; set $\alpha = 3$ for the scaling factor. For the token classification task, we use the 8-th layer representation, and for sentence uncertainty, we use the 12-th layer. Other layer settings show inferior performances on the test set.

For translation task, we use 6 encoder and decoder layers, and each layer employs a multi-head attention with 8 heads; the dimension of word embedding, and FFN layer is 2048; For inference, we choose the averaged model of the last 5 epoch checkpoints for beam search decoding with a beam width of 5.

4.2 Results of Uncertainty Estimation

As shown in Table 1, we test the uncertainty estimation at the token- and sentence-level. For the token level evaluation, we measure it using precision, recall and F1-score for the corrupted tokens. For the sentence-level, we use RMSE as the metric. A critical component in the estimation is the pooling operation that transforms the word piece representations into token-level representations. Therefore, we compare 4 types of pooling mechanisms in Table 1, and find that max pooling performs the best.

4.3 Results of Machine Translation

Table 2 shows the results of J $\rightarrow$ C task on the four test sets, and *All* denotes the union of them. We also experiment

Table 2 BLEU scores evaluated on test set (Japanese-to-Chinese)

Method	test-n1	test-n2	test-n3	test-n4	All
Transformer	41.53	42.68	26.15	60.92	48.29
Freq-Exp [4]	41.10	42.03	25.37	61.21	48.06
Freq-K2 [4]	40.63	42.05	24.21	60.82	47.82
BMI [5]	41.65	42.83	27.16	61.45	48.64
CBMI [15]	41.75	42.61	26.83	61.42	48.61
Proposed	42.52[†]	43.70[†]	26.32	61.67[†]	49.15[†]

Table 3 BLEU scores evaluated on test set (Chinese-to-Japanese)

Method	test-n1	test-n2	test-n3	test-n4	All
Transformer	49.54	48.51	22.21	66.13	54.00
Freq-Exp [4]	49.46	48.16	21.70	66.12	53.83
Freq-K2 [4]	49.12	47.53	20.98	65.64	53.33
BMI [5]	49.70	48.87	22.44	66.29	54.24
CBMI [15]	50.18	48.91	22.42	66.34	54.30
Proposed	50.46[†]	49.63[†]	23.39[†]	66.63	54.77[†]

Table 4 Ablation experiments for J→C and C→J tasks

No.	Method	J→C	C→J
0	Proposed method	49.15	54.77
1	w/o src token uncertainty	48.96(-0.19)	54.47(-0.30)
2	w/o tgt token uncertainty	48.69(-0.46)	54.54(-0.23)
3	w/o sentence uncertainty	48.73(-0.42)	54.46(-0.31)

several existing works. Among them, Freq-Exp and Freq-K2 show slight weaker results over the baseline, which can be attributed to the poor uncertainty measure provided by using token frequency only. Both BMI and CBMI show stronger performance than the baseline. In contrast, our proposed methods bring consistent improvements over the baseline on all the test sets, and obtain a large improvement of 0.86 BLEU points (48.29 vs. 49.15) on the overall.

Similar trends can be observed from the results of C→J task, as shown in Table 3. Compared to the baseline, our proposed methods achieve an improvement of 0.77 BLEU points (54 vs. 54.77) on the entire test data. From all these improvements, we can confirm the effectiveness of our methods for estimating multi-level data uncertainties and incorporating them into the adaptive training in NMT.

4.4 Ablation Study

As shown in Table 4, we conduct ablation study to investigate the contribution of different data uncertainties in the proposed method. We remove each uncertainty module alternately to check the performance drop on the entire test set. The first observation is that removing each of them leads to performance deteriorated, which indicates the effectiveness of such uncertainties. Another observation is that, the source token uncertainty seems only useful for the C→J direction. One possible explanation is that the quality of Chinese sentences in the JPO corpus is worse than that of the Japanese ones. Therefore, it’s more indispensable to utilize the data uncertainty on the Chinese side.

5. Conclusion

This paper presents a simple yet effective method for model-

ing the data uncertainties in NMT. We first propose a mBERT based model for estimating token- and sentence-level data uncertainties jointly; then, we propose heterogeneous ways of incorporating these derived uncertainties into NMT training. Extensive experiments on Japanese↔Chinese translation show that the proposed approach substantially outperforms the strong baseline, and verify the effectiveness of modeling data uncertainties for NMT. As our study focuses on Japanese-Chinese NMT currently, future work could include validating this approach on NMT for other languages and other natural language generation tasks.

Acknowledgments

This work was supported by Scientific Research Fund of Zhejiang Provincial Education Department (No. Y202456743)

References

- [1] H. Khayrallah and P. Koehn, “On the impact of various types of noise on neural machine translation,” *Proc. NMT@ACL*, Melbourne, Australia, pp.74–83, July 2018.
- [2] M. Ott, M. Auli, and et al, “Analyzing uncertainty in neural machine translation,” *Proc. ICML*, Stockholm, Sweden, pp.3953–3962, July 2018.
- [3] Y. Vyas, X. Niu, and M. Carpuat, “Identifying semantic divergences in parallel text without annotations,” *Proc. NAACL-HLT*, New Orleans, Louisiana, USA, pp.1503–1515, June 2018.
- [4] S. Gu, J. Zhang, F. Meng, Y. Feng, W. Xie, J. Zhou, and D. Yu, “Token-level adaptive training for neural machine translation,” *Proc. EMNLP*, Online, pp.1035–1046, Nov. 2020.
- [5] Y. Xu, Y. Liu, F. Meng, J. Zhang, J. Xu, and J. Zhou, “Bilingual mutual information based adaptive training for neural machine translation,” *Proc. ACL/IJCNLP*, Online, pp.511–516, Aug. 2021.
- [6] Y. Zhou, B. Yang, D.F. Wong, Y. Wan, and L.S. Chao, “Uncertainty-aware curriculum learning for neural machine translation,” *Proc. ACL*, Online, pp.6934–6944, July 2020.
- [7] E. Briakou and M. Carpuat, “Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation,” *Proc. ACL/IJCNLP*, Online, pp.7236–7249, Aug. 2021.
- [8] J. Devlin, M.W. Chang, and et al, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *Proc. NAACL-HLT*, Minneapolis, Minnesota, pp.4171–4186, July 2019.
- [9] A. Conneau and G. Lample, “Cross-lingual language model pretraining,” *Proc. NIPS*, Minneapolis, Minnesota, pp.7057–7067, 2019.
- [10] N. Tiyajamorn, T. Kajiwara, Y. Arase, and M. Onizuka, “Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation,” *Proc. EMNLP*, Online, pp.7764–7774, Nov. 2021.
- [11] Y. Lu, J. Zeng, J. Zhang, S. Wu, and M. Li, “Learning confidence for transformer-based neural machine translation,” *Proc. ACL*, Dublin, Ireland, pp.2353–2364, 2022.
- [12] M. Schuster and K. Nakajima, “Japanese and korean voice search,” *Proc. ICASSP*, Kyoto, Japan, pp.5149–5152, March 2012.
- [13] Z.-Y. Dou and G. Neubig, “Word alignment by fine-tuning embeddings on parallel corpora,” *Proc. EACL*, Online, pp.2112–2128, 2021.
- [14] S. Zhang, H. Huang, J. Liu, and H. Li, “Spelling error correction with soft-masked BERT,” *Proc. ACL*, Online, pp.882–890, 2020.
- [15] S. Zhang, Y. Liu, F. Meng, Y. Chen, J. Xu, J. Liu, and J. Zhou, “Conditional bilingual mutual information based adaptive training for neural machine translation,” *Proc. ACL*, Online, pp.2377–2389, May 2022.