Token-Level Self-Evolution Training for Sequence-to-Sequence Learning

Keqin Peng[1]*, **Liang Ding**[2]*, **Qihuang Zhong**[3]
Yuanxin Ouyang[1]†, **Wenge Rong**[1], **Zhang Xiong**[1], **Dacheng Tao**[4]
[1]Beihang University [2]Zhejiang University [3]Wuhan University [4]The University of Sydney
{keqin.peng,oyyx,w.rong,xiongz}@buaa.edu.cn
zhongqihuang@whu.edu.cn, {liangding.liam,dacheng.tao}@gmail.com

Abstract

Adaptive training approaches, widely used in sequence-to-sequence models, commonly reweigh the losses of different target tokens based on priors, e.g. word frequency. However, most of them do not consider the variation of learning difficulty in different training steps, and overly emphasize the learning of difficult one-hot labels, making the learning deterministic and sub-optimal. In response, we present *Token-Level Self-Evolution Training* (SE), a simple and effective dynamic training method to fully and wisely exploit the knowledge from data. SE focuses on dynamically learning the under-explored tokens for each forward pass and adaptively regularizes the training by introducing a novel token-specific label smoothing approach. Empirically, SE yields consistent and significant improvements in three tasks, i.e. machine translation, summarization, and grammatical error correction. Encouragingly, we achieve averaging +0.93 BLEU improvement on three machine translation tasks. Analyses confirm that, besides improving lexical accuracy, SE enhances generation diversity and model generalization.

1 Introduction

Sequence-to-sequence learning (Seq2Seq) with neural networks (Sutskever et al., 2014) has advanced the state-of-the-art in various NLP tasks, e.g. translation (Bahdanau et al., 2015; Vaswani et al., 2017), summarization (Cheng and Lapata, 2016), and grammatical error correction (Yuan and Briscoe, 2016). Generally, Seq2Seq models are trained with the cross-entropy loss, which equally weighs the training losses of different target tokens.

However, due to the token imbalance nature (Piantadosi, 2014) and the truth that different tokens contribute differently to the sentence meaning (Church and Hanks, 1990; Chen et al., 2020),

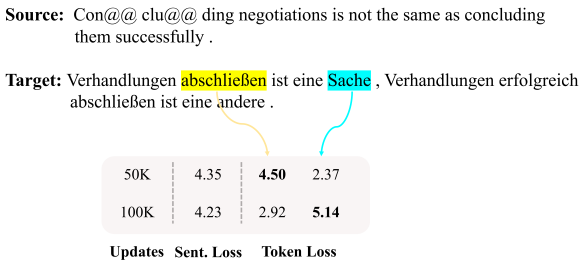

Figure 1: An example to illustrate the **changing token difficulties in different training steps** in WMT'14 En-De. The token "abschließen/ Sache" is hard/ easy to learn at 50K while the trend is totally reversed at 100K.

several works are developed to reweigh the token-level training loss according to explicit (e.g. frequency) or implicit (uncertainty estimated by off-the-shelf language models) priors (Gu et al., 2020; Xu et al., 2021; Zhang et al., 2022a). For example, Gu et al. (2020) proposed two heuristic criteria based on word frequency to encourage the model to learn from larger-weight low-frequency tokens. Zhang et al. (2022a) introduce target-context-aware metric based on an additional target-side language model to adjust the weight of each target token.

Despite some success, there are still limitations in these adaptive training approaches. First, most of them predetermine the difficult tokens and fix such prior to guiding the training. However, in our preliminary study, we find the hard-to-learn tokens are dynamically changing during training, rather than statically fixed. As shown in Figure 1, as the training progress goes, although the sentence-level loss is nicely converging, the difficult token is changing from "*abschließen*" to "*Sache*" in terms of the token-level loss. Second, these adaptive training methods overly emphasize fitting the difficult tokens' one-hot labels by reweighing the loss, which empirically may cause overfitting and limit the generalization (Norouzi et al., 2016; Szegedy et al., 2016; Xiao et al., 2019; Miao et al., 2021). Also, a more recent study (Zhai et al., 2023) provides

*Keqin and Liang contributed equally.
†Corresponding Author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 841–850

theoretical evidence to support that reweighting is not that effective to improve the generalization.

Correspondingly, we design a simple and effective *Token-Level Self-Evolution Training* (SE) strategy to encourage Seq2Seq models to learn from difficult words that are dynamically selected by the model itself. Specifically, SE contains two stages: ❶*self-questioning* and ❷*self-evolution training*. In the first stage, the Seq2Seq models dynamically select the hard-to-learn tokens based on the token-level losses, then we encourage the Seq2Seq models to learn from them in the second stage, where, rather than adopting reweighing, we introduce a novel *token-specific label smoothing* approach to generate easily digestible soft label, which considers both the ground truth and model's prediction.

Experiments across tasks, language pairs, data scales, and model sizes show that SE consistently and significantly outperforms both the vanilla Seq2Seq model and the re-implemented advanced baselines. Analyses confirm that besides improved lexical accuracy, SE generates diverse and human-like generations with better model generalization.

2 Methodology

Preliminary Sequence-to-sequence (Seq2Seq) learning aims to maximize the cross-entropy (CE) loss of the log-likelihood of each target word in $\mathbf{y} = \{y_1, \ldots, y_N\}$, conditioned on source $\mathbf{x}$, where the optimization treats all tokens equally:

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{j=1}^{N} \log p(y_j|\mathbf{y}_{<j}, \mathbf{x}; \theta) \quad (1)$$

However, due to the different learning difficulties of each token, it is sub-optimal to treat all tokens equally (Gu et al., 2020). To address this limitation, a series of token-level adaptive training objectives were adopted to re-weight the losses of different target tokens (Xu et al., 2021; Zhang et al., 2022a). The common goal of these methods is to facilitate the model training by fully exploiting the informative but underexplored tokens.

However, our preliminary study shows that the hard tokens are dynamically changing (see Figure 1) in different training steps (or model structures), thus it is sub-optimal to employ static token priors (e.g. frequency) during training. Also, recent studies (Zhai et al., 2023) in the ML community theoretically show that reweighting is not that effective to improve the generalization. Based on the above evidence, we present the self-evolution learning (SE) mechanism to encourage the model to adaptively and wisely learn from the informative yet under-explored tokens dynamically determined by the model itself (Stage❶ in §2.1), with an easy-to-learn label distribution (Stage❷ in §2.1). A similar work to ours is Hahn and Choi (2019). However, their method mainly considers the situation where the predicted answer is incorrect but close to the golden answer, while our method focuses on all dynamic hard tokens.

2.1 Token-Level Self-Evolution Learning

❶ Self-questioning Stage. The goal is to select the hard-to-learn tokens that are questioned by the Seq2Seq model itself during training dynamics. Previously, these difficult tokens are predetermined by external models or specific statistical metrics. However, inspired by the finding of dynamic change of difficult tokens during the training stage as shown in Figure 1 and the finding that the trained model contains useful information (Li and Lu, 2021), e.g. synonym, we propose to straightforwardly leverage the behavior of the model to dynamically select target tokens. In practice, we first calculate the token-level CE loss, denoted as $\{l_1, l_2, ..., l_n\}$, for each token for each forward pass. Then we set a loss threshold Γ and select the tokens whose losses exceed Γ as the target tokens, *i.e.*, $D = \{t_i | l_i > \Gamma\}$ where $i \in N = \{1, 2, ..., n\}$.

❷ Self-evolution Training Stage. After selecting the difficult tokens, we encourage the model to carefully learn from them. Given the theoretical shortage (Zhai et al., 2023) and potentially caused overfitting or overconfidence problem (Miao et al., 2021) of reweighting and deliberately learning from difficult tokens, we propose to strengthen the learning from these tokens with a newly designed *Token-specific Label Smoothing* (TLS) approach. Specifically, motivated by the effect of label smoothing (LS) regularization (Szegedy et al., 2016), we combine the ground truth p_i and the model's prediction $\hat{p}_i$ to form a new soft label $\widetilde{p}_i$ for the i-th token. Then we use $\widetilde{p}$ to guide the difficult tokens D, while leaving label-smoothing CE loss for the other tokens. It is worth noting that we also apply the traditional label smoothing technique to $\hat{p}_i$ to activate the information in the predicted distribution. Analogous to human learning, it is often easier for humans to grasp new things described by their familiar knowledge (Reder et al., 2016),

Model	WMT16 En→Ro	WMT14 En→De	WMT14 En→Fr
Transformer (Vaswani et al., 2017)	35.11	27.08	40.65
+ Freq-Exponential (Gu et al., 2020)	35.86 (+0.75)	27.60 (+0.52)	41.05 (+0.40)
+ Freq-Chi-Square (Gu et al., 2020)	35.74 (+0.63)	27.51 (+0.43)	40.99 (+0.34)
+ D2GPo (Li et al., 2020)	35.89 (+0.78)	27.66 (+0.58)	41.05 (+0.40)
+ BMI-adaptive (Xu et al., 2021)	35.89 (+0.78)	27.65 (+0.57)	41.10 (+0.45)
+ MixCrossEntropy (Li and Lu, 2021)	35.88 (+0.74)	27.61 (+0.53)	41.07 (+0.42)
+ CBMI-adaptive (Zhang et al., 2022a)	35.90 (+0.79)	27.69 (+0.61)	41.13 (+0.48)
+ SPL (Wan et al., 2020)	35.92 (+0.81)	27.88 (+0.80)	41.30 (+0.65)
+ Self-Evolution (ours)	**36.02 (+0.91)**†	**28.02 (+0.94)**†	**41.60 (+0.95)**†

Table 1: **BLEU scores (%) on three translation tasks spanning different data scales**, i.e. 0.6M, 4.5M, 36M. "†" indicates a statistically significant difference from the powerful Transformer baseline ($p < 0.05$).

	Ro-En	XSUM			GEC		
	BLEU	*RG-1*	*RG-2*	*RG-L*	*Prec.*	*Recall*	$F_{0.5}$
Baseline	37.3	43.2	19.8	34.0	**59.1**	39.8	53.9
+ SE	**37.7**†	**43.8**	**20.4**	**34.7**†	58.9	**46.2**	**55.8**†

Table 2: **Performance on more tasks** including translation, summarization, and grammar error correction, upon larger model BART (Lewis et al., 2020).

therefore the new soft label fused both accurate ground truth and model's self-distribution is easily digestible. Mathematically, for difficult tokens t_i, $\widetilde{p_i}$ is formulated as:

$$\widetilde{p_i} = (p_i + \hat{p_i})/2. \quad (2)$$

Then we calculate the losses of difficult tokens and the others, and combine the two losses:

$$L = -(\sum_i \widetilde{p_i} \cdot log(\hat{p_i}) + \sum_j p_j \cdot log(\hat{p_j})), \quad (3)$$

where $i \in D$ and $j \in N \setminus D$.

3 Evaluation

Machine Translation on three widely-used benchmarks (Ding et al., 2020, 2021c, 2022): small-scale WMT16 English-Romanian (En-Ro; 0.6M), medium-scale WMT14 English-German (En-De; 4.5M), and large-scale WMT14 English-French (En-Fr; 36.0M). We implement the baselines and our approach under Transformer-base settings. We follow the previous adaptive training approach (Gu et al., 2020) to pretrain with the cross-entropy loss with N steps, and further finetune the same steps with different adaptive training objectives, including **Freq-Exponential** (Gu et al., 2020), **Freq-Chi-Square** (Gu et al., 2020), **D2GPo** (Li et al., 2020), **BMI-adaptive** (Xu et al., 2021), **MixCrossEntropy** (Li and Lu, 2021), **CBMI-adaptive** (Zhang et al., 2022a), and **SPL** (Wan et al., 2020). For N, we adopt 100K and 30K for larger datasets, e.g. En-De and En-Fr, and small dataset, i.e. En-Ro, respectively. We empirically adopt 32K tokens per batch for large datasets, the learning rate warms up to 1e-7 for 10K steps, and then decays 90K, while for small dataset En-Ro, The learning rate warms up to 1e-7 for 4K steps, and then decays 26K steps. All the experiments are conducted on 4 NVIDIA Tesla A100 GPUs. The SacreBLEU (Post, 2018) was used for evaluation. Besides translation, we also follow previous works (Liu et al., 2021b; Zhong et al., 2022; Zhang et al., 2022b) to validate the universality of our method on more sequence-to-sequence learning tasks, e.g., summarization and grammatical error correction.

Text Summarization on XSUM corpus (0.2M). We follow fairseq (Ott et al., 2019) to preprocess the data and train the model, then finetune them for the same steps. We evaluated with the ROUGE (Lin, 2004), i.e. R-1, R-2, and R-L.

Grammatical Error Correction on CoNLL14 (1.4M). We follow Chollampatt and Ng (2018) to preprocess the data and train the model, then finetune them for the same steps. The MaxMatch (M^2) scores (Dahlmeier and Ng, 2012) were used for evaluation with precision, recall, and $F_{0.5}$ values.

3.1 Main Results

SE brings gains across language pairs and scales. Results on machine translation across different data sizes ranging from 0.6M to 36M in Table 1 show that our SE-equipped Transformer "+ Self-Evolution (ours)" 1) considerably improves the performance by averaging +0.92 BLEU points; 2) out-

	Valid Loss Scale			
	0-1	*1-2*	*2-3*	*>3*
Transformer	63.3	10.5	6.7	19.5
+ SE	**65.6**	9.5	5.8	19.1

Table 3: The **token distribution (%) on different loss scales**. Shadowed areas mean accurate token prediction estimated with lower cross-entropy loss, i.e. "0-1".

Method	**WMT22 De⇒En**			
	BLEU	Δ	*COMET*	Δ
Transformer	29.98	-	45.1	
+SE	**30.38**	+0.4	**46.3**	+1.2

Table 4: **Performance on extremely large dataset** WMT22 De-En (236M).

performs previous competitive method "+ CBMI-adaptive" by up to +0.47 BLEU points on large dataset WMT14 En-Fr. These results demonstrate the effectiveness and universality of our SE.

SE brings gains across tasks and backbone sizes. Table 2 lists the performance on more tasks, including translation, summarization, and grammar error correction, upon large pretrained backbone - BART (Lewis et al., 2020), which has above 600M parameters. Compared to a stronger baseline, our SE significantly and incrementally improves the generation quality in all tasks, i.e. +0.4 BLEU, + 0.7 RG-L, and + 1.9 $F_{0.5}$, respectively, showing our SE is robustly applicable to general scenarios.

SE works well on extremely large dataset. To further verify the effectiveness of SE on extremely large dataset, we conducted an experiment on WMT22 De-En processed by Zan et al. (2022b), which contains 236M training examples. The results in Table 4 show that our method can achieve +0.4 and +1.2 improvement in BLEU and COMET respectively, which proves that our SE also works on extremely large datasets.

3.2 Analysis

We provide some insights to better understand the effectiveness of our approach. The ablation of important modules and parameters is in Appendix A.

SE learns better token representation. To verify whether our method helps learn better tokens representation, we conduct analysis on WMT14 En-De from learning loss and fine-grained generation

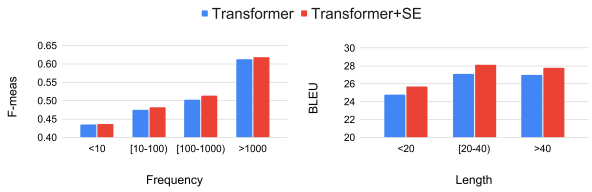

Figure 2: **Fine-grained translation quality** across word frequencies and sentence lengths.

perspectives, respectively.

First, we count the token ratios distributed in different cross-entropy loss scales in Table 3 following Zan et al. (2022a). Cross-entropy is a good indicator to quantify the distance between the predicted distribution and the ground truth in the valid dataset, and a lower value means a more similar distribution. As shown, our method improves the low-loss token ratios by +2.3%, indicating SE helps the model ***learn better token representations by reducing the token uncertainty***. In addition, we follow Ding et al. (2021a); Liu et al. (2021a) to break the translation down into different granularities and measure their fined-grained performance. In particular, we calculate[1] the F-measure of words by different frequency buckets and BLEU scores of buckets of different lengths in Figure 2. We see SE achieves better performance in all frequencies and sentence buckets, demonstrating our method can ***improve the performance of different granularities***.

SE encourages diverse generations. Lacking generation diversity is a notorious problem for Seq2Seq learning tasks (Sun et al., 2020; Lin et al., 2022). Benefiting from better exploring the model's prediction with corrected soft labels, SE is expected to improve generation diversity. We follow Wang et al. (2022) to examine this by analyzing the performance in an additional multiple-reference test of WMT'14 En-De (Ott et al., 2018). We choose additional references for each of the 500 test sentences taken from the original test. Table 5 shows SE consistently outperforms the baseline with the average improvement being 0.9/1.0 BLEU, which indicates that ***our SE can effectively generate diverse results.***

SE enhances model generalization. Benefiting from better hard token exploration, SE-equipped Transformers are expected to own better generalizations. We examine it by testing on domain shift

[1]Using compare-mt (Neubig et al., 2019).

Ref.	Avg.		Top	
	Transformer	+SE	Transformer	+SE
#1	42.5	**43.7 (+1.2)**	44.9	**45.7 (+0.8)**
#2	28.6	**29.3 (+0.7)**	30.2	**31.2 (+1.0)**
#3	31.2	**32.1 (+0.9)**	33.2	**34.4 (+1.2)**
#4	28.1	**28.8 (+0.7)**	29.6	**30.5 (+0.9)**
Mean	32.6	**33.5 (+0.9)**	34.5	**35.5 (+1.0)**

Table 5: **Multi-reference** performance. 'Avg./ Top" means the averaging/ most-matching performance.

Model	Law	Med.	Kor.	Sub.	*Avg.*
Transformer	41.2	30.9	7.4	14.5	**23.5**
+SE	**42.6**†	**32.3**†	**7.8**†	**15.0**†	**24.4**

Table 6: Performance on **domain shift setting**. Models are trained on the news but evaluated on out-of-domain test sets, including law, medicine, koran, and subtitle. "†" indicates statistically significance ($p < 0.05$).

scenarios following Ding et al. (2021b). In particular, we evaluate WMT14 En-De models over four out-of-domain test sets (Müller et al., 2020) in Table 6 and find that SE improves the translation by averaging +0.9 BLEU points, showing a ***better lexical generalization ability***.

SE encourages human-like generations. We design two types of evaluation on WMT14 En-Fr: 1) AUTOMATIC EVALUATION with **COMET** (Rei et al., 2020) and **BLEURT** (Sellam et al., 2020), which have a high-level correlation with human judgments. 2) HUMAN EVALUATION with three near-native French annotators who hold DALF C2 certificate[2]. Specifically, for human evaluation, we randomly sample 50 sentences from the test set to evaluate the translation **adequacy** and **fluency**, scoring 1∼5. For adequacy, 1 represents irrelevant to the source while 5 means semantically equal. For fluency, 1 means unintelligible while 5 means fluent and native. Table 7 shows the automatic and human evaluation results, where we find that ***our SE indeed achieves human-like translation***.

	AUTOMATIC EVAL.		HUMAN EVAL.	
	COMET	*BLEURT*	*Adequacy*	*Fluency*
Transformer	61.6	68.6	4.32	4.58
+ SE	**63.7**	**69.5**	**4.50**	**4.68**

Table 7: **Human evaluation** on WMT14 En-Fr.

4 Conclusion

In this paper, we propose a self-evolution learning mechanism to improve seq2seq learning, by exploiting the informative-yet-underexplored tokens dynamically. SE follows two stages, i.e. self-questioning and self-evolution training, and can be used to evolve any pretrained models with a simple recipe: continue train with SE. We empirically demonstrated the effectiveness and universality of SE on a series of widely-used benchmarks, covering low, medium, high, and extremely-high data volumes.

In the future, besides generation tasks, we would like to verify the effectiveness of SE on language understanding tasks (Wu et al., 2020; Zhong et al., 2023). Also, it will be interesting to design SE-inspired instruction tuning or prompting strategy like Lu et al. (2023) to enhance the performance of large language models, e.g. ChatGPT[3], which after all have already been fully validated on lots of conditional generation tasks (Hendy et al., 2023; Jiao et al., 2023; Peng et al., 2023; Wu et al., 2023).

Limitations

Our work has several potential limitations. First, we determine the threshold Γ by manual selection, which may limit the performance of Seq2Seq models, it will make our work more effective and elegant if we dynamically select the threshold. Second, besides the improvement on three widely used tasks, we believe that there are still other abilities, like code generation, of Seq2Seq models that can be improved by our method, which are not fully explored in this work.

Ethics Statement

We take ethical considerations very seriously and strictly adhere to the ACL Ethics Policy. This paper focuses on effective training for sequence-to-sequence learning. The datasets used in this paper are publicly available and have been widely adopted by researchers. We ensure that the findings and conclusions of this paper are reported accurately and objectively.

Acknowledgement

We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions.

[2]http://www.delfdalf.fr/dalf-c2-en.html

[3]https://chat.openai.com/

References

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

Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020. Content word aware neural machine translation. In *ACL*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *ACL*.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *AAAI*.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *CL*.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *NAACL*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2021a. Progressive multi-granularity training for non-autoregressive translation. In *Findings of ACL*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2021b. Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation. In *ACL*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2021c. Understanding and improving lexical choice in non-autoregressive translation. In *ICLR*.

Liang Ding, Longyue Wang, Shuming Shi, Dacheng Tao, and Zhaopeng Tu. 2022. Redistributing low-frequency words: Making the most of monolingual data in non-autoregressive translation. In *ACL*.

Liang Ding, Longyue Wang, and Dacheng Tao. 2020. Self-attention with cross-lingual position representation. In *ACL*.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *EMNLP*.

Sangchul Hahn and Heeyoul Choi. 2019. Self-knowledge distillation in natural language processing. In *RANLP*.

Amr Hendy, Mohamed Abdelrehim, et al. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Haoran Li and Wei Lu. 2021. Mixed cross entropy loss for neural machine translation. In *ICML*.

Zuchao Li, Rui Wang, et al. 2020. Data-dependent gaussian prior objective for language generation. In *ICLR*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Huan Lin, Baosong Yang, Liang Yao, Dayiheng Liu, Haibo Zhang, Jun Xie, Min Zhang, and Jinsong Su. 2022. Bridging the gap between training and inference: Multi-candidate optimization for diverse neural machine translation. In *Findings of NAACL*.

Xuebo Liu, Longyue Wang, Derek F Wong, Liang Ding, Lidia S Chao, Shuming Shi, and Zhaopeng Tu. 2021a. On the copying behaviors of pre-training for neural machine translation. In *Findings of ACL*.

Xuebo Liu, Longyue Wang, Derek F Wong, Liang Ding, Lidia S Chao, and Zhaopeng Tu. 2021b. Understanding and improving encoder layer fusion in sequence-to-sequence learning. In *ICLR*.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint*.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. In *ACL*.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *AMTA*, Virtual.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *NAACL*.

Mohammad Norouzi, Samy Bengio, Zhifeng Chen, et al. 2016. Reward augmented maximum likelihood for neural structured prediction. In *NeurIPS*.

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

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 Demonstration*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arxiv preprint*.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *WMT*.

Lynne M Reder, Xiaonan L Liu, Alexander Keinath, and Vencislav Popov. 2016. Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. *Psychonomic bulletin & review*.

Ricardo Rei, Craig 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*.

Zewei Sun, Shujian Huang, Hao-Ran Wei, Xinyu Dai, and Jiajun Chen. 2020. Generating diverse translation by manipulating multi-head attention. In *AAAI*.

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

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

Ashish Vaswani, Noam Shazeer, et al. 2017. Attention is all you need. In *NeurIPS*.

Yu Wan, Baosong Yang, et al. 2020. Self-paced learning for neural machine translation. In *EMNLP*.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael R. Lyu. 2022. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *ACL*.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. In *EMNLP*.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint*.

Fengshun Xiao, Yingting Wu, Hai Zhao, Rui Wang, and Shu Jiang. 2019. Dual skew divergence loss for neural machine translation. *CoRR*.

Yangyifan Xu, Yijin Liu, Fandong Meng, Jiajun Zhang, Jinan Xu, and Jie Zhou. 2021. Bilingual mutual information based adaptive training for neural machine translation. In *ACL*.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *NAACL*.

Changtong Zan, Liang Ding, Li Shen, Yu Cao, Weifeng Liu, and Dacheng Tao. 2022a. On the complementarity between pre-training and random-initialization for resource-rich machine translation. In *COLING*.

Changtong Zan, Keqin Peng, Liang Ding, et al. 2022b. Vega-mt: The jd explore academy machine translation system for wmt22. In *WMT*.

Runtian Zhai, Chen Dan, J Zico Kolter, and Pradeep Kumar Ravikumar. 2023. Understanding why generalized reweighting does not improve over ERM. In *ICLR*.

Songming Zhang, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, Jian Liu, and Jie Zhou. 2022a. Conditional bilingual mutual information based adaptive training for neural machine translation. In *ACL*.

Zheng Zhang, Liang Ding, Dazhao Cheng, Xuebo Liu, Min Zhang, and Dacheng Tao. 2022b. Bliss: Robust sequence-to-sequence learning via self-supervised input representation. *arXiv preprint*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2022. E2s2: Encoding-enhanced sequence-to-sequence pretraining for language understanding and generation. *arXiv preprint*.

Qihuang Zhong, Liang Ding, Keqin Peng, Juhua Liu, Bo Du, Li Shen, Yibing Zhan, and Dacheng Tao. 2023. Bag of tricks for effective language model pretraining and downstream adaptation: A case study on glue. *arXiv preprint*.

A Appendix

Parameter Analysis on Γ As stated in §2.1, we use the loss threshold Γ to dynamically select the hard-to-learn tokens. Here, we analyze the influence of different Γ in detail. In practice, we train the Transformer models with different Γ (in {3,4,5,6}) and evaluate the performance of the WMT14 En-De test set. Table 8 lists the performance of different Γ. The results of Table 8 show that ***SE is stable and insensitive to*** Γ ***within a certain range***. Noting that we select $\Gamma = 5$ for all experiment settings based on the results in Table 8.

	Γ=3	Γ=4	Γ=5	Γ=6
BLEU	27.7	27.8	**28.0**	27.8

Table 8: Parameter analysis of Γ on WMT14 En-De.

Ablation Study

Metric. In this work, we use the loss-based metric to dynamically select the hard-to-learn tokens. To validate the effectiveness of the metric, we use a simple adaptive training method ("+ ADD") that adds 1 to the weighting term of loss of the hard-to-learn tokens. The results on WMT16 En-Ro are shown in Table 9, the simple Add method can achieve +0.3 BLEU improvement compared to the baseline model, which proves that ***our proposed self-questioning stage indeed mines informative difficult tokens.*** Also, we can observe that learning these dynamic difficult tokens with our SE framework ("+ SE") could outperform "+ ADD" by +0.6 BLUE points, demonstrating ***the superiority of our token-specific label smoothing approach***.

	Baseline	**+ ADD**	**+ SE**
BLEU	35.1	35.4	**36.0**

Table 9: Ablation performance of our SE. on Metric.

Learning objective. As stated in §2.1, our learning objective is the combination of the ground truth and the model's prediction. To validate the effectiveness of predicted distribution, we conduct ablation experiments on WMT16 En-Ro and WMT14 En-De. The results in Table 10 show that adding the predicted distribution will consistently improve the model's performance, which proves the effectiveness of the predicted distribution.

Method	**BLEU**	
	EN⇒DE	**EN⇒Ro**
Transformer	27.08	35.11
SE	**28.02**	**36.02**
-w/o predicted results	27.89	35.71

Table 10: Ablation performance of our SE. on learning objective.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The last section of the paper.

☒ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The abstract and the introduction section.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3.2

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.