Single Model Ensemble for Subword Regularized Models in Low-Resource Machine Translation

Sho Takase and **Tatsuya Hiraoka** and **Naoaki Okazaki**
Tokyo Institute of Technology
{sho.takase@nlp., tatsuya.hiraoka@nlp., okazaki@}c.titech.ac.jp

Abstract

Subword regularizations use multiple subword segmentations during training to improve the robustness of neural machine translation models. In previous subword regularizations, we use multiple segmentations in the training process but use only one segmentation in the inference. In this study, we propose an inference strategy to address this discrepancy. The proposed strategy approximates the marginalized likelihood by using multiple segmentations including the most plausible segmentation and several sampled segmentations. Because the proposed strategy aggregates predictions from several segmentations, we can regard it as a single model ensemble that does not require any additional cost for training. Experimental results show that the proposed strategy improves the performance of models trained with subword regularization in low-resource machine translation tasks.

1 Introduction

Subword regularizations are the technique to make a model robust to segmentation errors by using multiple subword segmentations instead of only the most plausible segmentation during the training process (Kudo, 2018; Provilkov et al., 2020). Previous studies demonstrated that subword regularizations improve the performance of LSTM-based encoder-decoders and Transformers in various machine translation datasets, especially in low-resource settings (Kudo, 2018).

However, previous subword regularizations contain the discrepancy between the training and inference. In the training process, we stochastically re-segment a given sequence into subwords based on statistics such as the uni-gram language model (Kudo, 2018). Thus, we use multiple segmentations for each input sequence. In contrast, we use only the most plausible segmentation in the inference phase. We expect that we can improve the performance by solving this discrepancy.

To solve this discrepancy, we propose an inference strategy that uses multiple subword segmentations. We construct multiple subword segmentations for an input in the same manner as that in the training process, and then aggregate the predictions from each segmentation. Therefore, our proposed inference strategy can be regarded as a single model ensemble using multiple segmentations. Figure 1 illustrates the overview of previous methods and our proposed inference strategy.

We conduct experiments on several machine translation datasets. Experimental results show that the proposed strategy improves the performance of a subword regularized model without any additional costs in the training procedure when the subword regularization significantly contributes to the performance, i.e., in low-resource settings. Moreover, we indicate that our strategy can be combined with a widely used model ensemble technique.

2 Subword Regularization

Our proposed strategy is based on a model trained with subword regularization. Thus, we briefly describe subword regularization in this section.

Kudo (2018) proposed subword regularization to improve the robustness of a neural machine translation model. Let X and Y be the source and target sentences, $\mathbf{x} = (x_1, ..., x_S)$ and $\mathbf{y} = (y_1, ..., y_T)$ be the most plausible subword segmentations corresponding to X and Y. In the vanilla training strategy, i.e., without subword regularization, we train the parameters of a neural machine translation model $\boldsymbol{\theta}$ to maximize the following log-likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{(X,Y)\in\mathcal{D}} \log P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}), \quad (1)$$

$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \prod_{t=1}^{T} P(y_t|\mathbf{x}, \mathbf{y}_{<t};\boldsymbol{\theta}), \quad (2)$$

where $\mathcal{D}$ is the training data and $\mathbf{y}_{<t} = (y_1, ..., y_{t-1})$.

Findings of the Association for Computational Linguistics: ACL 2022, pages 2536 - 2541

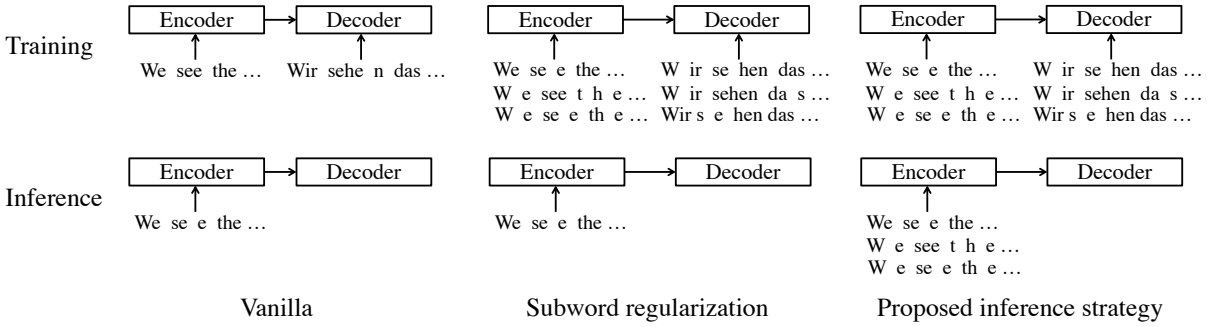

Figure 1: Overview of previous methods and the proposed inference strategy for an English-German pair "We see the ..." and "Wir sehen das ...". In the vanilla setting, we use the most plausible segmentation only in the training and inference. In the subword regularization, we use multiple segmentations during training but use only the most plausible segmentation in the inference phase. In the proposed inference strategy, we use multiple segmentations in both the training and inference phases.

In contrast, subword regularization uses multiple subword segmentations during training. Let $P(\mathbf{x}'|X)$ and $P(\mathbf{y}'|Y)$ be segmentation probabilities for sequences X and Y, respectively. We optimize the parameters $\boldsymbol{\theta}$ with the following marginalized likelihood in subword regularization:

$$\mathcal{L}'(\boldsymbol{\theta}) = \sum_{(X,Y)\in\mathcal{D}} \mathbb{E}_{\substack{\mathbf{x}'\sim P(\mathbf{x}'|X) \\ \mathbf{y}'\sim P(\mathbf{y}'|Y)}}[\log P(\mathbf{y}'|\mathbf{x}';\boldsymbol{\theta})]. \tag{3}$$

Because the number of possible segmentations increases exponentially with respect to the sequence length, it is impractical to optimize Equation (3) exactly. Thus, Kudo (2018) approximated Equation (3) with sampled segmentations from $P(\mathbf{x}'|X)$ and $P(\mathbf{y}'|Y)$,

$$\mathcal{L}'(\boldsymbol{\theta}) \cong \sum_{(X,Y)\in\mathcal{D}} \log P(\mathbf{y}_j|\mathbf{x}_i;\boldsymbol{\theta}), \tag{4}$$

$$\mathbf{x}_i \sim P(\mathbf{x}'|X), \tag{5}$$

$$\mathbf{y}_j \sim P(\mathbf{y}'|Y). \tag{6}$$

We sample $\mathbf{x}_i$ and $\mathbf{y}_j$ for every mini-batch during training to yield a good approximation.

In the inference phase, we input the most plausible segmentation $\mathbf{x}$ and search a sequence $\mathbf{y}^*$ that maximizes the log-likelihood $\log P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$. In other words, we input one segmentation to the model[1] even though we use multiple segmentations during training.

[1] Kudo (2018) also proposed n-best decoding. This strategy uses n segmentations but inputs them separately. In other words, a model receives only one segment and generates the corresponding output n times in this strategy. We compare this strategy in experiments.

Language	Vocab	Train	Dev	Test
En-De	6K	160K	7283	6750
En-Vi	4K	133K	1553	1268

Table 1: Details of each dataset.

3 Proposed Method

3.1 Proposed Inference Strategy

As described, previous subword regularizations use multiple segmentations during training but only one segmentation in the inference. To solve this discrepancy, we propose an inference strategy that uses multiple segmentations as inputs. In the proposed strategy, we search a sequence $\mathbf{y}^*$ that maximizes the following approximated marginalized likelihood:

$$\sum_{k=1}^{n} \log P(\mathbf{y}|\mathbf{x}_k;\boldsymbol{\theta}), \tag{7}$$

$$\mathbf{x}_k = \begin{cases} \mathbf{x} & k = 1 \\ \mathbf{x}_i \sim P(\mathbf{x}'|X) & \text{Otherwise.} \end{cases} \tag{8}$$

In short, we approximate the marginalized likelihood in Equation (3) with the most plausible segmentation and sampled $n-1$ segmentations.

3.2 Relation to Model Ensemble

We often apply the model ensemble technique to achieve better performance (Barrault et al., 2019). In the model ensemble, we aggregate the predictions from M models as follows:

$$\sum_{m=1}^{M} \log P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_m), \tag{9}$$

Method	En-De	De-En	En-Vi	Vi-En
Single Model				
Vanilla	28.89	34.87	31.09	31.43
+ w/ subword regularization (1)	29.51	35.53	31.86	**31.60**
(1) + n-best decoding	29.59	35.55	31.94	31.44
(1) + Proposed strategy	**29.72**	**35.68**	**32.16**	**31.60**
Model Ensemble				
Vanilla	30.03	36.04	32.22	32.46
+ w/ subword regularization (2)	30.83	36.83	33.22	32.83
(2) + n-best decoding	30.81	36.83	33.29	32.76
(2) + Proposed strategy	**30.86**	**36.95**	**33.44**	**33.04**

Table 2: BLEU scores on English-German and English-Vietnamese datasets.

where $\boldsymbol{\theta}_m$ denotes parameters of the m-th model.

In comparison to this model ensemble, the proposed strategy does not use multiple models but aggregates predictions from multiple segmentations. Thus, our proposed strategy can be regarded as the single model ensemble with multiple inputs. In addition, we can combine the proposed strategy with the model ensemble. We investigate the effect of this combination through experiments.

4 Experiments

4.1 Datasets

Kudo (2018) reported that subword regularization is especially effective in low-resource settings. Thus, we focus on low-resource machine translation tasks. We used IWSLT 2014 English-German (En-De) data in the same pre-processing manner as Ranzato et al. (2016)[2] because this dataset is widely-used as the low-resource setting (Sennrich and Zhang, 2019; Takase and Kiyono, 2021). In addition, we used IWSLT 2015 English-Vietnamese (En-Vi) data which were pre-processed by Luong and Manning (2015)[3].

We used SentencePiece (Kudo and Richardson, 2018) to construct a vocabulary set. We set the vocabulary sizes to 6k and 4k for En-De and En-Vi, respectively. Table 1 summarizes the dataset sizes.

4.2 Methods

We used Transformer (Vaswani et al., 2017) as our encoder-decoder architecture because Transformers are widely used as strong baselines in sequence-to-sequence problems including machine translation. We investigate the performance of the following configurations.

Vanilla: We trained Transformer (Vaswani et al., 2017) without subword regularization. For hyper-parameters, we adopted the IWSLT setting in `fairseq`[4] (Ott et al., 2019).

Subword regularization: We trained Transformer, whose hyper-parameters are identical to Vanilla, with subword regularization. We set the hyper-parameter α for sampling segmentations in subword regularization 0.2 in the same as Kudo (2018).

n-best decoding: Kudo (2018) proposed n-best decoding that generates n sequences corresponding to n-best segmentations and then outputs the most plausible sequence. We used this strategy for the model trained with subword regularization in the inference phase.

Proposed: We applied the proposed strategy to the model trained with subword regularization. To ensure fair comparison, we used the identical number, $n = 5$, for the number of sampled segmentations and n-best decoding.

4.3 Results

Table 2 shows BLEU scores of each configuration. For each configuration, we trained three models with different random seeds, and reported the averaged scores except for the proposed strategy. When we used the proposed strategy, we generated sequences three times with different random seeds for each model[5], and averaged the 9 (3 models $\times$

[2] github.com/pytorch/fairseq/blob/master/examples/translation/
[3] https://nlp.stanford.edu/projects/nmt/
[4] https://github.com/pytorch/fairseq
[5] Because the generated sequence mainly depends on the trained model, our inference strategy generates almost the same sequences even if we vary random seeds for samplings. However, we reported the averaged BLEU of 9 sequences to make the results more reliable.

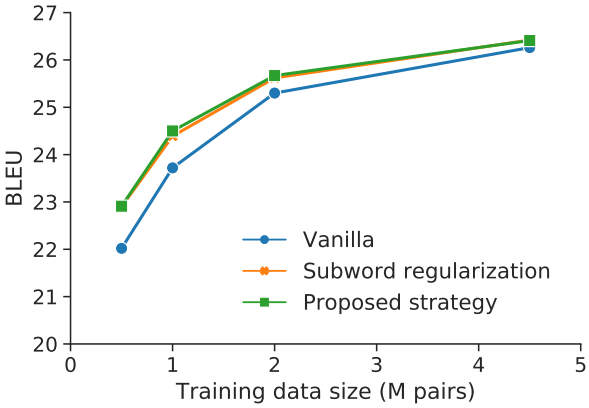

Figure 2: BLEU scores on newstest2013 when we vary the training data size.

3 sequences) scores. Table 2 also indicates BLEU scores with the ensemble of the above 3 models.

For the single model setting, Table 2 shows that subword regularization improved BLEU scores in all language pairs. In particular, subword regularization gained more than 0.5 BLEU score from Vanilla except for Vi-En. In these language pairs, the proposed strategy provided further improvements. The proposed strategy achieved better performance than n-best decoding when we used the same number of segmentations as inputs. Thus, our proposed method is more effective as the inference strategy. Moreover, our strategy maintained the score in Vi-En although n-best decoding degraded the score slightly. Therefore, the proposed strategy had no negative effect on the inference.

For the model ensemble setting, Table 2 indicates that subword regularization also improved BLEU scores in all language pairs. In this setting, the proposed strategy also provided further improvements in all language pairs. Thus, the proposed strategy is effective even if we conduct the model ensemble technique.

5 Performance in Enough Training Data

Section 4 shows the results in low-resource settings but previous studies reported that subword regularizations can improve the performance if we use sufficient training data. Thus, we investigate the performance of subword regularization and proposed strategy by varying the size of training data.

We used the WMT 2016 English-to-German training dataset, which is widely used in previous studies (Vaswani et al., 2017; Provilkov et al., 2020; Ott et al., 2018). This dataset contains 4.5M sentence pairs, that are more than 25 times as many as IWSLT datasets. We conducted pre-processing in the same manner as that in Ott et al. (2018). We trained the Transformer (base) model in Vaswani et al. (2017). For subword regularization, we set $\alpha = 0.5$ in the same as Kudo (2018). We evaluated BLEU scores on newstest2013, which is widely used as a valid data.

Figure 2 shows BLEU scores of each method for each training data size. This figure indicates that the model trained with subword regularization outperformed Vanilla in all training data sizes but the improvement decreased in accordance with the increase in the training data. The proposed strategy slightly improved the performance from subword regularization for the small training data but the improvement also decreased as the training data increased. When we used the entire training data (4.5M translation pairs), the BLEU score of the proposed strategy was identical to that of subword regularization. This result implies that the impact of the proposed strategy on the performance is small when the improvement by subword regularization is small. In other words, the proposed strategy is effective especially in low-resource settings because subword regularization probably provides much improvement in low-resource settings. However, we emphasize that the proposed strategy has no negative effect on the BLEU score for sufficient training data fortunately.

6 Related Work

In this study, we proposed the inference strategy to mitigate the discrepancy between the training and inference in subword regularizations. In experiments, we focused the subword regularization proposed by Kudo (2018) but we can apply the proposed inference strategy to variants of the subword regularization such as BPE dropout (Provilkov et al., 2020) and compositional word replacement (Hiraoka et al., 2022). Takase and Kiyono (2021) reported that simple perturbations such as word dropout are effective in a large amount of training data. Thus, we might improve the performance of the model trained with such simple perturbations if we use multiple inputs constructed by the same perturbation during the inference.

We focused on an input of a neural encoder-decoder. In contrast, Gal and Ghahramani (2016) focused on internal layers. For neural network methods, we often apply the dropout during the training but do not use it in the inference. Gal

and Ghahramani (2016) proposed the variational inference to mitigate this gap on the dropout.

As described in Section 1, our proposed inference strategy can be regarded as a single model ensemble. Huang et al. (2017) and Kuwabara et al. (2020) also proposed single model ensemble methods. Huang et al. (2017) proposed the snapshot ensemble that uses multiple models in the middle of the training. Kuwabara et al. (2020) used pseudo-tags and predefined distinct vectors to obtain multiple models virtually during the training of a single model. Since these methods are orthogonal to ours, we can combine our proposed strategy.

7 Conclusion

We proposed an inference strategy to address the discrepancy between the training and inference in subword regularizations. Our proposed strategy uses multiple subword segmentations as inputs to approximate the marginalized likelihood used as the objective function during training. The proposed strategy improved the performance of the model trained with subword regularization in cases where subword regularization provided the significant improvement, i.e., in low-resource settings. Moreover, the proposed strategy outperformed the n-best decoding strategy (Kudo, 2018). Experimental results show that our proposed strategy has no negative effect on the BLEU score even if the improvement by subword regularization is small. Because the proposed inference strategy does not require any additional training cost, we encourage using the strategy to highlight the potential of models trained with subword regularization.

Ethical Considerations

Limitations: The proposed method improves the performance of encoder-decoders in the inference phase in the situation where subword regularizations are effective. Thus, if subword regularizations are ineffective, the proposed method also might be ineffective. Since subword regularizations are especially effective when the training data size is small (Hiraoka et al., 2021), the proposed method is effective in low-resource settings. In contrast, as in Section 5, the improvements of both methods are small when we have an enough training data.

Risks: Since the proposed method uses the standard neural encoder-decoder architecture without any modification, the proposed method also contains the risks of neural encoder-decoders. For example, the under translation, that ignores some information in a source sentence during the translation, might happen.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21K17800 and JST ACT-X Grant Number JPMJAX200I. The first author is supported by Microsoft Research Asia (MSRA) Collaborative Research Program.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (WMT)*, pages 1–61.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. 2021. Joint optimization of tokenization and downstream model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 244–255.

Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. 2022. Word-level perturbation considering word length and compositional subwords. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. 2017. Snapshot ensembles: Train 1, get M for free. In *Proceedings of the Fifth International Conference on Learning Representations (ICLR)*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 66–75.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71.

Ryosuke Kuwabara, Jun Suzuki, and Hideki Nakayama. 2020. Single model ensemble using pseudo-tags and distinct vectors. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics (ACL), pages 3006–3013.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation (IWSLT)*.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 48–53.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*, pages 1–9.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1882–1892.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the Fourth International Conference on Learning Representations (ICLR)*.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 211–221.

Sho Takase and Shun Kiyono. 2021. Rethinking perturbations in encoder-decoders for fast training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5767–5780.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008.