Life Is Short, Train It Less: Neural Machine Tibetan-Chinese Translation Based on mRASP and Dataset Enhancement

Hao Wang[1], Yongbin Yu[1(✉)], Nyima Tashi[2(✉)], Rinchen Dongrub[2], Ekong Favour[1], Mengwei Ai[1], Kalzang Gyatso[2], Yong Cuo[2], and Qun Nuo[2]

[1] University of Electronic Science and Technology of China, Chengdu 610000, China
ybyu@uestc.edu.cn
[2] Engineering Research Center for Tibetan Information Processing, School of Information Science and Technology, Tibet University, Lhasa 850000, China
nmzx@utibet.edu.cn

Abstract. This paper highlights a multilingual pre-trained neural machine translation architecture as well as a dataset augmentation approach based on curvature selection. The multilingual pre-trained model is designed to increase the performance of machine translation with low resources by bringing in more common information. Instead of repeatedly training several checkpoints from scratch, this study proposes a checkpoint selection strategy that uses a cleaned optimizer to hijack a midway status. Experiments with our own dataset on the Chinese-Tibetan translation demonstrate that our architecture gets a 32.65 BLEU score, while in the reverse direction, it obtains a 39.51 BLEU score. This strategy drastically reduces the amount of time spent training. To demonstrate the validity of our method, this paper shows a visualization of curvature for a real-world training scenario.

Keywords: Neural machine translation · Dataset enhancement · Chinese-Tibetan translation · Curvature

1 Introduction

Many fields, such as education, publishing, and information security, have a strong demand for Chinese-Tibetan translation algorithms. During the last few years, neural machine translation (NMT) tasks have had a great deal of success thanks to the transformer architecture [11]. However, one of the key drawbacks of this approach is that transformer is greedy in terms of both quality and quantity of data. This study set out with the aim of investigating a training method including cross-language transferable pre-trained models and dataset

The demo of this paper is available at http://mt.utibet.edu.cn.

T. Xiao and J. Pino (Eds.): CCMT 2022, CCIS 1671, pp. 54–59, 2022.
https://doi.org/10.1007/978-981-19-7960-6_6

enhancement algorithm to achieve better results in low-resource Chinese-Tibetan translation tasks.

Many studies focus on cross-language unified models, which may increase translation quality in low-resource languages, on the assumption that a cross-language unified model will acquire common knowledge between languages to boost performance on unseen data [2,3]. Johnson et al. add an artificial token at the beginning of the sentence to set the required language [4]. Zhang et al. suggest that the off-target translation issue is the main reason for unexpected zero-shot performance. [12] Xiao et al. involve contrastive learning in their mRASP framework which achieve good results on other language pair yet misses the Tibetan language dataset [7].

The use of dataset enhancement as a solution to the data hungriness problem is another alternative. One of the merits of such methods is that most of them proceed in parallel with model architecture, saving time on model modifications. Aside from data augmentation, the back translation method [1,8] is a simple yet effective way to generate synthetic data to improve efficiency. Although this methodology is highly effective, it involved the use of additional monolingual data. Nguyen et al. [5] propose an interesting method that generates a diverse set of synthetic data to augment original data. This method is powerful and effective yet still required training multiple loops.

2 Prerequisite

2.1 Neural Machine Translation with mRASP

mRASP uses a standard Transformer with both 6 layers encoder and decoder pre-trained jointly on 32 language pairs. In this paper, we use the pre-trained mRASP model to finetune our Chinese-Tibetan translation model. Following the symbols in mRASP, we denote the Chinese-Tibetan parallel dataset as (L_{src}, l_{tgt}), the finetuned loss is

$$\mathcal{L}^{finetune} = \mathbb{E}_{(\mathbf{x}^i, \mathbf{x}^j) \sim \mathcal{D}_{src,tgt}} \left[-\log P_\theta \left(\mathbf{x}^i \mid \mathbf{x}^j \right) \right]. \tag{1}$$

where the θ is the pretrained mRASP model.

2.2 Diversification Method

Data diversification is a simple yet effective data augmentation method. It trains predictions from multiple bi-direction models to diversify training data which is ideal for the low-source Chinese-Tibetan translation task. This strategy is formulated as:

$$\mathcal{D} = (S, T) \bigcup \cup_{i=1}^{k} \left(S, M^i_{S \to T,1}(S) \right) \bigcup \cup_{i=1}^{k} \left(M^i_{T \to S,1}(T), T \right) \tag{2}$$

M denotes the model and k is the diversification factor. In this paper, we propose an accelerating hijack method to reduce this training burden significantly.

2.3 Curvature

In this work, we choose the curvature as the metric of the sharpness of the perplexity curve for the validation dataset in the whole training process. Denote K as the curvature, for a continuous curve it can be calculated as:

$$K = \frac{1}{r} = \frac{|f''(x_0)|}{\left(1 + (f'(x_0))^2\right)^{\frac{3}{2}}} \tag{3}$$

However, the valid perplexity averaged within an epoch is discrete and the direct finite difference may bring relatively large error. In this work, we use the curvature of the quadratic curve determined by the nearest three points to estimate the curvature of a valid perplexity curve [13].

3 Methodology

3.1 Overall Structure

In this work, mRASP pretrained on 32 language pairs is utilized to provide a good starting point than plain Transformer. The vocabulary for our 115k dataset is merged into the provided vocabulary of mRASP. Then the private Tibetan-Chinese parallel dataset is utilized to generate an enhanced dataset. As shown in Fig. 1, the whole fine-tune stage is divided into three parts based on the valid perplexity averaged on each epoch. We hijack the checkpoint from the key points and then continue training using a cleaned optimizer to generate k more checkpoints. Along with the main checkpoint, the enhanced dataset is generated to train the final model.

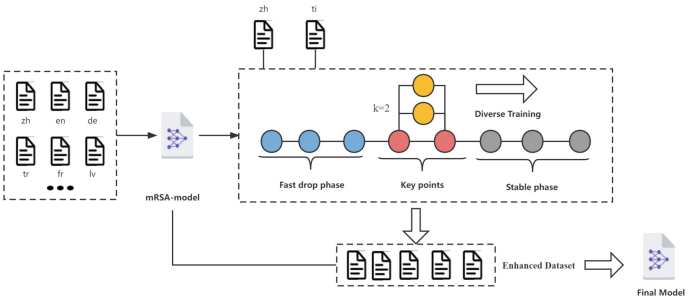

Fig. 1. The overall architecture of this work. The proposed work can be divided into three stages. (1) In the pre-trained stage, the multi-lingual pre-trained mRASP model is prepared for further finetune. (2) In the dataset enhancement stage, $m + 1$ checkpoints are trained and inference to generate an enhanced dataset. (3) The final translation model is finetuned based on mRASP at the enhanced dataset.

3.2 Curvature Based Checkpoint Hijack

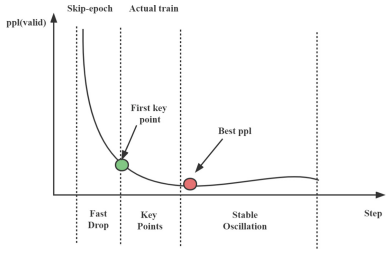

Fig. 2. The curvature change for an ideal training process. The green point denotes the first key point which is used to re-train. The red point denotes the best perplexity which is an ideal end-point. (Color figure online)

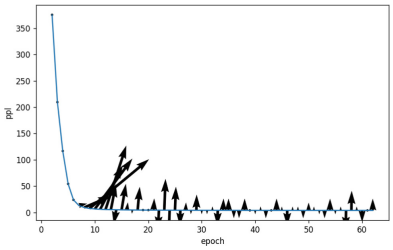

Fig. 3. Actual ppl change during training. The very first epoch is deleted for better visualization without changing the shape of the curve. The curvature is visualized as black arrows.

In this paper, we argue that it is not necessary to train the entire procedure for a large pre-trained model like mRASP. Figure 2 illustrates the perplexity of the valid set will go under three stages. In the fast drop stages, the perplexity will sharply drop to fit the new dataset. Then in the key points stage, the perplexity will gradually get smooth to the minimal value. The final stage is the stable oscillation stage where the perplexity will not change fast. Instead of training from scratch, the curvature is involved to quantify the key points. To ensure the model status is as far as possible from the best point, a few checkpoints before the first key checkpoint are averaged along with the key checkpoint to ensure maximum diversity (Fig. 3).

Formally, denote the training epoch as N. The calculated curvature for valid set is denoted as a sequence $\mathscr{A} := \{k_1, k_2, \ldots k_i, \ldots k_N\}, k_j \in \mathbb{R}$ where i denotes the very first key point. By setting the threshold for curvature as hyperparameter T, the key points can be formulated as:

$$\mathscr{S} := \{k \in \mathscr{A} \mid k_i \geq k_j, k_i \geq T, \forall i, j \in \mathbb{R}, i < j\}. \tag{4}$$

The generated parallel dataset is:

$$\mathcal{D} = (S, T) \bigcup \cup_{i=1}^{k} \left(S, \frac{1}{m} \sum_{i-m}^{i} M^{i}_{S \to T, i}(S) \right) \bigcup \cup_{i=1}^{k} \left(\frac{1}{m} \sum_{i-m}^{i} M^{i}_{T \to S, i}(T), T \right) \tag{5}$$

where m is the total averaged checkpoints numbers and i is the smallest index in $\mathscr{S}$.

4 Experiments

4.1 Dataset Description and Finetune Parameters

This paper uses the Chinese-Tibetan parallel dataset constructed by Tibet University and Qinghai Normal University. It contains high-quality parallel sentences checked and approved by professionals. The Chinese segment tool is Jieba and the Tibetan segment tool is based on perceptron and CRF developed by Tsering et al. [10] In the fine-tuning process, both the input and output length is restricted to 300. The optimizer is Adam and the learning rate is set to 1e−4. Label smoothing is set to 0.1 and mixed precision is used. The diversion factor k is set to 2 and the average checkpoint number m is 3. We perform our experiments in RTX 3090 and A5000 with fairseq [6].

4.2 Experiment Result

Table 1. BLEU score reported on test set

Task	Direction	BLEU	Epoch
Base [9]	zh-ti	30.46	X
	ti-zh	X	X
Train-scratch	zh-ti	27.17	90
	ti-zh	37.34	90
mRASP-finetuned	zh-ti	32.65	90
	ti-zh	39.51	90
Hijack-enhanced (our)	zh-ti	33.04	90(47)
	ti-zh	39.87	90(47)

Table 1 shows the BLEU score on the test set. Compared to baseline, the mRASP-based pre-trained model indeed performs better. For the training epochs, 90(47) means that we first train an entire loop and then use the best ppl as a stopping point so the next m training will stop at it. The hijack-enhanced dataset brings slightly better benefits than mRASP. However, it is worth mentioning that it only takes dozens of extra epochs to fine-tune, which is faster than the original diversity approach.

5 Conclusion

In this paper, a neural machine translation architecture is proposed for Chinese-Tibetan translation. The involvement of curvature selection reduces the training time significantly. The experiments demonstrate that a multilingual pre-trained model can boost low resources language translation performance. More discussion of curvature in neural networks is desirable for future work.

Acknowledgement. This paper is supported by the Chinese Tibetan English neural machine translation system, the artificial intelligence industry innovation task of the Ministry of industry and information technology with the open competition mechanism to select the best candidates to undertake key research projects. The authors thank you for the guidance of the reviewers!

References

1. Edunov, S., Ott, M., Ranzato, M., Auli, M.: On the evaluation of machine translation systems trained with back-translation. arXiv preprint arXiv:1908.05204 (2019)
2. Gu, J., Wang, Y., Cho, K., Li, V.O.: Improved zero-shot neural machine translation via ignoring spurious correlations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1258–1268. Association for Computational Linguistics, Florence (2019)
3. Ji, B., Zhang, Z., Duan, X., Zhang, M., Chen, B., Luo, W.: Cross-lingual pre-training based transfer for zero-shot neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 115–122 (2020)
4. Johnson, M., et al.: Google's multilingual neural machine translation system: enabling zero-shot translation. Trans. Assoc. Comput. Linguist. **5**, 339–351 (2017)
5. Nguyen, X.P., Joty, S., Kui, W., Aw, A.T.: Data diversification: a simple strategy for neural machine translation. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2020)
6. Ott, M., et al.: Fairseq: a fast, extensible toolkit for sequence modeling. In: Proceedings of NAACL-HLT 2019: Demonstrations (2019)
7. Pan, X., Wang, M., Wu, L., Li, L.: Contrastive learning for many-to-many multilingual neural machine translation. arXiv preprint arXiv:2105.09501 (2021)
8. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709 (2015)
9. Cairang, T., Dongzhu, R., Zhaxi, N., Yongbin, Y., Quanxin, D.: Research on Chinese-Tibetan machine translation model based on improved byte pair encoding. J. Univ. Electron. Sci. Technol. **50**(02), 249–255+293 (2021)
10. Tsering, T., Dhondub, R., Tashi, N.: Research on Tibetan location name recognition technology under CRF. Comput. Eng. Appl. **55**(18), 111 (2019)
11. Vaswani, A., et al.: Attention is all you need. In: Advances in neural information processing systems, vol. 30 (2017)
12. Zhang, B., Williams, P., Titov, I., Sennrich, R.: Improving massively multilingual neural machine translation and zero-shot translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1628–1639. Association for Computational Linguistics (2020)
13. Zhang, P., Wang, C.B., Ye, L.: A type iii radio burst automatic analysis system and statistic results for a half solar cycle with nançay decameter array data. Astron. Astrophys. **618**, A165 (2018)