Bilingual Mutual Information Based Adaptive Training for Neural Machine Translation

Yangyifan Xu[1]*, Yijin Liu[1,2], Fandong Meng[2], Jiajun Zhang[3,4], Jinan Xu[1]†and Jie Zhou[2]
[1]Beijing Jiaotong University, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
[3]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences
xuyangyifan2021@ia.ac.cn, adaxry@gmail.com
{fandongmeng, withtomzhou}@tencent.com
jjzhang@nlpr.ia.ac.cn, jaxu@bjtu.edu.cn

Abstract

Recently, token-level adaptive training has achieved promising improvement in machine translation, where the cross-entropy loss function is adjusted by assigning different training weights to different tokens, in order to alleviate the token imbalance problem. However, previous approaches only use static word frequency information in the target language without considering the source language, which is insufficient for bilingual tasks like machine translation. In this paper, we propose a novel bilingual mutual information (BMI) based adaptive objective, which measures the learning difficulty for each target token from the perspective of bilingualism, and assigns an adaptive weight accordingly to improve token-level adaptive training. This method assigns larger training weights to tokens with higher BMI, so that easy tokens are updated with coarse granularity while difficult tokens are updated with fine granularity. Experimental results on WMT14 English-to-German and WMT19 Chinese-to-English demonstrate the superiority of our approach compared with the Transformer baseline and previous token-level adaptive training approaches. Further analyses confirm that our method can improve the lexical diversity.

1 Introduction

Neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017; Chen et al., 2018; Meng and Zhang, 2019; Zhang et al., 2019; Yan et al., 2020; Liu et al., 2021) has achieved remarkable success. As a data-driven model, the performance of NMT depends on training corpus. Balanced training data is a crucial factor in building a superior model. However, natural languages conform to the Zipf's law (Zipf, 1949), the frequencies of words exhibit the long tail characteristics, which brings an imbalance in the distribution of words in training corpora. Some studies (Jiang et al., 2019; Gu et al., 2020) assign different training weights to target tokens according to their frequencies. These approaches alleviate the token imbalance problem and indicate that tokens should be treated differently during training.

However, there are two issues in existing approaches. First, these approaches believe that low-frequency words are not sufficiently trained and thus amplify the weight of them. Nevertheless, low-frequency tokens are not always difficult as the model competence increases (Wan et al., 2020). Second, previous studies only use monolingual word frequency information in the target language without considering the source language, which is insufficient for bilingual tasks, e.g., machine translation. The mapping between bilingualism is a more appropriate indicator. As shown in Table 1, word frequency of *pleasing* and *bearings* are both 847. Corresponding to Chinese, *pleasing* has multiple mappings, while *bearings* is relatively single. The more multivariate the mapping is, the less confidence in predicting the target word given the source context. He et al. (2019) also confirm this view that words with multiple mappings contribute more to the BLEU score.

To tackle the above issues, we propose bilingual mutual information (BMI), which has two characteristics: 1) BMI measures the learning difficulty for each target token by considering the strength of association between it and the source sentence; 2) for each target token, BMI can dynamically adjust according to the context. BMI-based adaptive training can dynamically adjust the learning granularity on tokens. Easy tokens are updated with coarse granularity while difficult tokens are updated with

*This work was done when Yangyifan Xu was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China
†Jinan Xu is the corresponding author of the paper.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 511–516
August 1–6, 2021.

pleasing (847)	gāoxìng (81); yúkuài (74); xǐyuè (63); qǔyuè (49) ...
bearings (847)	zhóuchéng (671) ...

Table 1: An example from the WMT19 Chinese-English training set. The Chinese words are presented in pinyin style and the word frequency is shown in brackets. The two words have the same word frequency, while the mapping of *bearings* is more stable than that of *pleasing*.

fine granularity.

We evaluate our approach on both WMT14 English-to-German and WMT19 Chinese-to-English translation tasks. Experimental results on two benchmarks demonstrate the superiority of our approach compared with the Transformer baseline and previous token-level adaptive training approaches. Further analyses confirm that our method can improve the lexical diversity. The main contributions[1] of this paper can be summarized as follows:

- We propose a training objective based on bilingual mutual information (BMI), which can reflects the learning difficulty for each target token from the perspective of bilingualism, and assigns an adaptive weight accordingly to guide the adaptive training of machine translation.
- Experimental results show that our method can improve not only the machine translation quality, but also the lexical diversity.

2 Background

2.1 Neural Machine Translation

A NMT system is a neural network that translates a source sentence $\mathbf{x}$ with n words to a target sentence $\mathbf{y}$ with m words. During the training process, NMT models are optimized by minimizing cross entropy:

$$\mathcal{L} = -\frac{1}{m}\sum_{j=1}^{m} \log p(y_j|\mathbf{y}_{<j}, \mathbf{x}), \quad (1)$$

where y_j is the ground-truth token at the j-th position and $\mathbf{y}_{<j}$ is the translation history known before predicting token y_j.

[1]Reproducible code: https://github.com/xydaytoy/BMI-NMT

BMI= 2.29	In ball **bearings**, as the radial clearance increases , the axial clearance increases as well. zài qiú **zhóu chéng**, jìng xiàng jiàn xì de zēng jiā, zhóu xiàng yóu xì zēng zhǎng.
BMI= 1.83	One of his crowd **pleasing** notions is that migrants will infect Americans with terrible diseases. tā **qǔ yuè** qún méng de gài niàn zhī yī shì yí mín huì jǐ měi guó rén dài lái kě pà de chuán rǎn bìng.

Figure 1: An example from WMT19 Chinese-to-English training set. Words with Red and Bold fonts have the same word frequency while different BMI.

2.2 Token-level Adaptive Training Objective

Following (Gu et al., 2020), the token-level adaptive training objective is

$$\mathcal{L} = -\frac{1}{m}\sum_{j=1}^{m} w_j \cdot \log p(y_j|\mathbf{y}_{<j}, \mathbf{x}), \quad (2)$$

where w_j is the weight assigned to the target token y_j . Gu et al. (2020) used monolingual word frequency information in the target language to calculate the w_j. The weight does not contain the information of the source language, and cannot be dynamically adjusted with the context.

3 BMI-based Adaptive Training

In this section, we start with the definition of the bilingual mutual information (BMI). Then we analyze the relationship between BMI and translation difficulty. Based on this, we introduce our BMI-based token-level adaptive training objective.

3.1 Definition of BMI

Mutual information measures the strength of association between two random variables by comparing the number of their individual and joint occurrences. We develop BMI, which is calculated by summarizing the mutual information of the target token and each token in the source sentence, to measure the learning difficulty of the model. Token pairs with high BMI are considered easy, since they have high co-occurrence relative to the frequency. Given the source sentence $\mathbf{x}$ and target token y_j, we define the bilingual mutual information as[2]:

$$\mathrm{BMI}(\mathbf{x}, y_j) = \sum_{i=1}^{n} \log \frac{f(x_i, y_j)}{f(x_i) \cdot f(y_j)/K}, \quad (3)$$

[2]To ensure comparability of two probability distribution, the tokens that appear multiple times in a sentence and the token pairs that appear multiple times in a sentence pair are not counted repeatedly.

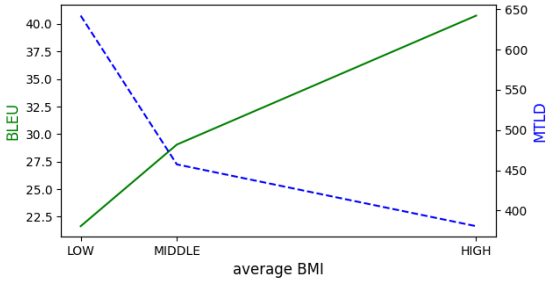

Figure 2: The BLEU (green solid line) and MTLD (blue dotted line) values on the subsets of WMT14 English-German training set, divided according to the average BMI. All target sentences of the training set are divided into three subsets according to the average BMI of the tokens in the sentence, which are equal in size and denoted as LOW, MIDDLE, and HIGH, respectively. BLEU indicates the learning difficulty of the model. The measure of textual lexical diversity (MTLD) (McCarthy and Jarvis, 2010) represents the lexical diversity of the data set. The results show that high BMI means relatively stable mapping, which is easy to be learned by the model and has low lexical diversity.

where $f(x_i)$ and $f(y_j)$ are total number of sentences in the corpus containing at least one occurrence of x_i and y_j, respectively, $f(x_i, y_j)$ represents total number of sentences in the corpus having at least one occurrence of the word pair (x_i, y_j), and K denotes total number of sentences in the corpus.

3.2 What BMI Measures?

We use an example to illustrate our idea. Figure 1 shows two sentence pairs. Words with Red and Bold fonts have the same word frequency. As shown in Table 1, *pleasing* has multiple mappings, while the mapping of *bearings* is relatively single. As a result, the appearance of corresponding English word brings different confidence of the appearance of the Chinese word, which can be reflected by BMI. Further statistical results are shown in Figure 2, high BMI means relatively stable mapping, which is easy to be learned by the model and has low lexical diversity.

3.3 BMI-based Objective

We calculate the token-level weight by scaling BMI and adjusting the lower limit as follows:

$$w_j = S \cdot \mathrm{BMI}(\mathbf{x}, y_j) + B. \quad (4)$$

The two hyperparameters S (scale) and B (base) influence the magnitude of change and the lower limit, respectively.

In training process, the loss of simple tokens will be amplified, the model updates simple tokens with coarse granularity, because our strategy thinks the model can easily predict these target tokens given the source sentence, and it needs to increase the penalty if the prediction is wrong. For difficult tokens, the model has a higher tolerance because their translation errors may not be absolute. As a result, the loss is small due to the small weight and the difficult tokens are always updated in a fine-grained way.

4 Experiments

We evaluate our method on the Transformer (Vaswani et al., 2017) and conduct experiments on two widely-studied NMT tasks, WMT14 English-to-German (En-De) and WMT19 Chinese-to-English (Zh-En).

4.1 Data Preparation

EN-DE. The training data consists of 4.5M sentence pairs from WMT14. Each word in the corpus has been segmented into subword units using byte pair encoding (BPE) (Sennrich et al., 2016) with 32k merge operations. The vocabulary is shared among source and target languages. We select newstest2013 for validation and report the BLEU scores on newstest2014.

ZH-EN. The training data is from WMT19 which consists of 20.5M sentence pairs. The number of merge operations in byte pair encoding (BPE) is set to 32K for both source and target languages. We use newstest2018 as our validation set and newstest2019 as our test set, which contain 4k and 2k sentences, respectively.

4.2 Systems

Transformer. We implement our approach with the open source toolkit THUMT (Zhang et al., 2017) and strictly follow the setting of Transformer-Base in (Vaswani et al., 2017).

Exponential (Gu et al., 2020). This method adds an additional training weights to low-frequency target tokens:

$$w_j = A \cdot e^{-T \cdot Count(y_j)} + 1. \quad (5)$$

Chi-Square (Gu et al., 2020). The weighting function of this method is similar to the form of chi-square distribution

$$w_j = A \cdot Count^2(y_j) e^{-T \cdot Count(y_j)} + 1. \quad (6)$$

	B	S	EN-DE	ZH-EN
BMI	1.0	0.05	26.87	23.52
		0.10	26.89	**23.61**
		0.15	26.93	23.49
		0.20	26.98	23.39
		0.25	26.91	23.24
		0.30	26.85	23.50
	0.9	0.15	26.93	23.31
		0.20	26.88	23.31
		0.25	26.96	23.41
	0.8	0.15	**27.01**	23.40
		0.20	26.81	23.25
		0.25	26.93	23.50
	0.7	0.15	26.92	23.44
		0.20	26.90	23.35
		0.25	26.89	23.34

Table 2: Performance of our methods on the validation sets with different hyperparameters S and B.

BMI. Our system is first trained with normal cross entropy loss (Equation 1) for 100k steps. Then the model is further trained with BMI-based adaptive objective (Equation 4) for 100k steps. The same procedure was used for the competing methods. In order to eliminate the influence of noise, we assign the weight of tokens with BMI lower than 0.4 to zero during the training process.

4.3 Hyperparameters

We introduce two hyperparameters, B and S, to adjust the weight distribution based on BMI, as shown in Equation 4. In our experiments, we fixed B to narrow search space [0.7, 1]. We tuned another hyperparameter S on the validation sets. The results are shown in Table 2. Finally, we use the best hyperparameters found on the validation set for the final evaluation of the test set. For En-De, $B = 0.8$ and $S = 0.15$, for Zh-En, $B = 1.0$ and $S = 0.1$.

4.4 Main Results

As shown in Table 3, compared with (Vaswani et al., 2017), our Transformer outperforms it by 0.67 BLEU points. We use a strong baseline system in this work in order to make the evaluation convincing. Improvement of existing methods (Gu et al., 2020) is limited over strong baseline. Exponential objective achieves 28.17 (+0.2) BLEU on En-De and Chi-Square objective achieves 24.62 (+0.25) BLEU on Zh-En. Our method yields 28.53 (+0.56) and 25.19 (+0.82) BLEU on the En-De task and Zh-En task, respectively. The significant and consistent improvement on the two large-scale dataset demonstrates the effectiveness of our method.

System	EN-DE	ZH-EN
Existing NMT systems		
Vaswani et al. (2017)	27.3	-
Chi-Square	27.51	-
Exponential	27.60	-
Our NMT systems		
Transformer	27.97	24.37
+ Chi-Square	28.08(+*0.11*)	24.62(+*0.25*)
+ Exponential	28.17(+*0.20*)	24.33(-*0.04*)
+ BMI	**28.53(+*0.56*)***	**25.19(+*0.82*)***

Table 3: BLEU scores (%) on the WMT14 En-De test set and the WMT19 Zh-En test set. Results of our method marked with '*' are statistically significant (Koehn, 2004) by contrast to all other models ($p<0.01$).

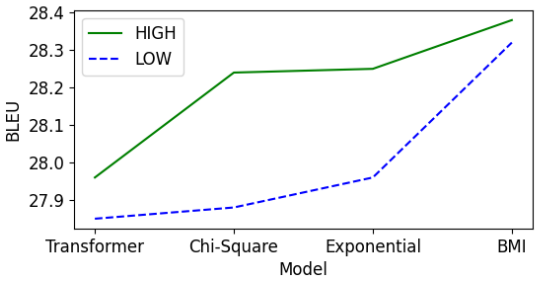

Figure 3: BLEU scores (%) on different WMT14 En-De test subsets which are grouped by their average BMI. Sentences in the HIGH subset contains more tokens with high BMI.

4.5 Results on Different BMI Intervals

We score each target sentence of newstest2014 by calculating the average BMI of each token in the sentence, and then divide newstest2014 into two subsets with equal size according to the score, denoted as HIGH and LOW, respectively. As shown in Figure 3, compared to Transformer, frequency-based methods outperform on the HIGH subset but have no obvious improvement on the LOW subset. By contrast, our method can not only bring a stable improvement on the HIGH subset, the improvement is even more obvious on the LOW subset. Low BMI means relatively rich mapping. We believe that the model should have a higher tolerance for these tokens because their translation errors may not be absolute. For example, the model outputs another token with similar meaning. Therefore, our method improves more on LOW subset.

Models	MATTR	HD-D	MTLD
Transformer	89.41	94.05	230.36
+ Chi-Square	89.37	94.02	230.02
+ Exponential	89.41	94.08	232.98
+ BMI	**89.45**	**94.10**	**236.43**
Reference	90.92	94.88	259.98

Table 4: The lexical diversity of WMT14 En-De translations measured by MATTR (%), HD-D (%) and MTLD. A larger value means a higher diversity.

4.6 Effects on Lexical Diversity

Vanmassenhove et al. (2019) suggest that the vanilla NMT systems exacerbate bias presented in corpus, resulting in lower vocabulary diversity. We use three measures of lexical diversity, namely, moving-average type-token ratio (MATTR) (Covington and McFall, 2010), the approximation of hypergeometric distribution (HD-D) and the measure of textual lexical diversity (MTLD) (McCarthy and Jarvis, 2010). Results in Table 4 show that, on improving the lexical diversity of translation, our method is superior to existing methods (Chi-Square and Exponential) based on word frequency.

4.7 Contrast with Label Smoothing

There are similarities between token-level adaptive training and label smoothing, because they both adjust the loss function of the model by token weighting. In particular, for some smoothing methods guided by prior or posterior knowledge of training data (Gao et al., 2020; Pereyra et al., 2017), different tokens are treated differently. But these similarities are not the key points of the two methods, and they are essentially different. The first and very important point is that the motivations of the two methods are different. Label smoothing is a regularization method to avoid overfitting, while our method treats samples of different difficulty differently for adaptive training. Second, the two methods work in different ways. Label smoothing is used when calculating the cross-entropy loss. It emphasizes how to assign the weight of tokens other than the golden one, and indirectly affects the training of the golden token. While our method is used after calculating the cross-entropy loss. It is calculated according to the golden token at each position in the reference, which is more direct. In all experiments, we employed uniform label smoothing of value $\epsilon_{ls} = 0.1$, the results show that the two methods does not conflict when used together.

5 Conclusion

We propose a novel bilingual mutual information based adaptive training objective, which can measure the learning difficulty for each target token from the perspective of bilingualism, and adjust the learning granularity dynamically to improve token-level adaptive training. Experimental results on two translation tasks show that our method can bring a significant improvement in translation quality, especially on sentences that are difficult to learn by the model. Further analyses confirm that our method can also improve the lexical diversity.

Acknowledgments

The research work descried in this paper has been supported by the National Key R&D Program of China (2020AAA0108001), the National Nature Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130), the Beijing Academy of Artificial Intelligence (BAAI2019QN0504) and the Youth Innovation Promotion Association CAS No. 2017172. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Yingbo Gao, Weiyue Wang, Christian Herold, Zijian Yang, and Hermann Ney. 2020. Towards a better understanding of label smoothing in neural machine translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 212–223.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046.

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 952–961.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference*, pages 2879–2885.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Yijin Liu, Fandong Meng, Jie Zhou, Yufeng Chen, and Jinan Xu. 2021. Faster depth-adaptive transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Fandong Meng and Jinchao Zhang. 2019. DTMT: A novel deep transition architecture for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 224–231.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112. Curran Associates, Inc.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Yu Wan, Baosong Yang, Derek F Wong, Yikai Zhou, Lidia S Chao, Haibo Zhang, and Boxing Chen. 2020. Self-paced learning for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1074–1080.

Jianhao Yan, Fandong Meng, and Jie Zhou. 2020. Multi-unit transformers for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1047–1059, Online.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017. Thumt: An open source toolkit for neural machine translation. *arXiv preprint arXiv:1706.06415*.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343.

George Kingsley Zipf. 1949. Human behavior and the principle of least effort.