Focus on the Target's Vocabulary: Masked Label Smoothing for Machine Translation

Liang Chen, Runxin Xu, Baobao Chang*
Key Laboratory of Computational Linguistics, Peking University, MOE, China
leo.liang.chen@outlook.com
runxinxu@gmail.com chbb@pku.edu.cn

Abstract

Label smoothing and vocabulary sharing are two widely used techniques in neural machine translation models. However, we argue that simply applying both techniques can be conflicting and even leads to sub-optimal performance. When allocating smoothed probability, original label smoothing treats the source-side words that would never appear in the target language equally to the real target-side words, which could bias the translation model. To address this issue, we propose Masked Label Smoothing (MLS), a new mechanism that masks the soft label probability of source-side words to zero. Simple yet effective, MLS manages to better integrate label smoothing with vocabulary sharing. Our extensive experiments show that MLS consistently yields improvement over original label smoothing on different datasets, including bilingual and multilingual translation from both translation quality and model's calibration. Our code is released at PKUnlp-icler.

1 Introduction

Recent advances in Transformer-based (Vaswani et al., 2017) models have achieved remarkable success in Neural Machine Translation (NMT). For most NMT studies (Vaswani et al., 2017; Song et al., 2019; Lin et al., 2020; Pan et al., 2021), there are two widely used techniques to improve the quality of the translation: Label Smoothing (LS) and Vocabulary Sharing (VS). Label smoothing (Pereyra et al., 2017) turns the *hard* one-hot labels into a *soft* weighted mixture of the golden label and the uniform distribution over the whole vocabulary, which serves as an effective regularization technique to prevent over-fitting and over-confidence (Müller et al., 2019) of the model. In addition, vocabulary sharing (Xia et al., 2019) is another commonly used technique, which unifies the vocabulary of both source and target language into a whole vocabulary, and therefore the vocabulary is shared. It enhances the semantic correlation between the two languages and reduces the number of total parameters of the embedding matrices.

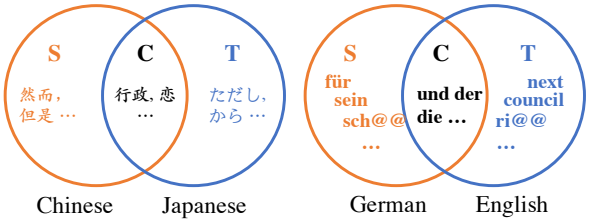

Figure 1: Venn diagram showing the structure of the shared vocabulary, which can be divided into three parts: Source (S), Common (C), and Target (T).

Model	DE-EN	VI-EN
Transformer	33.54	29.95
- w/ Label Smoothing (LS)	**34.76**	**30.73**
- w/ Vocabulary Sharing (VS)	33.83	29.36
- w/ LS+VS †	34.56	30.41

Table 1: Results in IWSLT'14 DE-EN and IWSLT'15 VI-EN datasets.† denotes consistent setting to (Vaswani et al., 2017). Jointly adopting label smoothing and vocabulary sharing techniques cannot achieve further improvements, but leads to sub-optimal performance.

However, in this paper, we argue that jointly adopting both label smoothing and vocabulary sharing techniques can be conflicting, and leads to sub-optimal performance. Specifically, with vocabulary sharing, the shared vocabulary can be divided into three parts as shown in Figure 1. But with label smoothing, the soft label still considers the words at the source side that are impossible to appear at the target side. This would mislead the translation model and exerts a negative effect on the translation performance. As shown in Table 1, although introducing label smoothing or vocabulary sharing alone can improve the vanilla Transformer, jointly

*Corresponding author

adopting both of them cannot obtain further improvements but achieves sub-optimal results.

To address the conflict of label smoothing and vocabulary sharing, we first propose a new mechanism named Weighted Label Smoothing (WLS) to control the smoothed probability distribution and its parameter-free version Masked Label Smoothing (MLS). Simple yet effective, MLS constrains the soft label not to assign soft probability to the words only belonging to the source side. In this way, we not only keeps the benefits of both label smoothing and vocabulary sharing, but also address the conflict of these two techniques to improve the quality of the translation.

According to our experiments, MLS leads to a better translation not only in scores like BLEU but also reports improvement in model's calibration. Compared with original label smoothing with vocabulary sharing, MLS outperforms in WMT'14 EN-DE(+0.47 BLEU), WMT'16 EN-RO (+0.33 BLEU) and other 7 language pairs including DE,RO-EN multilingual translation task.

2 Background

Label Smoothing The original label smoothing can be formalized as:

$$\hat{\boldsymbol{y}}^{LS} = \hat{\boldsymbol{y}}(1-\alpha) + \boldsymbol{\alpha}/K \tag{1}$$

K denotes the number of classes, α is the label smoothing parameter, $\boldsymbol{\alpha}/K$ is the soft label, $\hat{\boldsymbol{y}}$ is a vector where the correct label equals to 1 and others equal to zero and $\hat{\boldsymbol{y}}^{LS}$ is the modified targets.

Label smoothing is first introduced to image classification (Szegedy et al., 2016) task. Pereyra et al. (2017); Edunov et al. (2018) explore label smoothing's application in Sequence generation from token level and Norouzi et al. (2016) propose sentence level's label smoothing. Theoretically, Müller et al. (2019); Meister et al. (2020) all point out the relation between label smoothing and entropy regularization. Gao et al. (2020) explores the best recipe when applying label smoothing to machine translation. To generate more reliable soft labels, Lukasik et al. (2020) takes semantically similar n-grams overlap into consideration level label smoothing. Wang et al. (2020) proposes Graduate Label Smoothing that generate soft label according to the different confidence scores of model. To the best of our knowledge, we are the first to investigate label smoothing's influence on machine translation from the perspective of languages.

Category	DE->EN	RO->EN	VI->EN
Source	39%	50%	36%
Common	20%	8%	11%
Target	41%	42%	53%

Table 2: The distribution of different categories of the shared vocabulary forWMT'14 DE-EN, WMT'16 RO-EN, and IWSLT'15 VI-EN datasets. The proportion of tokens belonging to source category is up to 50%, which might mislead the translation model.

Vocabulary Sharing Vocabulary sharing is widely applied in most neural machine translation studies (Vaswani et al., 2017; Song et al., 2019; Lin et al., 2020). Researchers have conducted in-depth studies in Vocabulary Sharing. Liu et al. (2019) propose shared-private bilingual word embeddings, which give a closer relationship between the source and target embeddings. While Kim et al. (2019) point out that there is an vocabulary mismatch between parent and child languages in shared multilingual word embedding.

3 Conflict Between Label Smoothing and Vocabulary Sharing

Words or subwords in a language pair's joint dictionary can be categorized into three classes: **source**, **common** and **target** using Venn Diagram according to their belonging to certain language as depicted in Figure 1. This can be achieved by checking whether one token in the joint vocabulary also belongs to the source/target vocabulary. We formalized the categorization algorithm in Appendix A.

Then we compute the tokens' distribution in different translation directions as shown in Table 2. Tokens in source class account for a large proportion up to 50%. When label smoothing and vocabulary sharing are together applied, the smoothed probability will be allocated to words that belong to the source class. Those words have zero overlap with the possible target words, therefore they have no chance to appear in the target sentence. Allocating smoothed probability to them might introduce extra bias for the translation system during training process, unavoidably leading to a higher translation perplexity as also revealed by Müller et al. (2019).

Table 3 reveals the existence of conflict, that the joint use of label smoothing and vocabulary sharing doesn't compare with solely use one technique in all language pairs with a maximum loss of 0.32 BLEU score.

4 Methods

4.1 Weighted Label Smoothing

To deal with the conflict when executing label smoothing, we propose a plug-and-play Weighted Label Smoothing mechanism to control the smoothed probability's distribution.

Weighted Label Smoothing(WLS) has three parameters β_t, β_c, β_s apart from the label smoothing parameter α, where the ratio of the three parameters represents the portion of smoothed probability allocated to the target, common and source class and the sum of the three parameters is 1. The distribution within token class follows a uniform distribution. WLS can be formalized as:

$$\hat{\boldsymbol{y}}^{WLS} = \hat{\boldsymbol{y}}(1-\alpha) + \boldsymbol{\beta} \tag{2}$$

where $\hat{\boldsymbol{y}}$ is a vector where the element corresponding to the correct token equals to 1 and others equal to zero. $\boldsymbol{\beta}$ is a vector that controls the distribution of probability allocated to incorrect tokens. We use t_i, c_i, s_i to represent probability allocated to the i-th token in the target,common,source category, all of which form the distribution controlling vector $\boldsymbol{\beta}$ with $\sum_i^K \beta_i = \alpha$. The restriction can be formalized as:

$$\sum t_i : \sum c_i : \sum s_i = \beta_t : \beta_c : \beta_s \tag{3}$$

4.2 Masked Label Smoothing

Based on the Weight Label Smoothing mechanism, we can now implement Masked Label Smoothing by set β_s to 0 and regard the target and common category as one category. In this way, Masked Label Smoothing is parameter-free and implicitly injects external knowledge to the model. And we have found out that this simple setting can reach satisfactory results according our experiments.

We illustrate different label smoothing methods in Figure 2. It is worth noticing that MLS is different from setting WLS's parameters to 1-1-0 since there might be different number of tokens in the common and target vocab.

5 Experiments

5.1 Task Settings

For bilingual translation, we conduct experiments on 7 translation tasks. We choose language pairs that have different ratio of common subwords. These include WMT'14 DE-EN,EN-DE, IWSLT'14 DE-EN, IWSLT'15 VI-EN, WMT'16 RO-EN,EN-RO and CASIA ZH-EN.

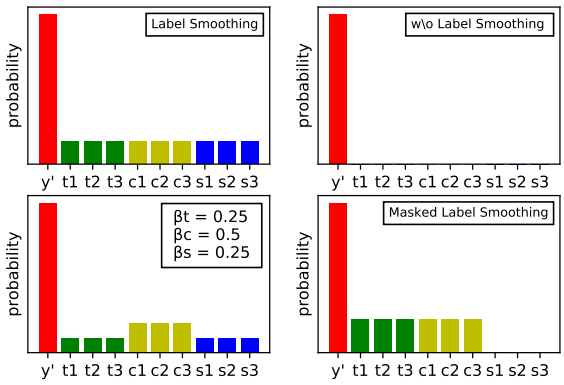

Figure 2: Illustration of different label smoothing methods. The height of each bar in the graph denoted the probability allocated to each token. y' is the current token during current decoding phase. We assume that there are only 10 tokens in the joint vocabulary and t1-t3 belongs to target class, c1-c3 belongs to common class and s1-s3 belongs to source class.

We use the official train-dev-test split of WMT'14, 16 and IWSLT'14, 15 datasets. For CASIA ZH-EN dataset, we randomly select 5000 sentences as development set and 5000 sentences as test set from the total dataset.

For multilingual translation, we combine the WMT'16 RO-EN and IWSLT'14 DE-EN datasets to formulate a RO,DE-EN translation task. We also make a balanced multilingual dataset that has equal numbers of DE-EN and RO-EN training examples to reduce the impact of imbalance languages and to explore how MLS performs under different data distribution condition in multilingual translation.

We apply the Transformer base (Vaswani et al., 2017) model as our baseline model. We fix the label smoothing parameter α to 0.1 in the main experiments and individually experiment and examine the performance of MLS under different α.

We use compound_split_bleu.sh from fairseq to compute the final bleu scores. The inference ECE score[1] and chrF score[2] are computed through open source scripts. We list the concrete training and evaluation settings in Appendix B.

5.2 Results

Bilingual Table 3 shows the results of bilingual translation experiments. The results reveal the conflict between LS and VS that models with only LS

[1]https://github.com/shuo-git/InfECE
[2]https://github.com/m-popovic/chrF

(a) Bilingual Translation

	WMT'16		IWSLT'14	WMT'14		IWSLT'15	CASIA
Model	RO-EN	EN-RO	DE-EN	DE-EN	EN-DE	VI-EN	ZH-EN
Transformer	22.03	19.61	33.54	30.85	27.21	29.95	20.66
- w/ VS	22.20	19.91	33.83	31.08	27.51	29.36	20.88
- w/ LS	22.96	20.68	34.76	31.14	27.53	**30.73**	21.10
- w/ LS+VS	22.89	20.59	34.56	30.98	27.44	30.41	21.04
- w/ MLS (ours)	**23.22****	**20.88****	**35.04****	**31.43***	**27.91***	30.57*	**21.23***

(b) Multilingual Translation

	IWSLT'14+WMT'16			IWSLT'14+WMT'16†		
Model	**DE,RO-EN**	**DE-EN**	**RO-EN**	**DE,RO-EN**	**DE-EN**	**RO-EN**
- w/ LS+VS	33.78	37.24	23.15	33.25	37.44	20.40
- w/ MLS (ours)	**34.10****	**37.53****	**23.19**	**33.53****	**37.77****	**20.86****

Table 3: Results of bilingual translation tasks (a) and multilingual translation (b). † denotes the balanced version of multilingual translation data. Same conflict between LS and VS occurs in all language pairs. Our MLS outperforms the original label smoothing with vocabulary sharing with significance levels when of $p < 0.01$ (**), $p < 0.05$ (*) and also beats individually using LS or VS in most cases.

surpass models with both LS and VS in all experiments. Our Masked Label Smoothing obtained consistent improvements over original LS+VS in all tested language pairs significantly.

The effectiveness of MLS maintained under different α value as shown in Table 4 for both BLEU and chrF scores. Similar to Gao et al. (2020)'s conclusion, we find that a higher α can generally improve the bilingual translation quality. And applying MLS can further improve the results. It shows that not only the probability increase in target vocabulary, but also the allocation of smoothed probabilities in different languages matters in the improvement of translation performance.

(a) EN-RO

Scores	BLEU(chrF)		
α	0.1	0.3	0.5
LS+VS	20.54(45.54)	20.65(45.79)	20.62(45.7)
MLS	**20.57(45.68)**	**20.99(46.29)**	**21.10(46.4)**

(b) RO-EN

Scores	BLEU(chrF)		
α	0.1	0.3	0.5
LS+VS	22.54(47.09)	22.95(47.29)	22.98(47.23)
MLS	**22.89(48.23)**	**23.10(48.36)**	**23.07(47.39)**

Table 4: Individual experiment on α. BLEU and chrF scores are reported under different label smoothing α on WMT'16 EN-RO (a) and RO-EN (b) datasets.

Multilingual As shown in Table 3, MLS achieves consistent improvement over the original label smoothing in both the original and the balanced multilingual translation dataset under all translation directions. In the original combined dataset, direction RO-EN (400K) has much more samples than DE-EN (160K). We do not apply a resampling strategy during training in order to investigate how the imbalance condition affects different models' performance. The balanced version cuts down samples in RO-EN direction to the same number as in DE-EN direction.

Compared with the imbalance version, the balanced version gave better BLEU scores in DE-EN direction while much worse performance in RO-EN translation for both the original label smoothing and MLS. It indicates that the cut down on RO-EN training examples does weaken the generalization of model in RO-EN translation however doesn't influence the DE-EN translation quality since the RO-EN data might introduce bias to the training process for DE-EN translation.

Even under imbalance condition, MLS can give a better performance (37.53) compared to original LS in the balance condition (37.44). It implies that MLS can relieve the imbalance data issue in multilingual translation. However, the improvement in relative high-resources direction (RO-EN) is not as significant as in the balanced condition. We guess that label smoothing has more complex influence on multilingual model due to the increase of languages and relation among different languages. We leave those questions for future exploration.

β_t	β_c	β_s	RO-EN	EN-RO	DE-EN
-	-	-	22.80	23.15	30.94
1/3	1/3	1/3	22.68	23.19	**31.40**
1/2	1/2	0	**23.05**	23.19	31.18
1/2	0	1/2	22.86	23.01	31.33
0	1/2	1/2	22.22	**23.33**	30.85
1/2	1/4	1/4	22.73	23.16	30.92

Table 5: Value "-" denotes the original label smoothing. WLS generally can improve the translation quality with appropriate parameters. Scores are computed using the development set of each direction.

6 Discussion

6.1 Exploring of Weighted Label Smoothing

As reported in Table 5, we explore the influence of different WLS on multiple tasks including WMT'16 RO-EN,EN-RO and WMT'14 DE-EN.

According to the result, though the best BLEU score's WLS setting vary from different tasks and there seems to exist a more complex relation between the probability allocation and the BLEU score, we still have two observations. First, applying WLS can generally boost the quality of translation compared to the original label smoothing. Second, only WLS with β_t, β_c, β_s each equals to 1/2-1/2-0 can outperform the original label smoothing on all tasks, which suggests the setting is the most robust one. Thus we recommend using this setting as the initial setting when applying WLS.

Furthermore, the most robust setting agrees with the form of MLS since they both allocate zero probability to the source category's tokens, which further proves the robustness of MLS.

6.2 Improvement in Model's Calibration and Translation Perplexity

Müller et al. (2019) have pointed out label smoothing prevents the model from becoming overconfident therefore improve the calibration of model. Since there is a training-inference discrepancy in NMT models, inference ECE score (Wang et al., 2020) better reflects models' real calibration.

To compute the ECE scores, we need to split the model's predictions into M bins according to the output confidence and calculate the weighted average of bin's confidence/accuracy difference as the ECE scores considering the number of samples in each bin.

$$ECE = \sum_{i=1}^{M} \frac{|B_i|}{N} \left|\text{acc}\left(B_i\right) - \text{confidence}\left(B_i\right)\right|$$

where N is the number of total prediction samples and B_i is the number of samples in the i-th bin. $\text{acc}\left(B_i\right)$ is the average accuracy in the i-th bin.

The score denotes the difference between accuracy and confidence of models' output during inference. Less ECE implies better calibration.

The inference ECE scores of our models are shown in Table 6. It turns out that models with MLS have lower Inference ECE scores on different datasets. The results indicate that MLS will lead to better model calibration.

We also find out that MLS leads to a significantly lower perplexity than LS during the early stage of training in all of our experiments. It's not surprising since zeroing the source side words' smoothed probability can decrease the perplexity. It can be another reason for model's better translation performance since it gives a better training initialization.

Model	DE-EN	VI-EN	DE,RO-EN	DE,RO-EN*
- w/ LS+VS	9.77	13.07	11.62	10.77
- w/ MLS	**9.67**	**12.63**	**11.37**	**8.82**

Table 6: Inference ECE score (less is better) on different translation tasks. $*$ denotes the balanced version of multilingual data. MLS leads to an average of 0.7 lower ECE score, suggesting better model calibration.

7 Conclusion

We reveal the conflict between label smoothing and vocabulary sharing techniques in NMT that jointly adopting the two techniques can lead to suboptimal performance. To address this issue, we introduce Masked Label Smoothing to eliminate the conflict by reallocating the smoothed probabilities according to the languages' differences. Simple yet effective, MLS shows improvement over original label smoothing from both translation quality and model's calibration on a wide range of tasks.

8 Acknowledgements

We thank all reviewers for their valuable suggestions for this work. This paper is supported by the National Science Foundation of China under Grant No.61876004 and 61936012, the National Key Research and Development Program of China under Grant No. 2020AAA0106700.

9 Ethics Consideration

We collect our data from public datasets that permit academic use. The open-source tools we use for training and evaluation are freely accessible online without copyright conflicts.

References

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *NAACL*.

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, Suzhou, China. Association for Computational Linguistics.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Xuebo Liu, Derek F. Wong, Yang Liu, Lidia S. Chao, Tong Xiao, and Jingbo Zhu. 2019. Shared-private bilingual word embeddings for neural machine translation. In *ACL*.

M. Lukasik, Himanshu Jain, A. Menon, Seungyeon Kim, Srinadh Bhojanapalli, F. Yu, and Sanjiv Kumar. 2020. Semantic label smoothing for sequence to sequence problems. In *EMNLP*.

Clara Meister, Elizabeth Salesky, and Ryan Cotterell. 2020. Generalized entropy regularization or: There's nothing special about label smoothing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6870–6886, Online. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *NeurIPS*.

Mohammad Norouzi, Samy Bengio, zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. Reward augmented maximum likelihood for neural structured prediction. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of ACL 2021*.

Gabriel Pereyra, G. Tucker, J. Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions.

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, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*.

Christian Szegedy, V. Vanhoucke, S. Ioffe, Jonathon Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Yingce Xia, Tianyu He, Xu Tan, Fei Tian, Di He, and Tao Qin. 2019. Tied transformers: Neural machine translation with shared encoder and decoder. In *AAAI*.

A Algorithm

Algorithm 1 Divide Token Categories

Input: List: S, T, J
Output: List: A,B,C
Description: S is the vocabulary list for source language, T for target language, J for joint vocabulary. A is the output vocabulary for source tokens, B for common tokens, C for target tokens.

```
Initialize empty list A,B,C
for i in J do
    if i in S and i in T then
        B.add(i)
    else
        if i in S then
            A.add(i)
        else
            C.add(i)
return A,B,C
```

B Experiment Details

We evaluate our method upon Transformer-Base (Vaswani et al., 2017) and conduct experiments under same hyper-parameters for fair comparison.

Before training, we first apply BPE(Sennrich et al., 2016) to tokenize the corpus for 16k steps each language and then learn a joint dictionary. During training, the label smoothing parameter α is set to 0.1 except for Table 4's exploration in alpha values. We use Adam optimizer with betas to be (0.9,0.98) and learning rate is 0.0007. During warming up steps, the initial learning rate is 1e-7 and there are 1000 warm-up steps. We use a batch-size of 2048 together with an update-freq of 4 on two NVIDIA 3090 GPUs. Dropout rate is set to 0.3 and weight decay is set to 0.0001 for all experiments. We average the last 3 checkpoints to generate the final model in the main bilingual experiments before inferring on the test set. We use beam size as 5 during all testing.