HW-TSC's Participation in the IWSLT 2022 Isometric Spoken Language Translation

Zongyao Li, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Minghan Wang,
Ting Zhu, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Lizhi Lei, Hao Yang, Ying Qin
Huawei Translation Service Center, Beijing, China
{lizongyao,guojiaxin1,weidaimeng,shanghengchao,wangminghan,
zhuting20,wuzhanglin2,yuzhengzhe,chenxiaoyu35,
leilizhi,yanghao30,qinying}@huawei.com

Abstract

This paper presents our submissions to the IWSLT 2022 Isometric Spoken Language Translation task. We participate in all three language pairs (English-German, English-French, and English-Spanish) under the constrained setting, and submit an English-German result under the unconstrained setting. We use the standard Transformer model as the baseline and obtain the best performance via one of its variants that shares the decoder input and output embedding. We perform detailed pre-processing and filtering on the provided bilingual data. Several strategies are used to train our models, such as Multilingual Translation, Back Translation, Forward Translation, R-Drop, Average Checkpoint, and Ensemble. We experiment on three methods for biasing the output length: i) conditioning the output to a given target-source length-ratio class; ii) enriching the transformer positional embedding with length information and iii) length control decoding for non-autoregressive translation etc. Our submissions achieve 30.7, 41.6 and 36.7 BLEU respectively on the tst-COMMON test sets for English-German, English-French, English-Spanish tasks and 100% comply with the length requirements.

1 Introduction

This paper introduces our submissions to the IWSLT 2022 Isometric Spoken Language Translation task. To train our models, we perform multiple data filtering strategies to enhance data quality. In addition, we leverage Multilingual model (Johnson et al., 2017), Forward (Wu et al., 2019) and Back Translation (Edunov et al., 2018), and R-Drop (Wu et al., 2021) strategies to further enhance training effects. We also adopt Length Token (Lakew et al., 2019), Length Encoding (Takase and Okazaki, 2019) and Non-Autoregressive Translation (NAT) to further enhance system performances. We compare and contrast different strategies in light of our experiment results and conduct analysis accordingly.

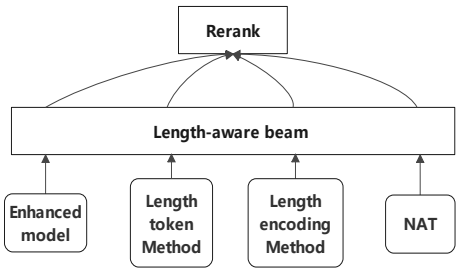

Figure 1: The training process for the IWSLT 2022 Isometric Spoken Language Translation.

The overall training process is illustrated in Figure 1. Section 2 focuses on our training techniques, including model architecture, data processing and training strategies. Section 3 describes our experiment settings and training process. Section 4 presents the experiment results while section 5 analyzes the effects of different model enhancement and length control strategies on the quality and length of translation outputs.

2 Method

2.1 Model Architecture

2.1.1 Autoregressive NMT Model

Transformer-based model with the self-attention mechanism (Vaswani et al., 2017) has achieved the state-of-the-art translation performance. The Transformer architecture is a standard encoder-decoder model. The encoder can be viewed as a stack of N layers, including a self-attention sub-layer and a feed-forward (FFN) sub-layer. The decoder shares a similar architecture as the encoder but integrates an encoder-decoder attention sub-layer to capture the mapping between two languages.

For autoregressive translation (AT) models we trained in this shared task, Transformer-Base architecture is used, which features 6-layer encoder, 6-layer decoder, 512 dimensions of word vec-

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 361 - 368

tor, 2048-hidden-state, 8-head self-attention, post-norm, share decoder input, and output embedding.

2.1.2 Non-autoregressive NMT Model

Non-autoregressive models generate all outputs in parallel and break the dependency between output tokens. For AT models, EOS (end of sentence) token is used to indicate the end of a sentence and thus determines the length of the sequence. On the contrary, for NAT models, the output length should be predicted in advance. We believe such mechanism is more suitable for this task.

CMLM (Ghazvininejad et al., 2019) adopts a masked language model to progressively generate the sequence from entirely masked inputs and has achieved stunning performance among non-autoregressive NMT models. HI-CMLM (Wang et al., 2021a) extends CMLM using a novel heuristic hybrid strategy, i.e. fence-mask, to improve the translation quality of short texts and speed up early-stage convergence. In the constrained task, HI-CMLM is used, which features 6-layer encoder, 6-layer decoder, 512 dimensions of word vector, 1024-hidden-state, and 4-head self-attention.

AT and NAT models have distinctive superiorities and drawbacks in terms of performance and latency. We try to combine the two strategies into one model, hoping to leverage advantages of both. Diformer (Wang et al., 2021b) (Directional Transformer), with a newly introduced direction variable, is a unified framework that jointly models Autoregressive and Non-autoregressive settings into three generation directions (left-to-right, right-to-left and straight). It works by controlling the prediction of each token to have specific dependencies under that direction. In the unconstrained task, Diformer is used, which features 6-layer encoder, 6-layer decoder, 512 dimensions of word vector, 2048-hidden-state, and 8-head self-attention.

2.2 Data Processing and Augmentation

As for the constrained task, we use only the officially provided data, MuST-C v1.2. As for the unconstrained task, we additionally apply WMT2014 data to the English-German task for NAT model training.

2.2.1 Data Filtering

We perform the following steps to cleanse all data:

- Filter out repeated sentences (Khayrallah and Koehn, 2018; Ott et al., 2018).
- Convert XML escape characters.
- Normalize punctuations using Moses (Koehn et al., 2007).
- Delete HTML tags, non-UTF-8 characters, unicode characters and invisible characters.
- Filter out sentences with mismatched parentheses and quotation marks; sentences of which punctuation exceeds 30%; sentences with the character-to-word ratio greater than 12 or less than 1.5; sentences of which the source-to-target token ratio higher than 3 or lowers than 0.3; sentences with more than 120 tokens.
- Apply langid (Joulin et al., 2016b,a) to filter sentences in other languages.
- Use fast-align (Dyer et al., 2013) to filter sentence pairs with poor alignment, and about 10% of the data is filtered out.

Language pair	Raw data	Data filtering
en-de	229.7K	211.1K
en-fr	275.1K	253.9K
en-es	265.6K	247.8K

Table 1: Data sizes before and after filtering.

Data sizes before and after filtering are listed in Table 1.

2.2.2 Data Diversification

Nguyen et al. (2020) introduce Data Diversification, a simple but effective strategy to enhance neural machine translation (NMT) performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging the generated text with the original dataset on which the final NMT model is trained.

In terms of back translation, we adopt top-k sampling to translate data (BT sampling). With regard to forward translation, we translate data using beam search. Through sampling, we ensure that the sizes of data generated by forward and back translation are relatively equal. In this paper, we refer to the combination of forward and backward translation sampling as FBTS.

Inspired by Iterative Joint Training (Zhang et al., 2018), we first adopt multiple copies of BT sampling data for model training in this task. Then, we further perform model augmentation training by

merging multiple copies of FBTS data generated by the optimized model with the authentic bilingual data. Since model performance (Zhang et al., 2019) will be affected due to length control, we generate a great amount of synthetic parallel data to enrich data diversity, in hope of minimizing the effect of length control.

2.2.3 Data Distillation and Self-Distillation Mixup Training

Knowledge distillation trains a student model to perform better by learning from a stronger teacher model. This method has been proved effective for NAT models training by Zhou et al. (2019). In this work, we use enhanced AT models as teacher models to generate distilled data, and use self-distillation mixup training (Guo et al., 2021) strategy to train the NAT student models.

2.3 Model Augmentation

2.3.1 Multilingual Model

Johnson et al. (2017) proposes a simple solution that uses a single neural machine translation model to translate across multiple languages, without architecture changes. The model introduces an artificial token at the beginning of the input sentence to specify the required target language. All languages use a shared vocabulary. No additional parameters are required. The experiments surprisingly show that such model design can achieve better translation qualities across languages. In the task, we use only constrained data of the particular language pair for training. Taking en2de as an example, we use only English-to-German and German-to-English data.

2.3.2 R-Drop Training

R-Drop (Wu et al., 2021) uses a simple dropout twice method to construct positive samples for comparative learning, significantly improving the experimental results in supervised tasks. We apply R-Drop with a = 5 to regularize the model so as to prevent over-fitting.

2.3.3 Ensemble

Model ensemble is a widely used technique in previous WMT workshops (Garmash and Monz, 2016), which enhances the performance by combining the predictions of several models at each decoding step. We train multiple models (generally four models) by shuffling training data and perform ensemble decoding with the above models in the inference phase.

2.4 Output Length Control

As described in the task, we define length compliance (LC) as the percentage of translations in a given test set falling in a predefined length threshold of ±10% of the number of characters in the source sentence.

2.4.1 Length Token

Lakew et al. (2019) classify bi-text into three classes based on the target-to-source character ratio (LR) of each sample (s; t) pair. The labels are defined based on LR thresholds: $short < 0.9 < normal < 1.1 < long$ in our experiment. We prepend the length token $v \epsilon \{short; normal; long\}$ at the beginning of the source sentence during training. The desired v is prepended on the input sentence during inference.

2.4.2 Length Encoding

Takase and Okazaki (2019) propose a simple but effective extension of sinusoidal positional encoding to constrain the length of outputs generated by a neural encoder-decoder model. We adopt the length-ratio positional encoding (LRPE) method mentioned in the paper. LRPE is expected to generate sentences of any length even if sentences of exact lengths are not included in the training data.

2.4.3 Length-control decoding for NAT

Traditional NAT models predict the output token numbers first and then generate all output tokens in parallel. Some prior work (Wang et al., 2021c) has analyzed how length prediction influences the performance of NAT. To further improve the length compliance, we propose length-control decoding (LCD), which sets the length of the target tokens as that of the source tokens. We assume that if the source and target sentences have the same number of tokens, their sentence lengths are also approximately the same.

2.4.4 Length-aware beam

In order to get better translation results, we generate n-best hypotheses with a multi-model ensemble. In this task, beam-size is set to 12, so that 12 candidate outputs are generated for one source sentence, among which we select the one that comply with the +-10% length requirements. The candidate output with the least loss value is selected when all

the 12 outputs fail to meet the length requirement. This method is called length-aware beam (LAB).

2.4.5 Rerank

We try various strategies in our experiments. With LAB strategy, each model has its own trade off on quality and length control. We ensemble several models of which BLEU is better on tst-COMMON test sets to score all the candidate outputs. Based on the scores, we rerank the candidates to select the best one.

3 Settings

3.1 Experiment Settings

We use the open-source fairseq (Ott et al., 2019) for training. BERTScore is used to measure system performances and the script officially provided is used to calculate the output lengths in the task. Each model is trained using 8 GPUs. The size of each batch is set to 2048, parameter update frequency to 2, and learning rate to 5e-4. The number of warmup steps is 4000, and the dropout is 0.3. We share vocabulary for source and target languages, and sizes of the vocabularies for English-German, English-French and English-Spanish are 30k, 27k, and 30k respectively. We use early stopping when validation loss stops improving and apply checkpoint averaging on last 5 checkpoints. In the inference phase, the beam-size is 12 and the length penalty is set to 0.6.

3.2 System Process

Our overall training strategy is to train a baseline model, conduct enhanced training with techniques such as multilingual translation, R-Drop, and data augmentation. After obtaining the optimized model, we add length token to the training data, adopt length encoding to the model, and use non-autoregressive decoding to control the output length. In addition, we ensemble multiple models to achieve the submitted results. Our training process is as follows:

1) We preprocess the training data using methods mentioned in section 2.2.1 and train four models using Multilingual Translation and R-Drop strategies with shuffled training data.

2) We perform data augmentation as described in section 2.2.2. We train four models with bilingual data and BT sampling data generated by the models mentioned in step 1. Then, we perform FBTS data augmentation on the basis of the enhanced models and train four more models. For the constrained setting, we use both source and target sides of the bilingual data to generate four copies of forward and backward translated pseudo bi-texts (one model generates one copy), respectively.

3) We add length token to authentic and synthetic parallel data as described in section 2.4.1, and train four models to ensemble. We also train a model using length encoding, as mentioned in section 2.4.2.

4) We train the NAT models using the method described in section 2.4.3 with authentic bilingual data and synthetic parallel data generated in step 2).

5) We average the last five checkpoints and perform separate inference on each model, and then ensemble the models. We change length token ($long, normal, short$) for models using Length Token strategy to generate multiple results.

6) We use the method described in section 2.4.4 and 2.4.5 rerank hypotheses generated from models trained by different strategies to get the final results.

4 Experiment Result

Table 2 lists the results of our submissions on the tst-COMMON test sets. The baseline models, trained on transformer-base architecture, achieve the poorest performances on BLEU and rather poor performance on LC. Our enhanced models (Enhanced), trained with data and model augmentation strategies, achieve the highest BLEU scores (33.3, 45.9, 37.1) but the lowest LC scores (36.9, 36.6, 57.9) on the three language pairs. Len-tok models are trained with Length Token strategy and the length token is set to $normal$, and an improvement on LC has been witnessed. Len-control decoding for nat models uses NAT Decoding. Length-aware beam strategy is demonstrated useful for all of the three types of models as we witness significant improvements on LC for those models by using the strategy. Rerank1 reranks hypotheses from the enhanced and Len-tok models; Rerank2 reranks hypotheses from the enhanced and len-control decoding for nat models; and Rerank3 reranks hypotheses from all of the three types of models. Accord-

Pairs	English-German				English-French				English-Spanish			
System	BLEU	F1	LR	LC	BLEU	F1	LR	LC	BLEU	F1	LR	LC
Baseline	28.9	0.828	1.12	41.0	35.6	0.812	1.22	33.1	30.5	0.809	1.11	44.0
Enhanced	**33.3**	**0.842**	1.14	36.9	**45.9**	**0.872**	1.14	36.6	**37.1**	**0.850**	1.04	57.9
+LAB	33.0	0.838	1.10	68.6	45.4	0.869	1.13	50.5	36.9	0.848	1.03	72.1
Len-tok	32.1	0.835	1.06	54.7	44.1	0.866	1.09	49.1	36.8	0.848	1.02	66.8
+LAB	31.2	0.830	**1.04**	80.8	42.9	0.859	1.07	73.s1	37.1	0.845	1.01	84.2
NAT	30.4	0.829	1.04	83.5	42.3	0.848	1.05	82.3	36.1	0.830	1.01	89.9
+LAB	29.8	0.826	1.05	**89.0**	41.6	0.848	1.05	**87.3**	35.9	0.833	1.01	**93.7**
Rerank1	30.7	0.830	1.03	99.8	41.5	0.851	1.03	98.7	36.8	0.845	1.01	98.9
Rerank2	29.9	0.829	1.02	100	40.9	0.849	1.02	100	36.0	0.844	1.01	100
Rerank3	30.7	0.830	1.04	**100**	41.6	0.851	**1.02**	**100**	36.7	0.845	**1.01**	**100**

Table 2: Experimental results of our submitted system. (F1 is short for BERTScore F1.)

Pairs	English-German				English-French				English-Spanish			
System	BLEU	F1	LR	LC	BLEU	F1	LR	LC	BLEU	F1	LR	LC
Enhanced	33.0	0.838	1.10	68.6	45.4	0.869	1.13	50.5	36.9	0.848	1.03	72.1
LT-normal	31.2	0.830	1.04	80.8	42.9	0.859	1.07	73.1	37.1	0.845	1.01	84.2
LT-short	27.2	0.818	0.94	82.0	38.0	0.845	0.98	85.3	36.3	0.841	0.95	83.3
LT-long	32.6	0.839	1.15	45.4	44.9	0.864	1.17	42.8	35.0	0.844	1.07	66.1
LRPC	28.0	0.822	1.06	79.3	40.6	0.843	1.04	78.7	34.8	0.842	1.00	90.5

Table 3: The experimental results of length token and encoding method.

ing to our experiment results, Rerank3 achieves the best BLEU and BERTScore scores and 100% comply with the length requirement. For details about the blind-test results submitted, see appendix A.

5 Analysis

5.1 Data Augmentation and Model Augmentation to Enhance Model Performance

Our experiment results demonstrate that model augmentation has positive effects on model performances. Table 4 lists the BLEU scores on the tst-COMMON test sets. Compared with the baseline models, other models obtain much higher BLEU on English-German, English-French and English-Spanish tasks. Our experiment on English-German task shows that strategies such as multilingual translation, decoder input and output embedding (Tied-embed) sharing, R-Drop, BT sampling, and FBTS, have significant impact on translation quality. Meanwhile, ensemble strategy can only result in little improvement due to the limited size of the training data. The final BLEU scores of en2de, en2fr, and en2es are 33.3, 45.9, and 37.1 respectively.

Strategy	En2de	En2fr	En2es
Baseline	28.9	35.6	30.5
+Tied-embed	29.5	-	-
+Multilingual	29.9	-	-
+R-Drop	30.6	43.0	34.3
+BT sampling	32.0	45.1	36.9
+FBTS	33.1	45.9	37.0
+Ensemble	**33.3**	**45.9**	**37.1**

Table 4: The experimental results of Model Augmentation.

5.2 Length Token and Length Encoding to Control Output Length

Our experiment demonstrates that the length token method is useful to control the output length. In order to enrich the diversity of results, we decode models using token $\{short; normal; long\}$ and LAB strategy, which correspond to LT-short, LT-normal and LT-long respectively. Table 3 shows that LT-normal model has the best overall quality. LT-short model leads to significantly shortened outputs and poor performance. LT-long model generates long outputs with relatively good performance. The above results further illustrate the shortening the length of outputs is the root cause of translation

Pairs	English-German				English-French				English-Spanish			
System	BLEU	F1	LR	LC	BLEU	F1	LR	LC	BLEU	F1	LR	LC
Enhanced	33.3	0.842	1.14	36.9	45.9	0.872	1.14	36.6	37.1	0.850	1.04	57.9
NAT	31.6	0.835	1.06	62.5	43.1	0.860	1.08	60.6	36.6	0.837	1.01	68.0
+LCD	30.4	0.829	1.04	83.5	42.3	0.848	1.05	82.3	36.1	0.830	1.01	89.9
+LAB	29.8	0.826	1.05	89.0	41.6	0.848	1.05	87.3	35.9	0.833	1.01	93.7
Unconstrained NAT	28.8	0.825	1.02	96.3	-	-	-	-	-	-	-	-

Table 5: The experimental result of Length-control decoding for NAT.

Pairs		English-German			
System	**Strategy**	BLEU	F1	LR	LC
Enhanced	LAB	33.0	0.838	1.10	68.6
LT-normal	LAB	31.2	0.830	1.04	80.8
LT-short	LAB	27.2	0.818	0.94	82.0
LT-long	LAB	32.6	0.839	1.15	45.4
NAT	LCD+LAB	29.8	0.826	1.05	89.0
Rerank1	-	30.7	0.830	1.03	99.8
Rerank3	-	30.7	0.830	1.04	100

Table 6: The experimental result of LAB and Rerank Method.

quality degradation. Although the LRPC method can dynamically adjust the length of the output, it negatively affects the translation quality, so we do not use the LRPC method in our submissions.

5.3 NAT to Control Output Length

Our experiments show that the model trained with NAT strategy can predict the output length based on the source length, so it outperforms the model trained with AT strategy on LC measurement, but underperforms the AT model on BLEU measurement. Table 5 illustrates that LCD strategy produces significantly improved LC scores but decreased BLEU scores. The LAB strategy leads to further improved LC scores but slightly decreased BLEU scores.

The unconstrained NAT model is trained along with the WMT14 English-German training data and fine-tuned with MuST-C. We witness significant improvements on LR and LC after increasing the data size. We believe data diversity is the reason for such improvement.

5.4 Effect of Length-aware beam and Rerank on Result

Table 2 shows that all systems achieve much higher LC scores when they are trained using LAB strategy. However, table 6 presents systems trained with various output length controlling methods without the rerank. Models without reranking can only achieve 89% LC at most. 100% LC can only be achieved by reranking all the above systems to minimize the deterioration of translation quality.

6 Conclusion

This paper presents HW-TSC's submission to IWSLT 2022 Isometric Spoken Language Translation Task. In general, we explore data and model augmentation methods, and achieve huge increases in BLEU scores when comparing with baseline models. In terms of length compliance, we use strategies such as Length Token, Length Encoding, NAT, Length-Aware Beam and Rerank. Our systems obtain 30.7, 41.6 and 36.7 BLEU respectively on the tst-COMMON test sets for English-German, English-French, English-Spanish tasks and 100% comply with the length requirements.

References

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel

decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Jiaxin Guo, Minghan Wang, Daimeng Wei, Hengchao Shang, Yuxia Wang, Zongyao Li, Zhengzhe Yu, Zhanglin Wu, Yimeng Chen, Chang Su, Min Zhang, Lizhi Lei, Shimin Tao, and Hao Yang. 2021. Self-distillation mixup training for non-autoregressive neural machine translation. *CoRR*, abs/2112.11640.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. *arXiv preprint arXiv:1910.10408*.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, 33:10018–10029.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965. PMLR.

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. *arXiv preprint arXiv:1904.01038*.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. *arXiv preprint arXiv:1904.07418*.

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 *Advances in neural information processing systems*, pages 5998–6008.

Minghan Wang, Jiaxin Guo, Yuxia Wang, Yimeng Chen, Chang Su, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2021a. HI-CMLM: improve CMLM with hybrid decoder input. In *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*, pages 167–171. Association for Computational Linguistics.

Minghan Wang, Jiaxin Guo, Yuxia Wang, Daimeng Wei, Hengchao Shang, Chang Su, Yimeng Chen, Yinglu Li, Min Zhang, Shimin Tao, and Hao Yang. 2021b. Diformer: Directional transformer for neural machine translation. *CoRR*, abs/2112.11632.

Minghan Wang, Guo Jiaxin, Yuxia Wang, Yimeng Chen, Su Chang, Hengchao Shang, Min Zhang, Shimin Tao, and Hao Yang. 2021c. How length prediction influence the performance of non-autoregressive translation? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 205–213, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. 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 4207–4216.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2019. Understanding knowledge distillation in non-autoregressive machine translation. *arXiv preprint arXiv:1911.02727*.

A Blind-test result

Table 7 presents the blind-test results for our submissions. isometric-slt-01, 02, 03, and 04 indicates Rerank1, Rerank2, Rerank3, and unconstrained

Pairs	English-German				English-French				English-Spanish			
System	BLEU	F1	LR	LC	BLEU	F1	LR	LC	BLEU	F1	LR	LC
isometric-slt-01	18.0	0.744	1.25	99.5	30.8	0.768	1.18	99.5	30.4	0.784	1.15	99.5
isometric-slt-02	17.8	0.753	1.18	100	27.8	0.763	1.17	100	28.7	0.788	1.15	100
isometric-slt-03	17.9	0.740	1.28	99.5	31.5	0.765	1.19	98.0	29.9	0.784	1.18	96.5
isometric-slt-04	20.2	0.759	1.03	96.0	-	-	-	-	-	-	-	-

Table 7: The experimental result of blind-test.

NAT results in our experiments. isometric-slt-03 post-processes punctuation over-translated, and as a result, it cannot 100% meets the length requirements.