A Holistic Approach to Reference-Free Evaluation of Machine Translation

Hanming Wu[1]*, **Wenjuan Han**[1]*, **Hui Di**[2], **Yufeng Chen**[1], **Jinan Xu**[1]†
[1] Beijing Jiaotong University, Beijing, China
[2] Toshiba (China) Co., Ltd., Beijing, China
21120416@bjtu.edu.cn, wjhan@bjtu.edu.cn
dihui@toshiba.com.cn, yfchen@bjtu.edu.cn, jaxu@bjtu.edu.cn

Abstract

Traditional machine translation evaluation relies on references written by humans. While reference-free evaluation gets rid of the constraints of labor-intensive annotations, it can pivot easily to new domains and is more scalable. In this paper, we propose a reference-free evaluation approach that characterizes evaluation as two aspects: (1) fluency: how well the candidate translation conforms to normal human language usage; (2) faithfulness: how well the candidate translation reflects the source data. We further split the faithfulness into word-level and sentence-level. Extensive experiments spanning WMT18/19/21 Metrics segment-level daRR and MQM datasets demonstrate that our proposed reference-free approach, `ReFreeEval`, outperforms SOTA reference-free metrics like YiSi-2, SentSim and BERTScore-MKD in most language directions. The code can be found at `ReFreeEval` Repo[1].

1 Introduction

Machine translation evaluation has conventionally relied on reference, where outputs are compared against translations written by humans. This is in contrast to the reference-free manner in which translation quality is directly assessed with the source text. Reference-free evaluation (Napoles et al., 2016; Thompson and Post, 2020; Agrawal et al., 2021) has the potential to free the evaluation model from the constraints of labor-intensive annotations, allowing it to pivot easily to new domains. In this way, reference-free evaluation metrics are substantially more scalable and have lately been in the spotlight.

The history of reference-free evaluation for MT can trace back to "QE as a Metric" track of WMT2019 Metrics Task (Ma et al., 2019). YiSi-2 (Lo, 2019) and XBERTScore (Zhang* et al., 2020; Leiter, 2021) are embedding-based methods that adopt contextual word embeddings to calculate the lexical similarity between the source and candidate translation words. Quality estimation (Fonseca et al., 2019) system metrics such as UNI+ (Yankovskaya et al., 2019) and COMET-QE (Rei et al., 2020a, 2021) also leverage contextual word embeddings and feed them into a feed-forward network. However, they are trained to regress on human scores that are expensive to collect, and gross discrepancies exist when different humans are asked to label the scores.

More challenging but worthwhile, we focus on dispensing with references as well as human scores. Nevertheless, embedding-based methods are limited to token-level semantic similarity while neglecting sentence-level faithfulness (Song et al., 2021). Besides, it's difficult for word embeddings to discriminate matched word pairs from random ones (Zhao et al., 2020a).

In addition, current reference-free evaluation methods rarely take fluency into account. For the unfluent candidates whose content is roughly consistent with the source, the embedding-based metrics can hardly discriminate and provide accurate evaluation scores[2]. Moreover, the general goal of evaluation metrics is to estimate not only the semantic equivalence between source and candidate but also the general quality (*i.e.*, fluency and naturalness) (Banchs et al., 2015; Feng et al., 2020; Yuan et al., 2021).

In this work, we propose a holistic approach (*i.e.*, `ReFreeEval`) to enhance the evaluation model in aspects of fluency and faithfulness, meanwhile on both word and sentence levels. With regard to fluency, we pose a data augmentation method and train a fluency discrimination module. For word-level faithfulness, we adopt a self-guided

* Equal contribution.

† Corresponding author.

[1] https://github.com/cocacola-lab/Reference-Free-Evaluation-of-Machine-Translation.git

[2] We provide more details and case studies in Appendix B.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 623–636
July 9-14, 2023 ©2023 Association for Computational Linguistics

contrastive word-alignment method. For sentence-level faithfulness, we execute knowledge distillation with SBERT (Reimers and Gurevych, 2019) to capture more fine-grained semantics. Our method builds on the framework of XBERTScore. Extensive experiments spanning WMT18/19/21 Metrics (Ma et al., 2018, 2019; Freitag et al., 2021) segment-level daRR and MQM datasets demonstrate that our proposed reference-free approach, `ReFreeEval`, outperforms SOTA reference-free metrics like YiSi-2, SentSim and BERTScore-MKD in most language directions.

2 Approach

Reference-free evaluation of MT can be characterized as two aspects: (1) fluency: how well it conforms to normal human language usage; and (2) faithfulness: how well the translated text reflects the source data. We assess faithfulness at different granularity: word level and sentence level. Figure 1 is the illustration of our `ReFreeEval` method.

2.1 Sentence-Level Fluency

We explore a data augmentation method to perturb the fluency of target sentences with noise which is difficult to be identified. Then we train a fluency discrimination module with contrastive learning (Gao et al., 2021; Zhang et al., 2021; Wu et al., 2022; Wang et al., 2022) to distinguish fluent samples from perturbed samples (namely, challenging negative samples).

Data Augmentation Using Clause Permutation A complex or compound sentence[3] has two or more clauses and relative clauses that are joined together with conjunctions or punctuation. As logical relations exist between these clauses, we manipulate and permute the clauses separated by punctuation, instead of words. In this way, the meaning is preserved inside the clauses, meanwhile, the sentence is often unfluent and unnatural. Similar to complex and compound sentences, for a simple sentence with only one clause[4], we randomly split it into two fragments and permute the two fragments. Compared to permutation on the token level, clause-level permutation has less influence on sentence fluency and semantic change. The clause-based permutation method brings perturbed samples that are more challenging and hard to be recognized.

Fluency Discrimination We denote a source and target sentence in parallel data as x and y. Perturbed samples augmented from y are $\hat{y}_1, \hat{y}_2, ..., \hat{y}_k$. A reliable metric has the ability to give the original fluent target y a higher evaluation score than those k perturbed unfluent samples.

As for the score, we adopt the same calculation measure as BERTScore but replace the pre-trained monolingual model (Devlin et al., 2019; Liu et al., 2019) with a cross-lingual model (Devlin et al., 2019; Conneau et al., 2019) to do reference-free evaluation (Zhou et al., 2020; Song et al., 2021) denominated as XBERTScore (Leiter, 2021). We use 9th layer of XLM-Roberta-Base to extract contextual word embeddings. Here we only use F_{BERT} as evaluation score between source x and target-side y or $\hat{y}_i$, which is represented as $s_w(x, y)$ or $s_w(x, \hat{y}_i)$. Then we can obtain word-level faithfulness scores $s_w(x, y), s_w(x, \hat{y}_1), ..., s_w(x, \hat{y}_k)$ of $(k + 1)$ pairs.

In order to discriminate fluent sentences from perturbed ones according to these scores, we treat the original target and its corresponding perturbed samples as opposite and assign them 1/0 hard labels. The cross-lingual model which produces XBERTScore is trained to classify target-side sentences with a cross-entropy loss function. The objective function on N training samples is as follows:

$$L_{fl} = -\frac{1}{N} \sum_{x,y} \log \frac{e^{s_w(x,y)}}{e^{s_w(x,y)} + \sum_{i=1}^{k} e^{s_w(x,\hat{y}_i)}} \quad (1)$$

2.2 Word-Level Faithfulness

As for word-level faithfulness, each word in the source sentence should have a corresponding cross-lingual representation in the target sentence and each word in the target sentence should be an accurate translation of its source word. This motivates us to do word-alignment training to enhance word-level evaluation.

This module shares similar architecture with sentence-level fluency where word embeddings are derived from 9th layer of XLM-Roberta-Base.

We take the same steps as (Dou and Neubig, 2021) to extract alignments. First, we compute the dot product between source and target word embeddings to obtain the similarity matrix S. Then S is normalized in source and target dimensions. And

[3] https://simple.wikipedia.org/wiki/Sentence#Types_of_sentence

[4] https://simple.wikipedia.org/wiki/Simple_sentence

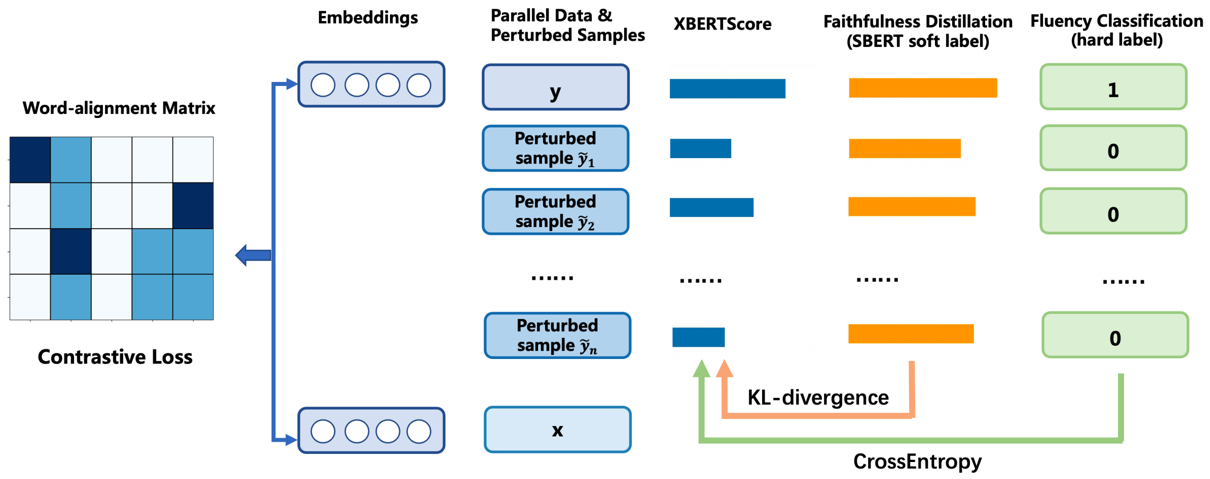
Embeddings
Parallel Data & Perturbed Samples
XBERTScore
Faithfulness Distillation (SBERT soft label)
Fluency Classification (hard label)
Word-alignment Matrix
y
Perturbed sample $\tilde{y}_1$
Perturbed sample $\tilde{y}_2$
Perturbed sample $\tilde{y}_n$
x
1
0
0
0
......
Contrastive Loss
KL-divergence
CrossEntropy

Figure 1: Illustration of `ReFreeEval` method with an example.

we get source-to-target alignment matrix S_{xy} and target-to-source alignment matrix S_{yx}. A source/-target token and a target/source token whose similarity value in alignment matrix S_{xy}/S_{yx} exceed threshold c_1 are regarded as aligned. The bidirectional alignment matrix A is deduced:

$$A = (S_{xy} > c_1) * (S_{yx}^T > c_1) \tag{2}$$

$A_{ij} = 1$ means x_i and y_j are aligned. Dou and Neubig (2021) also propose the self-training objective to align words with this bidirectional alignment, which improves alignment performance most.

Based on this objective, we adopt a self-guided contrastive cross-lingual word-alignment method. By contrast, we not only pull semantic aligned words to have closer contextual representations but also push unrelated words away (Luo et al., 2021; Su et al., 2022; Meng et al., 2022), which encourages the model to discriminate matched word embeddings from semantically unrelated ones.

The source token and target token are deemed to be unrelated if their similarity value is low. In our method, these unmatched pairs constitute negative samples and are pushed away. Moreover, we set threshold c_2 to further restrict the negative samples. The unmatched pairs whose similarity value is lower than c_2 are discarded from negatives as this unmatched relation can be easily distinguished by the model. In this way, we can control the difficulty of negative samples and only preserve those indistinguishable ones (hard negatives) to train the model.

$$B = (S_{xy} > c_2) * (S_{yx}^T > c_2) \tag{3}$$

$B_{ij} = 1$ means x_i and y_j are aligned or a part of hard negatives, which are preserved to train.

In Figure 1, the dark blue positions mean bidirectional alignment while the light blue positions are hard negative examples.

Finally, based on two dimensions of source and target, the positive and negative samples mentioned above, we construct a self-guided contrastive learning objective function on the word level as follows:

$$L_x = -\frac{1}{m}\sum_{i=1}^{m}\frac{\sum_{j=1}^{n}\mathbb{1}(A_{ij}=1)e^{S_{xy_{ij}}}}{\sum_{j=1}^{n}\mathbb{1}(B_{ij}=1)e^{S_{xy_{ij}}}} \tag{4}$$

$$L_y = -\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{m}\mathbb{1}(A_{ij}^T=1)e^{S_{yx_{ij}}^T}}{\sum_{j=1}^{m}\mathbb{1}(B_{ij}^T=1)e^{S_{yx_{ij}}^T}} \tag{5}$$

$$L_{word} = L_x + L_y \tag{6}$$

2.3 Sentence-Level Faithfulness

The main idea is to improve sentence-level faithfulness evaluation. Concretely, we distill sentence-level semantic meaning from SBERT into the word-level shared model.

We use SBERT to extract semantically meaningful sentence embeddings. Sentence semantic similarity between x and y is calculated with cosine-similarity between sentence embeddings $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$s_s(x, y) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \tag{7}$$

The semantic similarity reflects the sentence-level faithfulness from target to source. Then we can obtain sentence-level faithfulness scores $s_s(x, y)$, $s_s(x, \hat{y}_1), ..., s_s(x, \hat{y}_k)$. We use KL-divergence as the objective function to reduce the discrepancy between sentence-level and word-level similarity:

$$L_{fa} = \sum_{x,y' \in Y_x} s_s(x, y') \log \frac{s_s(x, y')}{s_w(x, y')} \quad (8)$$

In this distillation module, SBERT plays a role of a teacher. Sentence-level semantic knowledge is distilled into the word-level shared model through these sentence-level faithfulness scores. In this way, evaluation is no longer limited to word level but incorporated sentence semantics.

On the other hand, SBERT plays a role as a corrector. It is unreasonable that a disturbed sample with slightly changed semantics is considered to be completely contrary to the original sentence. We correct the binary classification and convert the 0/1 discrete value in the fluency discrimination module to continuous variables.

For sentence-level training, we combine fluency with faithfulness. This joint architecture is motivated by (Ren et al., 2021). The objective is:

$$L_{sent} = L_{fl} + \alpha L_{fa} \quad (9)$$

α is a hyper-parameter to control the weight that the sentence-level faithfulness module accounts for.

3 Experiment

3.1 Setup

Datasets We train and evaluate on four language pairs: English↔Chinese and English↔German. For training, we use the datasets following Awesome-Align (Dou and Neubig, 2021). The En-Zh training dataset is collected from the TsinghuaAligner[5] website and En-De training data is Europarl v7 corpus. For evaluation, we use the segment-level daRR dataset of WMT18/19 and MQM dataset of WMT21 Metrics Task. Details about datasets are introduced in Appendix C.1.

Embeddings We use the 9th layer of XLM-Roberta-Base to extract contextual word embeddings. This follows the default setting of BERTScore[6]. For sentence embeddings, we adopt *xlm-r-bert-base-nli-stsb-mean-tokens* model[7] the same as SentSim.

Baselines For reference-based metrics, we choose sentBLEU (Papineni et al., 2002) and YiSi-1 (Lo, 2019). For reference-free metrics, we choose XBERTScore (Leiter, 2021) , YiSi-2 (Lo, 2019), SentSim (Song et al., 2021) and BERTScore-MKD (Zhang et al., 2022). Most results of baseline models are reported in the original paper (Ma et al., 2018, 2019; Freitag et al., 2021; Zhang et al., 2022). We also implement experiments that have not been reported, such as XBERTScore, SentSim and BERTScore-MKD.

Training Process For `ReFreeEval`, sentence-level module is first trained. Then word-level faithfulness module is trained based on the best checkpoint of sentence-level training. Training details are in Appendix C.3.

Evaluation Measures For WMT18/19 segment-level evaluation, Kendall's Tau-like formulation is used to measure the scores against daRR.

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (10)$$

For WMT21 segment-level evaluation, conventional Kendall-tau statistic is used to measure the correlation between our scores and MQM scores.

3.2 Results

The main results are displayed in Table 1, 2, 3. First, we observe that fluency, word-level faithfulness, and sentence-level faithfulness module improve the evaluation performance respectively. We also find that the main improvement comes from sentence-level fluency indicating that XBERTScore as a token-level evaluation metric lacks sentence-level knowledge. Then, the ensemble model combining the advantages of the three modules achieves even better results. And compared with some reference-based baselines it achieves comparable results or even outperforms them. More details of experimental results are in Appendix C.4.

4 Conclusion

We propose a reference-free evaluation approach `ReFreeEval` that comprehensively considers three aspects: fluency, word-level faithfulness, and sentence-level faithfulness. Extensive experiments spanning datasets from WMT18/19/21 demonstrate the superiority of each module designed for each

[5] http://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html

[6] https://github.com/Tiiiger/bert_score

[7] https://huggingface.co/sentence-transformers/xlm-r-bert-base-nli-stsb-mean-tokens

[8] Reported by Zhang et al. (2022). Our implementations here are 0.1997 in Zh-En and 0.0623 in De-En.

Model	Zh-En	En-Zh	De-En	En-De
Reference-based				
SentBLEU	0.178	0.311	0.415	0.620
YiSi-1	0.211	0.323	0.488	0.691
Reference-free				
XBERTScore	0.0831	0.1129	0.3313	0.3143
SentSim	0.1213	0.1436	0.4127	0.4315
YiSi-2	0.091	0.101	0.279	0.359
BERTScore-MKD	0.1012	0.1102	0.4082	0.4329
word-level	0.0948	0.1355	0.3337	0.3413
fluency	0.1371	0.2503	0.3733	0.4751
sent-fa	0.1169	0.1759	0.3529	0.4319
sent-level	0.1798	0.2749	0.4144	0.5817
ReFreeEval	**0.1813**	**0.2920**	**0.4154**	**0.5884**

Table 1: Segment-level metric results for WMT18: absolute Kendall's Tau formulation on different evaluation metrics.

Model	Zh-En	En-Zh	De-En	En-De
Reference-based				
SentBLEU	0.323	0.270	0.056	0.248
YiSi-1	0.426	0.355	0.164	0.351
Reference-free				
XBERTScore	0.1482	0.0347	0.0488	0.1803
SentSim	0.2213	0.0771	0.0629	0.2334
YiSi-2	0.253	0.044	0.068	0.212
BERTScore-MKD	0.208[8]	0.0805	0.093[8]	0.2636
word-level	0.1864	0.0382	0.0517	0.1894
fluency	0.2435	0.1679	0.0682	0.2537
sent-fa	0.2346	0.0941	0.0497	0.2257
sent-level	0.3032	0.2387	**0.0807**	**0.3013**
ReFreeEval	**0.3173**	**0.2508**	0.0739	0.2995

Table 2: Segment-level metric results for WMT19: absolute Kendall's Tau formulation on different evaluation metrics.

Model	Zh-En w/o HT	En-De w/o HT	Zh-En w/ HT	En-De w/ HT
Reference-based				
SentBLEU	0.176	0.083	0.165	0.064
YiSi-1	0.302	0.172	0.289	0.145
Reference-free				
XBERTScore	0.2457	0.0367	0.2395	0.0176
SentSim	0.1938	0.0455	0.1867	0.0234
YiSi-2	**0.270**	0.098	**0.263**	0.071
BERTScore-MKD	0.2227	0.0503	0.2137	0.0290
word-level	0.2489	0.0388	0.2425	0.0196
fluency	0.2450	0.0482	0.2382	0.0281
sent-fa	0.2429	0.0448	0.2359	0.0238
sent-level	0.2601	0.0988	0.2520	0.0819
ReFreeEval	0.2628	**0.1008**	0.2543	**0.0828**

Table 3: Segment-level Kendall-Tau correlations for WMT21 MQM data.

aspect. ReFreeEval, combining the above three modules, achieves a higher correlation with human judgments, outperforming current SOTA reference-free metrics like YiSi-2, SentSim and BERTScore-MKD in most language directions.

Limitations

In this section, we discuss some limitations of our method and future work based on the limitations.

First, the enhancement of the word-level module is not as strong as the remedy of the sentence-level module. Our word-level module solely achieves improvement compared with XBERTScore but doesn't improve as much as the sentence-level module. The main reason is that the XBERTScore framework lacks sentence-level semantic knowledge. Besides, our word-level self-guided contrastive method doesn't resort to external information and only consolidates the alignment already existing in the pre-trained language model. Second, ReFreeEval performs comparably with baseline models on language pairs involving German. We guess it is due to the evaluation of QE. Ma et al. (2019) mention that the evaluation results across all language pairs are unstable in "QE as a Metric" track and can't explain yet.

In the future, we'll further explore valuable external information on word level. And we'll try to explore discrepancies among language pairs to optimize the results. In addition, our simple but effective data augmentation method - clause permutation doesn't rely on rules or toolkits, which is an initial attempt at modeling fluency. It could benefit from further refinement such as language-specific knowledge, syntactic and semantic parsing to recognize clauses. We'll conduct an in-depth investigation into further work.

Acknowledgements

The research work described in this paper has been supported by the National Key R&D Program of China (2020AAA0108001) and the National Nature Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130). Wenjuan Han is supported by the Talent Fund of Beijing Jiaotong University (2023XKRC006). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper. We would like to express our sincere gratitude to Hui Huang for guidance before this research. We are also grateful to Chunyou Li, Yu Xiang and Yu Zhang for their assistance during internship.

References

Sweta Agrawal, George Foster, Markus Freitag, and Colin Cherry. 2021. Assessing reference-free peer evaluation for machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1171, Online. Association for Computational Linguistics.

Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *CoRR*, abs/2101.08231.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics.

Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):59–66.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Christoph Wolfgang Leiter. 2021. Reference-free word- and sentence-level translation evaluation with token-matching metrics. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 157–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Ruikun Luo, Guanhuan Huang, and Xiaojun Quan. 2021. Bi-granularity contrastive learning for post-training in few-shot scene. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1733–1742, Online. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Zhao Meng, Yihan Dong, Mrinmaya Sachan, and Roger Wattenhofer. 2022. Self-supervised contrastive learning with adversarial perturbations for defending word substitution-based attacks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 87–101, Seattle, United States. Association for Computational Linguistics.

João Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André F. T. Martins. 2020. IST-unbabel participation in the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036, Online. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Catarina Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. *CoRR*, abs/2010.15535.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. SentSim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156, Online. Association for Computational Linguistics.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2022. TaCL: Improving BERT pre-training with token-aware contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2497–2507, Seattle, United States. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. SNCSE: contrastive learning for unsupervised sentence embedding with soft negative samples. *CoRR*, abs/2201.05979.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3898–3907, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021. Difficulty-aware machine translation evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 26–32, Online. Association for Computational Linguistics.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5786–5798, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Min Zhang, Hao Yang, Shimin Tao, Yanqing Zhao, Xiaosong Qiao, Yinlu Li, Chang Su, Minghan Wang, Jiaxin Guo, Yilun Liu, and Ying Qin. 2022. Incorporating multilingual knowledge distillation into machine translation evaluation. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy*, pages 148–160, Singapore. Springer Nature Singapore.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020a. Inducing language-agnostic multilingual representations. *CoRR*, abs/2008.09112.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020b. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Lei Zhou, Liang Ding, and Koichi Takeda. 2020. Zero-shot translation quality estimation with explicit cross-lingual patterns. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1068–1074, Online. Association for Computational Linguistics.

A Related Work

A.1 Reference-based Evaluation for MT

According to matching features, reference-based evaluation methods can be categorized as follows: (1) n-gram(e.g. BLEU (Papineni et al., 2002) and CHRF (Popović, 2015)); (2) edit distance(e.g. TER (Snover et al., 2006) and EED (Stanchev et al., 2019)); (3) word embedding(e.g. YiSi (Lo, 2019) and BERTScore (Zhang* et al., 2020)); (4)predictor-estimator model (Kim et al., 2017)(e.g. COMET (Rei et al., 2020a)). n-gram matching

	Sentence	DA	XBERTScore	ReFreeEval
SRC	但也有顾客认为，网站退款服务不是百分之百完美。			
REF	Nonetheless, some customers felt that website refund services are not perfect.			
MT1	But there are also customers who believe the site refund service is not 100 per cent perfect.	1.1059	0.8993	0.9249
MT2	But also some customers believe that website refunds money the service is not 100% perfect.	-1.5038	0.9031	0.8680

Table 4: An example of WMT18-Metrics dataset.

metrics are restricted to surface form and neglect semantic meaning.

Instead, embedding-based metrics adopt word embedding to explore word-level semantic meaning. WMDo (Chow et al., 2019) builds on Word Mover's Distance (Kusner et al., 2015) to measure the similarity of candidate and reference. It also introduces a word order penalty to take fluency into account. YiSi-1 aggregates the weighted lexical similarity to evaluate translation quality. BERTScore calculates the token-level semantic similarity between candidate translation tokens and reference tokens. DA-BERTScore (Zhan et al., 2021) takes translation difficulty into account and assigns difficulty weightings to each token in reference.

COMET leverages contextual word embeddings of the source sentence, MT hypothesis, and reference (or human post-edition) extracted from pretrained cross-lingual models. The embeddings are combined and fed into a feed-forward network. It's a quality estimation system and is trained with human assessments(DA, HTER, MQM).

A.2 Reference-Free Evaluation for MT

As reference is costly to be collected in practice, reference-free metrics attract more attention. Recent studies have explored evaluating translation quality only based on the source text.

YiSi-2 calculates similarities between cross-lingual word embeddings for aligned source and candidate translation words and outputs an F-measure statistic as the metric score. Zhao et al. (2020b) propose to re-align vector spaces and couple the semantic faithfulness scores with GPT-based fluency testing.

OpenKiWi-XLMR (Moura et al., 2020) and COMET-QE (Rei et al., 2020b) are quality estimation systems from "QE as a Metric" task (Mathur et al., 2020). They remove reference at the input but still require human assessments to train.

As reference-based BERTScore has achieved outstanding performance, many recent reference-free evaluation methods build on BERTScore. XBERTScore (Leiter, 2021) adopts the cross-lingual pre-trained language model to evaluate only based on source sentence without reference. SentSim (Song et al., 2021) combines semantic sentence similarity with token-level BERTScore. BERTScore-MKD (Zhang et al., 2022) also uses sentence embeddings to achieve cross-lingual word embedding alignment by multilingual knowledge distillation.

B Case Study

From Table 4, we can see there is a significant difference between the golden truth DA of MT1 and MT2. And the quality of MT1 is much better than MT2. But XBERTScore evaluates incorrectly and assigns MT1 with a lower score than MT2. Though MT2 is translated word by word which means poor fluency, almost all words in MT2 can be aligned with source. As XBERTScore method is evaluated on word-level matching, it can be easily confused. The model trained with our holistic approach can make up for this shortage and discriminate the fluency problem.

C Experimental Details

C.1 Data Analysis

Following the data setting of awesome-align (Dou and Neubig, 2021), we use the following parallel corpora to fine-tune our model. The English-Chinese(En-Zh) dataset is collected from the TsinghuaAligner webset and Englist-German(En-De) dataset is the Europarl v7 corpus. We only adopt a multilingual setting but use less data. We randomly sample 20k parallel sentence pairs from each dataset and mix them together.

In the word-level faithfulness module, we directly use mixed data to train. In the sentence-level fluency and faithfulness module, as only the target is perturbed, we randomly select 1/3 mixed data and swap the source and target in order to attend to all three languages.

To evaluate our method, we choose segment-level evaluation datasets of WMT Metrics Task. Two types of human assessments are included. Segment-level Metrics datasets of WMT18/19 use daRR(direct assessment relative ranking) as ground truth and WMT21 use MQM(multidimensional quality metrics) as ground truth.

C.2 Details of Sentence-Level Faithfulness

Before applying KL-divergence, the word-level and sentence-level similarity scores are processed as follows.

$$s_w(x,y) = \log(\frac{e^{s_w(x,y)}}{\sum_{y' \in Y_x} e^{s_w(x,y')}}) \qquad (11)$$

$$s_s(x,y) = \frac{e^{s_s(x,y)}}{\sum_{y' \in Y_x} e^{s_s(x,y')}} \qquad (12)$$

C.3 Training Details

Our model is fine-tuned based on the 9th layer of XLM-Roberta-Base. We implement our model with Pytorch (Paszke et al., 2019), Transformers (Wolf et al., 2020) and BERTScore (Zhang* et al., 2020) package. We use AdamW optimizer (Loshchilov and Hutter, 2017). The model is trained on up to 3 GeForce RTX 2080 Ti GPUs.

For sentence-level training, the hyperparameter settings are displayed in Table 5. We mainly search $\alpha \in \{0, 1, 5, 10, 20, 30, 40, 50, 100, 500\}$. The training process is on a single GPU with gradient accumulation. We evaluate the model with classification accuracy every 100 steps and save the checkpoint with the highest accuracy.

For word-level training, the hyperparameter settings are displayed in Table 6. We search batch size$\in \{8, 10, 15, 16, 24, 28, 32, 48\}$, learning rate$\in \{1e{-}5, 5e{-}6, 3e{-}6, 1e{-}6, 2e{-}6, 5e{-}7, 1e{-}7\}$ and $c_2 \in \{1e-5, 1e-10, 1e-15, 1e-20, 1e-30, 1e-50\}$. For dataset of WMT18/19 the training process is on 3 GPUs and the batch size on each GPU is 5. Specifically, for WMT21 MQM dataset the batch size is 32 and the learning rate is 2e-6. The training is on 4 GPUs and the batch size on each GPU is 8. The code of this module is implemented based on awesome-align (Dou and Neubig, 2021)[9]. This word-level faithfulness training continues on the basis of the best checkpoint of sentence-level training.

Hyperparameters	Values
Epoch	1
Evaluation Step	100
Batch Size	10
Learning Rate	1e-6
Warmup Steps	1000
α	30
k	7
Random Seed	42

Table 5: Hyperparameters for sentence-level training.

Hyperparameters	Values
Epoch	1
Batch Size	15(32)
Learning Rate	1e-6(2e-6)
Warmup Steps	200
c_1	1e-3
c_2	1e-20
Random Seed	42

Table 6: Hyperparameters for word-level training of WMT18/19.

C.4 Details of Experimental Results

In Section 3, as we want to demonstrate the improvement of our ReFreeEval in multilingual setting of all language directions, we report results corresponding to the highest "average" of all language pairs for each dataset.

Model	Zh-En	En-Zh	De-En	En-De
word-level	0.0948	0.1355	0.3337	0.3413
sent-level	0.1798	0.2749	0.4144	0.5817
`ReFreeEval`	**0.1857**	**0.2943**	**0.4154**	**0.5995**

Table 7: Segment-level best results of `ReFreeEval` on each language direction for WMT18

Table 7, 8, 9 are the best results of each language direction of WMT18/19/21 dataset.

[9] `https://github.com/neulab/awesome-align/tree/xlmr`

Model	Zh-En	En-Zh	De-En	En-De
word-level	0.1864	0.0382	0.0517	0.1894
sent-level	0.3032	0.2387	0.0807	0.3013
ReFreeEval	**0.3195**	**0.2561**	**0.0831**	**0.3041**

Table 8: Segment-level best results of ReFreeEval on each language direction for WMT19

Model	Zh-En w/o HT	En-De w/o HT	Zh-En w/ HT	En-De w/ HT
word-level	0.2489	0.0388	0.2425	0.0196
sent-level	0.2601	0.0988	0.2520	0.0819
ReFreeEval	**0.2684**	**0.1189**	**0.2603**	**0.1080**

Table 9: Segment-level best results of ReFreeEval on each language direction for WMT21.

D Analysis

D.1 Analysis of Data Augmentation

We compare our clause permutation with token-level data augmentation methods shuffling and repetition. The results are displayed in Table 10.

For the fluency module alone, our clause-based augmentation method performs much better than the others, which suggests that our method provides more proper and valuable fluency information than others. As for sentence-level faithfulness, we compare the variation of sentence semantic similarity in Table 11. The disturbance caused by token shuffling is too great while our clause permutation is small. The obvious disturbance is easy to be distinguished and learned. While the disturbance caused by our method can hardly be distinguished by sentence similarity thus only this module is not enough. However, with the clause permutation method, the combination of both fluency and sentence-level faithfulness outperforms others a lot. This verifies that our clause-based augmentation method is effective.

Based on the linguistic definition of clauses, our clause permutation approach can effectively incorporate perturbation to continuity and smoothness, which constitute the essence of fluency. This approach is simple and intuitive, making it a suitable choice for the preliminary step for more in-depth investigations about realistic perturbations.

D.2 Balance between Fluency Discrimination and Faithfulness Distillation

For sentence-level training, we adjust the hyper-parameter α to balance fluency and faithfulness.

A small α means the sentence-level training mainly focuses on classification, which may neglect the semantic meaning of perturbed samples as we explained in section2.3. While a large α weakens the effect of hard classification labels, the soft similarity is also not enough for sentence-level training.

From Table 12, we can conclude that only by keeping the balance between hard fluency discrimination and soft faithfulness distillation can we achieve excellent experimental results.

D.3 Control over Difficulty of Negative Samples in Word-Level Faithfulness

We experiment with different settings of threshold c_2 in word-level faithfulness to observe the influence of the difficulty of negative samples.

A small c_2 reduces the difficulty of contrastive learning. This setting includes negative samples whose unmatched relations can be easily distinguished. While a large c_2 restricts the negative samples extremely, which may lose some useful information. The results in Table 13 indicate that properly controlling the difficulty of negative samples can lead to great performance on the whole. However, for En-De, a small threshold is beneficial to improve the results. This may be because negatives without strict limitations are harmful to contrastive learning due to specific language features of En-De.

E Significance Test

Our experimental results above are based on the model training in a single run with random seed 42. In this section, we implement the statistical significance test following (Dror et al., 2018) to further compare the performance of our ReFreeEval with a strong baseline SentSim. We run both the models 10 times with random seed $\in \{19, 27, 42, 55, 76, 80, 99, 153, 178, 200\}$. The p-value of the statistical test is displayed in Table 14. As we can see, the p-value on each language pair is well below the significance level of 0.05, which indicates that the results of our ReFreeEval are significantly better than SentSim.

Data Aug Method	Model	Zh-En	En-Zh	De-En	En-De
Permutation	fluency	0.1371	0.2503	0.3733	0.4751
	sent-fa	0.1169	0.1759	0.3529	0.4319
	sent-level	0.1798	0.2749	0.4144	0.5817
Shuffle	fluency	0.1055	0.1815	0.3491	0.3749
	sent-fa	0.0809	0.0720	0.3100	0.3034
	sent-level	0.1469	0.2238	0.3729	0.4757
Repetition	fluency	0.1106	0.1586	0.3359	0.4048
	sent-fa	0.1305	0.1979	0.3642	0.4552
	sent-level	0.0654	0.0716	0.2846	0.2847

Table 10: Segment-level metric results for WMT18: absolute Kendall's Tau formulation with different data augmentation methods.

	Permutation	Shuffle	Repetition
Variation	-0.0213	-0.6285	-0.0380

Table 11: The variation of sentence similarity due to data augmentation compared with original data.

α	Zh-En	En-Zh	De-En	En-De
0	0.1425	0.2564	0.3774	0.4951
5	0.1636	0.2538	0.3971	0.5656
10	0.1664	0.2522	0.3979	0.5692
50	**0.1800**	**0.2798**	**0.4133**	**0.5843**
100	0.1608	0.2613	0.3918	0.5524
500	0.1277	0.2001	0.3615	0.4628

Table 12: The influence of different α setting on WMT18 segment-level metrics results.

c_2	Zh-En	En-Zh	De-En	En-De	Avg
1e-5	0.1707	0.2482	0.4039	**0.6019**	0.3562
1e-10	0.1768	0.2752	0.4133	0.5903	0.3639
1e-15	0.1778	0.2798	0.4134	0.5845	0.3639
1e-20	**0.1813**	**0.2920**	**0.4154**	0.5884	**0.3693**
1e-30	0.1806	0.2766	0.4072	0.5757	0.3600
1e-50	0.1744	0.2701	0.3898	0.5663	0.3502
0	0.0883	0.2077	0.2781	0.4998	0.2685

Table 13: The influence of different threshold c_2 setting on WMT18 segment-level metrics results.

	Zh-En	En-Zh	De-En	En-De
WMT18	7.02e-13	1.78e-14	3.43e-2	2.02e-16
WMT19	3.79e-17	5.72e-15	2.09e-2	1.93e-12
WMT21 w/o HT	5.36e-13	-	-	6.26e-16
WMT21 w/ HT	8.44e-13	-	-	1.36e-15

Table 14: p-value of significance test on WMT18/19/21 Metric datasets.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section: Limitations

☒ A2. Did you discuss any potential risks of your work?
No potential risks

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section: Abstract and Section1: Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Mainly in Section3: Experiments and Appendix D: Experimental Details. Also Section1: Introduction and Section2: Approach mention models.

☑ B1. Did you cite the creators of artifacts you used?
Section Reference and footnotes of Section2 and Section 3.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section Reference and footnotes of Section2 and Section 3.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section2 and Section3

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Our data doesn't have these information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section3 and Appendix D

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section3 and Appendix D.1

C ☑ Did you run computational experiments?

Section3 and AppendixE / F

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix D.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section3 and Appendix D.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix F

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix D.3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.