MEE4 and XLsim : IIIT HYD's Submissions' for WMT23 Metrics Shared Task

Ananya Mukherjee and **Manish Shrivastava**
Machine Translation - Natural Language Processing Lab
Language Technologies Research Centre
International Institute of Information Technology - Hyderabad
ananya.mukherjee@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

This paper presents our contributions to the WMT2023 shared metrics task, consisting of two distinct evaluation approaches: a) **Unsupervised Metric** (MEE4) and b) **Supervised Metric** (XLSim). MEE4 represents an unsupervised, reference-based assessment metric that quantifies linguistic features, encompassing lexical, syntactic, semantic, morphological, and contextual similarities, leveraging embeddings. In contrast, XLsim is a supervised reference-based evaluation metric, employing a Siamese Architecture, which regresses on Direct Assessments (DA) from previous WMT News Translation shared tasks from 2017-2022. XLsim is trained using XLM-RoBERTa (base) on English-German reference and mt pairs with human scores. Here are the links for MEE4 [1] and XLsim[2] metrics.

1 Introduction

In recent times, there has been a growing interest in Neural Machine Translation (NMT) systems, leading to significant improvements in machine translation (MT) quality. Over the past few years, the field of MT evaluation has seen substantial advancements. Each year, the WMT conference hosts a metrics-shared task, where new evaluation metrics are introduced and those demonstrating a strong correlation with human judgments are highlighted from the array of newly devised metrics. In the last three years of the WMT Metrics Task (Freitag et al., 2022, 2021; Mathur et al., 2020), neural-based metrics have predominantly taken the lead. Nevertheless, n-gram-based and lexical-based metrics (Papineni et al., 2002; Popović, 2015) continue to be favored as automatic MT evaluation tools due to their flexibility and efficiency.

As a result, this year we participated in the metrics shared task, evaluating machine translation outputs using two types of metrics: an unsupervised metric and a supervised metric.

Unsupervised Metric: Our unsupervised metric, MEE4 (Mukherjee and Shrivastava, 2022), relies on a combination of lexical and embedding similarity measures. Notably, MEE4 demonstrated strong performance in the previous year's shared task (Freitag et al., 2022), surpassing several baseline metrics such as BERTscore (Zhang* et al., 2020), BLEU (Papineni et al., 2002), F101SPBLEU (Goyal et al., 2022) , and CHRf (Popović, 2015). In our efforts to improve its performance further this year, we conducted experiments with two different sentence embedding models: LaBSE (Feng et al., 2022) and the stsb-xlm-r-multilingual [3]. Interestingly, our findings indicated that MEE4, when equipped with LaBSE as the sentence embedding model, exhibited superior performance compared to the alternatives.

Supervised Metric: Unlike the existing neural models which are huge in size, our goal was to build a more compact supervised training model (XLsim) that offers improved performance. To achieve this, we created a SentenceTransformer model by combining a pre-trained transformer model with a pooling layer. This hybrid approach enables the generation of sentence embeddings, which can be compared using cosine similarity to assess similarity between sentences.

2 MEE4

MEE4 is an improved version of MEE focusing on computing contextual and syntactic equivalences, along with lexical, morphological, and semantic similarity. The goal is to comprehensively evaluate the fluency and adequacy of MT outputs while also considering the surrounding context. Fluency is determined by analyzing syntactic correlations, while context is evaluated by comparing sentence

[1] https://github.com/AnanyaCoder/WMT22Submission

[2] https://github.com/AnanyaCoder/XLsim

[3] https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 800–805

similarities using sentence embeddings. The ultimate score is derived from a weighted amalgamation of three distinct similarity measures: a) Syntactic similarity, which is established using a modified BLEU score. b) Lexical, morphological, and semantic similarity, quantified through explicit unigram matching. c) Contextual similarity, gauged by sentence similarity scores obtained from the Language-Agnostic BERT model (Feng et al., 2022).

In our experiments this year, we made adjustments to MEE4 while maintaining the same underlying architecture. Specifically, we computed the evaluation scores using a different sentence embedding model.

In addition to our previous choice, we utilized the stsb-xlm-r-multilingual model. This particular sentence-transformers model is designed to map sentences and paragraphs into a 768-dimensional dense vector space, making it suitable for various tasks such as clustering and semantic search. It's worth highlighting that the version of XLM-R (Conneau et al., 2020) we employed is considered a state-of-the-art model for multilingual Semantic Textual Similarity (STS) (Reimers and Gurevych, 2020).

2.1 Multilingual Sentence Encoders

Numerous multilingual sentence encoders, including mBERT (Devlin et al., 2018), consist of single self-attention networks. These models are pretrained on monolingual corpora in over 100 languages and are optimized for masked language modeling. Here, the model is tasked with predicting randomly selected tokens in the original text that have been replaced by a placeholder.

However, these pretrained multilingual sentence encoders often exhibit limited sensitivity to cross-language semantic similarity. To address this issue, Reimers and Gurevych employed human Semantic Textual Similarity (STS) annotations to enhance a pretrained multilingual sentence encoder, specifically BERT resulting in *stsb-xlm-r-multilingual* model .

In contrast, LaBSE differs slightly as it has been trained not only for masked language modeling but also for translation language modeling.

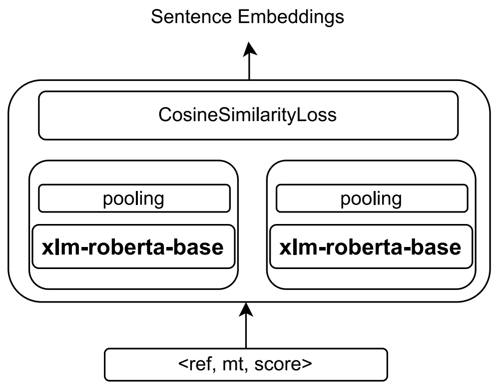

Figure 1: Illustration of our methodology using siamese network architecture
En: English IL- Indian Langauge

3 XLSim: MT Evaluation Metric based on Siamese Architecture

XLsim is a supervised reference-based metric that regresses on human scores provided by WMT (2017-2022). Using a cross-lingual language model XLM-RoBERTa-base[4] (Conneau et al., 2020), we train a supervised model using a Siamese network architecture with CosineSimilarityLoss.

3.1 Training Data

The WMT DA human evaluation data[5] (WMT17-WMT22) (Kocmi et al., 2022; Akhbardeh et al., 2021; Barrault et al., 2020, 2019; Bojar et al., 2018, 2017) contains raw score and z-score; we considered z-score for our training purpose by normalizing it to a range of 0-1.

3.2 Siamese Network Architecture

Similar to SBERT, we train the network with a Siamese Network Architecture (Reimers and Gurevych, 2019). In this siamese network, for each sentence pair, we pass *reference translation (ref)* and *hypothesis translation (mt)* through our network which yields the embeddings u und v. The similarity of these embeddings is computed using cosine similarity and the result is compared to the gold similarity score *(score)*. This allows our network to be fine-tuned and recognize sentence similarity. Figure 1 illustrates our XLsim training architecture.

While training, we used **CosineSimilarityLoss**, which automatically ensures training in a siamese network structure.

[4]https://huggingface.co/xlm-roberta-base
[5]https://huggingface.co/datasets/RicardoRei/wmt-da-human-evaluation

ref	I believe that financially, automakers are doing very well now, maintaining high sales margins.
mt	I believe car manufacturers are feeling very good financially right now, maintaining high sales margins.
score	0.77

Table 1: Input Example

3.3 CosineSimilarityLoss

CosineSimilarityLoss expects that the input consist of two texts and a float label. Refer Table 1.

It computes the vectors $u = model(input[0])$ and $v = model(input[1])$ and measures the cosine-similarity between the two. By default, it minimizes mean squared error loss.

3.4 Training Details

In our experiment, we focused on the en-de[6] language pair and utilized specific columns from the wmt-da-human-evaluation dataset, which included translation (mt), reference translation (ref), and z-score (score). Among the total 125,992 en-de samples available, we partitioned them as follows: 105,992 samples were used for training, 10,000 for validation, and another 10,000 for testing.

We employed a SentenceTransformer architecture to train our model, leveraging a multilingual pre-trained transformer model, XLM-RoBERTA base model. XLM-RoBERTa (Conneau et al., 2020) model is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages.

We utilized the CosineSimilarityLoss function for a total of 4 training epochs. Our training setup involved a batch size 16, employing the Adam optimizer with a learning rate 2e-5 and a linear learning rate warm-up strategy over 10% of the training data. The entire training process was carried out on NVIDIA GPUs, specifically T4 x2.

3.5 Inference

To assess translation quality based on reference, our trained model generates embeddings for reference and translation sentences and subsequently calculates the cosine similarity between these embeddings. This similarity measure serves as a metric for evaluating the quality and similarity between the translation and reference text (refer figure 2).

[6]we chose the language-pair having a more significant number of samples than other language-pairs.

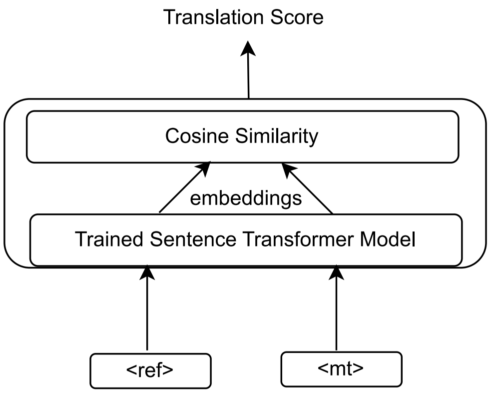

Figure 2: XLsim architecture at inference (to compute segment-level scores)

Model	COMET	XLsim
Size	2.32 GB	1.1 GB
Training Samples#	1,027,155	105,992
Pearson correlation	0.68	0.52

Table 2: Comparison with the SOTA neural metric based on Pearson Correlation with human scores.

Table 2 reports the comparison of our trained metric with the existing state-of-the-art metric, COMET (Rei et al., 2022) in terms of model size, total training samples and pearson correlation on the 10000 en-de samples (test samples see 3.4). It is worth noticing that the difference in correlation is 0.16 which is minute and model is 50% lesser in size.

4 WMT23 Metric Shared Task Submission

4.1 Segment Level Evaluation

For Segment-level task, we submitted the sentence-level scores obtained by our reference-based unsupervised metrics namely MEE4 (primary metric) and MEE4_stsb_xlm.

For the same Segment-level task, we also submitted the sentence-level scores obtained by our reference-based supervised evaluation metric (XLsim).

4.2 System Level Evaluation

To calculate the system-level score for each system, we take the average of the segment-level scores that we've derived. We employ a similar approach when computing system-level scores based on segment-level human annotations, such as DA's and MQM.

testset	lp	#sentences	XLsim	MEE4	MEE4_stsb_xlm
	en-de	6684	**0.67**	0.64	0.47
generaltest2023	zh-en	29640	0.68	**0.74**	0.59
	he-en	22920	0.76	**0.78**	0.62
	en-de	33470	**0.73**	0.71	0.64
challengeset	zh-en	6996	0.86	**0.91**	0.89
	he-en	9466	0.80	**0.86**	0.85

Table 3: Pearson correlation of evaluated scores on WMT23 submissions with COMET metric.

This suggests that a metric with a strong correlation at the segment level should also exhibit a robust correlation at the system level.

4.3 Results

Table 3 provides the details of the WMT23 Metric Shared Task test-set for the language pairs we investigated. However, it's important to note that the final and most comprehensive analysis will rely on the official results, where metric submissions are thoroughly compared to human judgments.

In our preliminary assessment, we have reported Pearson correlation scores for the submitted metrics when compared to COMET at the segment-level. This analysis helps us gauge the performance of the three metrics in relation to the state-of-the-art metric. In case of Unsupervised metrics, it appears that MEE4, which utilizes LaBSE, outperforms MEE4_stsb_xlm, which employs stsb-xlm-r-multilingual as its sentence embedding model. This difference in performance may be attributed to the training techniques applied to LaBSE, which involve both masked language modeling and translation language modeling, making it more effective for the task. Indeed, it's evident that XLsim exhibits a relatively strong correlation with COMET, almost exceeding 0.7. However, when compared to MEE4, there is a mild decrease in performance, particularly in the zh-en (Chinese to English) and he-en (Hebrew to English) language pairs, where the correlation drops by approximately 0.06.

This slight decline in performance for XLsim in certain language pairs could be attributed to the fact that even though XLsim utilizes the pre-trained multilingual XLM-Roberta model, the training data (ref, mt) was primarily in the German (de) language.

5 Conclusion and Future Work

In this paper, we describe our submissions to the WMT23 Metrics Shared Task. Our submission includes segment-level and system-level translation evaluation scores for sentences of three language pairs English-German (en-de), Chinese-English (zh-en) and Hebrew to English (he-en). We evaluate this year's test set using: a)two unsupervised metrics, *MEE4 and MEE4_stsb_xlm*. These metrics are based on lexical and embedding similarity match that evaluates the translation on various linguistic features (syntax,lexical, morphology, semantics and context) ; b) a supervised metric, XLsim that learns on en-de WMT DA human evaluation data from 2017-2022. It is observed that all the three metrics displayed a positive correlation (>0.5) with the baseline metric COMET.

Certainly, there are promising research directions to explore, especially in the realm of metric enhancement. In our future work, we intend to delve deeper into these areas:

MEE4 Metric Improvement: One of our primary objectives is to refine and enhance MEE4, seeking more efficient approaches that can better estimate translation quality while achieving higher agreement with human judgments. This might involve exploring novel techniques in sentence embedding, fine-tuning, or leveraging additional linguistic information.

XLsim Enhancement: For XLsim, we plan to boost its performance by optimizing the training data. This involves ensuring that it is trained on a more diverse set of languages and data to improve its cross-lingual capabilities. Simultaneously, we aim to maintain its compactness and ensure it remains trainable with fewer computational requirements.

These future research directions hold the potential to contribute significantly to the field of machine translation evaluation, ultimately leading to more robust and accurate metrics that align closely with human assessments.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. 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. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Ananya Mukherjee and Manish Shrivastava. 2022. Unsupervised embedding-based metric for MT evaluation with improved human correlation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 558–563, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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*.