Unsupervised Embedding-based Metric for MT Evaluation with Improved Human Correlation

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

In this paper, we describe our submission to the WMT22 metrics shared task. Our metric focuses on computing contextual and syntactic equivalences along with lexical, morphological and semantic similarity. The intent is to capture fluency and context of the MT outputs along with their adequacy. Fluency is captured using syntactic similarity and context is captured using sentence similarity leveraging sentence embeddings. The final sentence translation score is the weighted combination of three similarity scores: a) Syntactic Similarity b) Lexical, Morphological and Semantic Similarity and c) Contextual Similarity. This paper outlines two improved versions of MEE i.e., MEE2 and MEE4. Additionally, we perform our experiments on language pairs of en-de, en-ru and zh-en from WMT17-19 testset and further report the correlation with human assessments. Our submission will be made available at https://github.com/AnanyaCoder/WMT22Submission.

1 Introduction

Neural Machine Translation (NMT) systems have emerged with an increased research interest in recent times and significantly enhanced the MT quality. However, the MT research community still relies mainly on antiquated metrics and no new, universally adopted standard metric has emerged. In the last few years, research in Machine Translation (MT) evaluation has made significant progress. A metrics-shared task is held annually at the WMT conference, where new evaluation metrics are proposed and those which correlates highly with human judgements are presented from the pool of newly defined metrics. Neural-based metrics largely dominated the last two years of the WMT Metrics Task (Freitag et al., 2021; Mathur et al., 2020; Ma et al., 2019). Nevertheless, n-gram based or lexical-based metrics remain popular as automatic MT evaluation metric due to their agile and light-weighted nature. Traditionally, automatic metrics for evaluating MT quality have relied on estimating the similarity between machine outputs and reference sentences in the target language. However, advanced NMT methods yield high-quality translations that might have lexical, morphological, syntactic variations and different word choices having similar meanings. Typically, the machine output diverges from monotonic lexical transfer between the source and target languages. Widely used evaluation metrics rely on basic, lexical-level features as they calculate the surface similarity between the hypothesis and reference sentences by counting the number of matching n-grams (Papineni et al., 2002; Doddington, 2002). Metrics relying on n-gram overlap cannot appropriately capture morphological, syntactic and semantic variations as they are sensitive to only lexical variations. METEOR (Denkowski and Lavie, 2014; Gupta et al., 2010; Lavie and Denkowski, 2009; Lavie and Agarwal, 2007; Banerjee and Lavie, 2005) captures semantic variations but it is highly dependent on language specific tools . Hence, there is huge requirement for a robust, understandable, easy to use automatic MT evaluation metric which captures all the linguistic features to evaluate like humans. The better evaluation metric will be highly helpful to the development of better MT systems (Liu et al., 2011).

In this paper, we present our submission to the WMT2022 metrics shared task. We evaluate the translations of English-German (en-de), English-Russian (en-ru) and Chinese-English (zh-en) language pairs. However, the proposed metric is language independent and supports 100+ languages. Here, our submission includes scores of three metrics MEE (Mukherjee et al., 2020), MEE2 and MEE4 (MEE2 and MEE4 are extended versions of MEE). We have evaluated the testsets of WMT17 (Bojar et al., 2017), WMT18 (Bojar et al., 2018) and WMT19 (Bojar et al., 2019a,b,c), for the same

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 558 - 563
December 7-8, 2022 ©2022 Association for Computational Linguistics

language pairs (en-de, en-ru and zh-en) and reported the correlation with human assessments. The empirical results conclude that MEE4 shows better agreement with humans.

2 *M*etric for *E*valuation using *E*mbeddings (MEE)

2.1 MEE

MEE (Mukherjee et al., 2020) is an automatic evaluation metric that leverages the similarity between embeddings of words in candidate and reference sentences to assess translation quality focusing mainly on adequacy. Unigrams are matched based on their surface forms, root forms and meanings which aids to capture lexical, morphological and semantic equivalence. Semantic evaluation is achieved by using pretrained fasttext embeddings (Grave et al., 2018) provided by Facebook to calculate the word similarity score between the candidate and the reference words. MEE computes evaluation score using three modules namely exact match, root match and synonym match. In each module, fmean-score is calculated using harmonic mean of precision and recall by assigning more weightage to recall. Final translation score is obtained by taking average of fmean-scores from individual modules.

2.2 MEE2, MEE4

MEE2 and MEE4 are improved versions of MEE that capture lexical, morphological, semantic, contextual and syntactic similarity. These linguistic aspects are captured in different modules and the final sentence translation score is the weighted pool of these individual modules. Unlike MEE, these metrics capture fluency and sentence semantics. **Contextual similarity** (or sentence semantics) is obtained by computing a cosine similarity between sentence embeddings of reference sentence and system output. Whereas fluency is captured by performing **Syntactic Similarity** which is computed by using a modified BLEU score. **Lexical, Morphological and Semantic**[1] **Similarity** is measured by explicit unigram matching similar to MEE.

Figure 1 illustrates the segment-level computation of final translation score of based on a reference sentence.

2.2.1 Syntactic Similarity

Our approach assesses fluency by capturing the syntactic similarity between the reference and the hypothesis using BLEU (Papineni et al., 2002) since it follows the notion that longer n-gram scores account for the fluency of the translation. However, the length with the "highest correlation with monolingual human judgements" was found to be four (BLEU-4). Our experiments adopt the concept of BLEU with a slight variation i.e., dynamic n-gram (n depends on the sentence length). Here, while evaluating a hypothesis, the order of n-gram is based on the corresponding reference sentence length.

2.2.2 Lexical, Morphological and Semantic Similarity

In our work, *lexical, morphological and semantic* equivalence score is computed in similar to MEE metric [2]. MEE (Metric for Evaluation using Embeddings) contains three modules, namely *Exact Match, Root Match, and Synonym Match* which accounts for *lexical, morphological and semantic* features of the translation (Mukherjee et al., 2020).

2.2.3 Contextual Similarity

Contextual Similarity Score is computed by measuring the distance between the hypothesis sentence embedding and reference sentence embedding. Sentence Embedding models map text/sentences to a vector space, implying that related or similar sentences lie closer to each other in this embedding space. Sentence embedding captures the intention of the sentence. Our work is based on the assumption that contextual information of a given sentence can be captured from its vector (or embedding). We determine the context equivalence of two sentences by computing a cosine similarity (Foreman, 2014) between the embeddings of reference and hypothesis. Contextual equivalence is calculated by computing cosine similarity between the sentences embedded using LaBSE by Google AI. Out of several existing Language-Agnostic models, LaBSE (Feng et al., 2020), LASER (Artetxe and Schwenk, 2018), and Indic-Bert (Kakwani et al., 2020) we prefered to use LaBSE to embed the sentences as it is a multilingual BERT embedding model trained using MLM and TLM pre-training, resulting in a model that is effective even on low-resource languages

[1] word-level semantic similarity

[2] https://github.com/AnanyaCoder/MEE_WMT2021

for which there is no data available during training. Also, it produces language-agnostic cross-lingual sentence embeddings for 109 languages.

2.3 Score Computation

The segment-level evaluation score is computed as follows. Based on number of matched unigrams in candidate and reference sentence, individual fmean scores are computed at lexical, morphological and semantic levels. These fmean scores are achieved by parameterized harmonic mean (Sasaki, 2007) of precision and recall as per Equation 3. Ulitmately, MEE score is computed by averaging the individual fmean scores of three modules.

$$precision(P) = \frac{\#matched\_unigrams}{Total\#unigrams\_in\_hypothesis} \quad (1)$$

$$recall(R) = \frac{\#matched\_unigrams}{Total\#unigrams\_in\_reference} \quad (2)$$

$$f_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad (3)$$

MEE2 and MEE4 is computed using Equation 4 where LMS score is same as the MEE score (Mukherjee et al., 2020) i.e., $\beta = 3$ in Equation 3. Syn and Cxt are Syntax Simlarity score and Contextual Similarity score of reference and translation. The parameters in Equation 4 are manually tuned for computing MEE2 and MEE4 scores[3]. For MEE2: $\alpha = 2, \gamma = 1, \delta = 1, \epsilon = 1$ and for MEE4: $\alpha = 2, \gamma = 1, \delta = 1, \epsilon = 3$

$$score = \frac{\delta * \frac{\alpha * LMS + \gamma * Syn}{\alpha + \gamma} + \epsilon * Cxt}{\delta + \epsilon} \quad (4)$$

3 Experiments and Results

3.1 Results on WMT17-19 testset

Each year, the WMT Translation shared task organisers collect human judgements in the form of Direct Assessments. Those assessments are then used in the Metrics task to measure the correlation between metrics and therefore decide which metric works best. Therefore, we evaluated a total of 9K sentences from the testset of WMT17, WMT18, WMT19 for en-ru, en-de, zh-en language pairs and computed the pearson correlation (Benesty et al., 2009) of MEE, MEE2, MEE4 with human assessments. The segment level correlation scores are mentioned in Table 1. It is clearly evident that MEE4 correlates better with humans i.e., across the different testsets and language pairs, MEE4 demonstrates higher agreement with human judgements.

3.2 WMT22 task submission

During our experiments, we tested several techniques: averaging the module scores with different weights. Based on the agreement with humans on the WMT17-19 testset (refer Table 1, we decided to report the scores of MEE, MEE2 and MEE4 for the current WMT22 metric shared task submission. Table 2 shows the WMT22 test-set details we have experimented on.

3.2.1 Segment Level Evaluation

For Segment-level task, we submitted the sentence level scores obtained by our reference based metrics MEE2 and MEE4 for en-ru, en-de and zh-en language pairs.

3.2.2 System Level Evaluation

For the System-level task we compute the system-level score for each system by averaging the segment-level scores obtained. We observe an equivalent approach used to compute system-level scores based on segment-level human annotations such as DA's and MQM, implying that a metric that achieves a solid segment-level correlation should also gain strong system-level performances.

4 Conclusion and Future Work

In this paper, we present our participation to the WMT22 Metrics Shared Task. Our submission includes segment-level and system-level scores for sentences of three language pairs Chinese-English (zh-en), English-Russian (en-ru) and English-German (en-de). We evaluate this year's test set using our **unsupervised, reference-based** metrics: MEE2 and MEE4. Both the metrics are extended versions of MEE with improved correlation. From the last year's findings, it was evident that MEE2 was one among the better performing metrics as it was highlighted in the top significant cluster (Freitag et al., 2021). However, this year we present MEE4 along with MEE2 and MEE4 has proved to perform better in terms of correlation with humans when evaluated on testsets of WMT17, WMT18 and WMT19. We observe that this improvement in agreement to human experts level judgements is due to assigning more weightage to context information (sentence level semantics) when compared

[3] These scores range from 0-1.

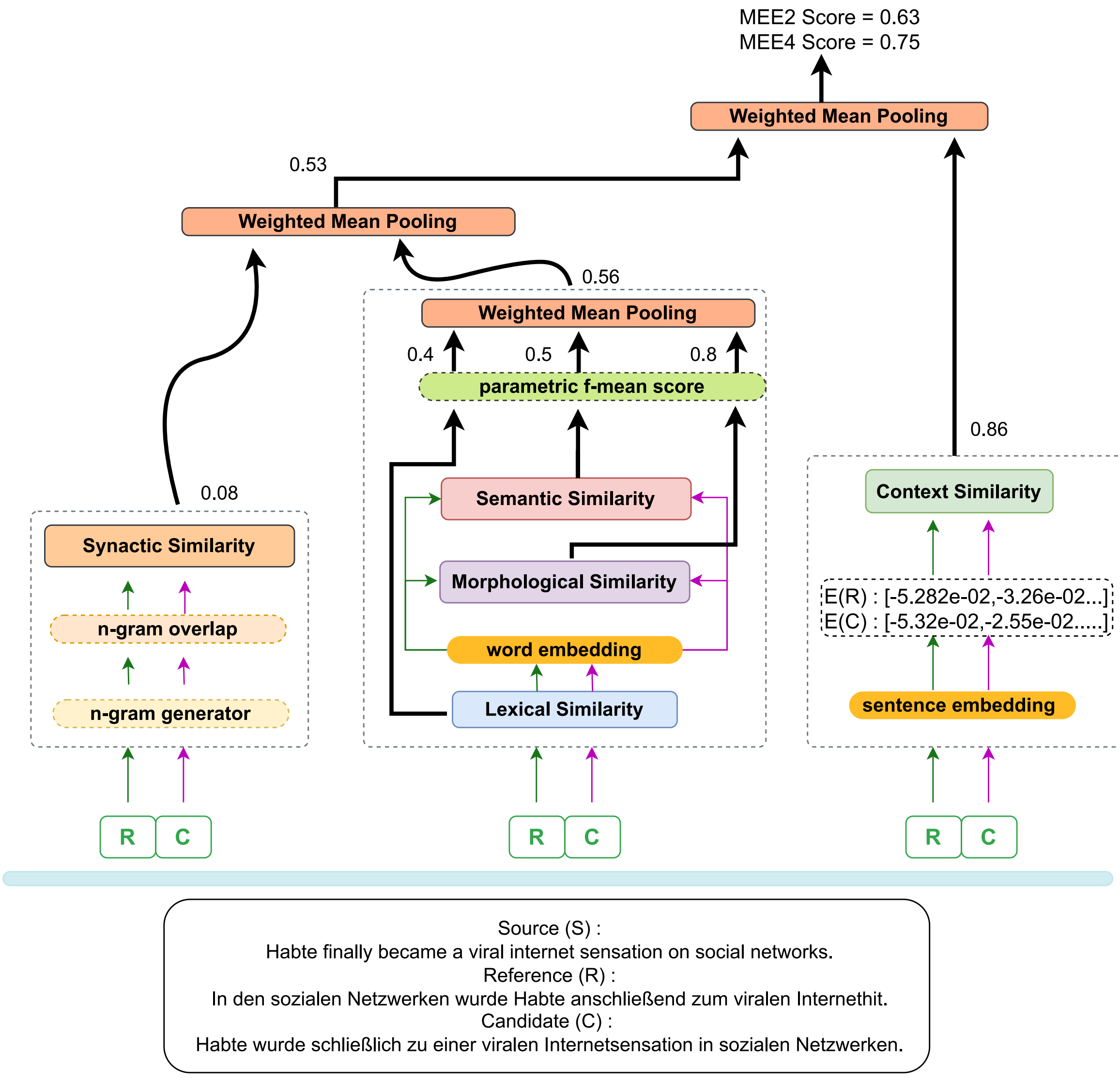

Figure 1: Illustration of our model Architecture.

Test-set	LP	#Sentences	BLEU	MEE	MEE2	MEE4
WMT17	zh-en	1000	0.22	0.261	0.383	0.402
	en-ru	1000	0.32	0.376	0.476	0.495
	en-de	1000	0.2	0.211	0.326	0.380
WMT18	zh-en	1000	0.18	0.189	0.273	0.290
	en-ru	1000	0.32	0.335	0.404	0.414
	en-de	1000	0.42	0.476	0.549	0.563
WMT19	zh-en	1000	0.33	0.328	0.5	0.555
	en-ru	1000	0.35	0.465	0.491	0.489
	en-de	1000	0.24	0.245	0.322	0.351

Table 1: Segment Level Correlation with Human Judgements on WMT17, WMT18 and WMT19 testset.

to other linguistic aspects. In future, we plan to further experiment on optimizing weights assigned to individual linguistic modules with an aim to evaluate the translations to have better correlation with humans.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

Language Pair	#Sentences	#Systems
en-ru	33988	82
en-de	97002	164
zh-en	73668	192

Table 2: Data statistics of WMT22 testset for en-ru, ende and zh-en pairs.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. *Pearson Correlation Coefficient*, pages 1–4. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors. 2017. *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*. Association for Computational Linguistics, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019a. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019b. *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019c. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Association for Computational Linguistics, Florence, Italy.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors. 2018. *Proceedings of the Third Conference on Machine Translation*. Association for Computational Linguistics, Belgium, Brussels.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *the second international conference*, page 138.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

John Foreman. 2014. COSINE DISTANCE, COSINE SIMILARITY, ANGULAR COSINE DISTANCE, ANGULAR COSINE SIMILARITY.

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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Ankush Gupta, Sriram Venkatapathy, and R. Sangal. 2010. Meteor-hindi : Automatic mt evaluation metric for hindi as a target language.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2–3):105–115.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2011. Better evaluation metrics lead to better machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 375–384, Edinburgh, Scotland, UK. 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.

Ananya Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. Mee : An automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299.

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*, pages 311–318.

Yutaka Sasaki. 2007. The truth of the f-measure. *Teach Tutor Mater*.