chrF-S: Semantics is All You Need

Ananya Mukherjee, Manish Shrivastava
MT-NLP Lab, LTRC, KCIS, IIIT Hyderabad, India
ananya.mukherjee@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Machine translation (MT) evaluation metrics like BLEU and chrF++ are widely used reference-based metrics that do not require training and are language-independent. However, these metrics primarily focus on n-gram matching and often overlook semantic depth and contextual understanding. To address this gap, we introduce chrF-S (Semantic chrF++), an enhanced metric that integrates sentence embeddings to evaluate translation quality more comprehensively. By combining traditional character and word n-gram analysis with semantic information derived from embeddings, chrF-S captures both syntactic accuracy and sentence-level semantics. This paper presents our contributions to the WMT24 shared metrics task, showcasing our participation and the development of chrF-S. We also demonstrate that, according to preliminary results on the leaderboard, our metric performs on par with other supervised and LLM-based metrics. By merging semantic insights with n-gram precision, chrF-S offers a significant enhancement in the assessment of machine-generated translations, advancing the field of MT evaluation. Our code and data will be made available at https://github.com/AnanyaCoder/chrF-S.

1 Introduction

In the rapidly advancing field of machine translation (MT), the need for robust and nuanced evaluation metrics has become increasingly critical. The evaluation landscape has expanded significantly in recent years, as evidenced by the WMT Metrics Shared Task, which provides a platform for meta-evaluating these metrics. Notably, in recent iterations of the WMT Metrics Shared Task, apart from learned metrics, lexical-based metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015, 2017) have consistently been regarded as baselines.

These metrics are widely appreciated for their language independence, which require no training and can be applied across diverse languages. However, they primarily address syntactic accuracy and often fall short in capturing the deeper semantic nuances and contextual relevance of translations.

The BLEU metric, with its reliance on modified precision of n-grams, provides a useful measure of how closely a machine-generated translation aligns with reference translations. Similarly, chrF enhances evaluation by incorporating character-level n-grams, offering greater sensitivity to morphological variations.

Despite these advancements, both metrics primarily focus on surface-level features, which can lead to incomplete assessments of translation quality, especially in complex linguistic contexts. To address these limitations, we propose **chrF-S** (Semantic chrF), an extension to the chrF++ metric, which leverages sentence embeddings to provide a more comprehensive evaluation by incorporating semantic analysis alongside traditional n-gram matching. Sentence embeddings (Reimers and Gurevych, 2019, 2020) encode entire sentences, thereby capturing the relationships between words, the structure of the sentence, and the broader contextual meaning. These offer rich semantic representations of sentences, enabling a deeper understanding of meaning and context. By merging these embeddings with chrF++'s character and word n-gram analysis, chrF-S aims to capture both the syntactic and semantic dimensions of translation quality.

This paper details our contributions to the WMT24 shared metrics task, where we have applied chrF-S to evaluate its effectiveness in comparison with existing metrics. We present our methodology of integrating semantic analysis into the chrF framework and discuss the preliminary results from the leaderboard, which indicate that chrF-S performs competitively with other super-

Proceedings of the Ninth Conference on Machine Translation, pages 470–474

vised and LLM-based metrics. Our findings suggest that chrF-S not only enhances the evaluation of translation quality by incorporating semantic understanding but also provides a significant advancement over traditional metrics.

2 chrF-S

The main idea behind chrF-S is to have a combination of character-level match, word-level match and sentence-level match to provide a more comprehensive evaluation of translation quality. While chrF++ (Popović, 2015, 2017) already accounts for character and word-level matches, we enhanced this metric by introducing a sentence-level matching component. **We achieved this by adding a sentence-level component that utilizes sentence embeddings to compute a cosine similarity score, representing the semantic closeness between the reference and translation**. This flow is clearly illustrated in the figure 1. This approach allows chrF-S to assess not only the surface-level accuracy of the translation but also its deeper semantic fidelity, making it a more robust and nuanced evaluation metric.

For our experiments, we employed the LaBSE (Feng et al., 2022) model to generate these sentence embeddings. The ChrF-S score is computed as per equation 2

chrF-S(ref, hyp) = $\alpha \cdot chrF++(ref, hyp) + (1-\alpha) \cdot CosSim\,(embed(ref), embed(hyp))$

In this equation, ref refers to the reference sentence, and hyp is the hypothesis (translation) sentence. $chrF++(ref, hyp)$ denotes the character- and word-level similarity from ChrF++. The function $embed$ represents the sentence embeddings, which are generated using a sentence embedding model[1]. $CosSim$ computes the cosine similarity between the sentence embeddings. Finally, α is the weighting factor used to balance these two components; in our experiments, we set $\alpha = 2$.

3 Experiments

In our experiments, we considered two datasets released by WMT i.e., Direct Assessments (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022) from 2017-2022 and MQM (Freitag et al., 2021a,b, 2022) assessments from 2020-2022. As the data is heavily skewed towards west-germanic languages.

[1]In this case, we used LaBSE

lp	#segments	#systems
en-de	187	1
en-ru	250	1
cs-uk	3322	17
en-es	240	1
en-zh	120	1
en-cs	202	1
en-ja	242	1
en-uk	226	1
en-is	157	1
en-hi	247	1
ja-zh	243	1
zh-en	239	1
de-en	212	1
es-en	217969	356
en-fr	27730	161
fr-en	17553	117
ru-en	4320	18
pt-en	146508	158
en-arz	27333	156
en-twi	3560	16
en-xho	3285	18
en-luo	1251	6
en-hau	2996	14
en-yor	4774	28
en-som	46340	140
yor-en	25948	156
en-kik	18962	114
ary-fr	19960	120
en-swh	22954	138
en-ibo	19960	120
Total	617290	1865

Table 1: WMT24 Metrics Shared Task Test Set Statistics

Test-set	#sentences	BLEU	BERTScore	chrF++	chrF-S
MQM-A	457	0.210	0.333	0.478	**0.481**
MQM-B	790	0.180	0.282	0.410	**0.423**
MQM-C	1399	0.140	0.222	0.329	**0.355**
MQM-D	2425	0.117	0.188	0.272	**0.313**
MQM-E	4242	0.099	0.110	0.200	**0.242**

Table 2: Pearson Correlation scores on five different test sets curated from WMT-MQM (20-22) data

Test-set	#sentences	BLEU	BERTScore	chrF++	chrF-S
DA-A	8903	0.186	0.208	0.290	**0.328**
DA-B	17663	0.183	0.209	0.290	**0.336**
DA-C	34715	0.180	0.191	0.288	**0.333**
DA-D	67487	0.180	0.179	0.285	**0.333**
DA-E	126957	0.191	0.188	0.291	**0.336**

Table 3: Pearson Correlation scores on five different test sets curated from WMT-DA (17-22) data

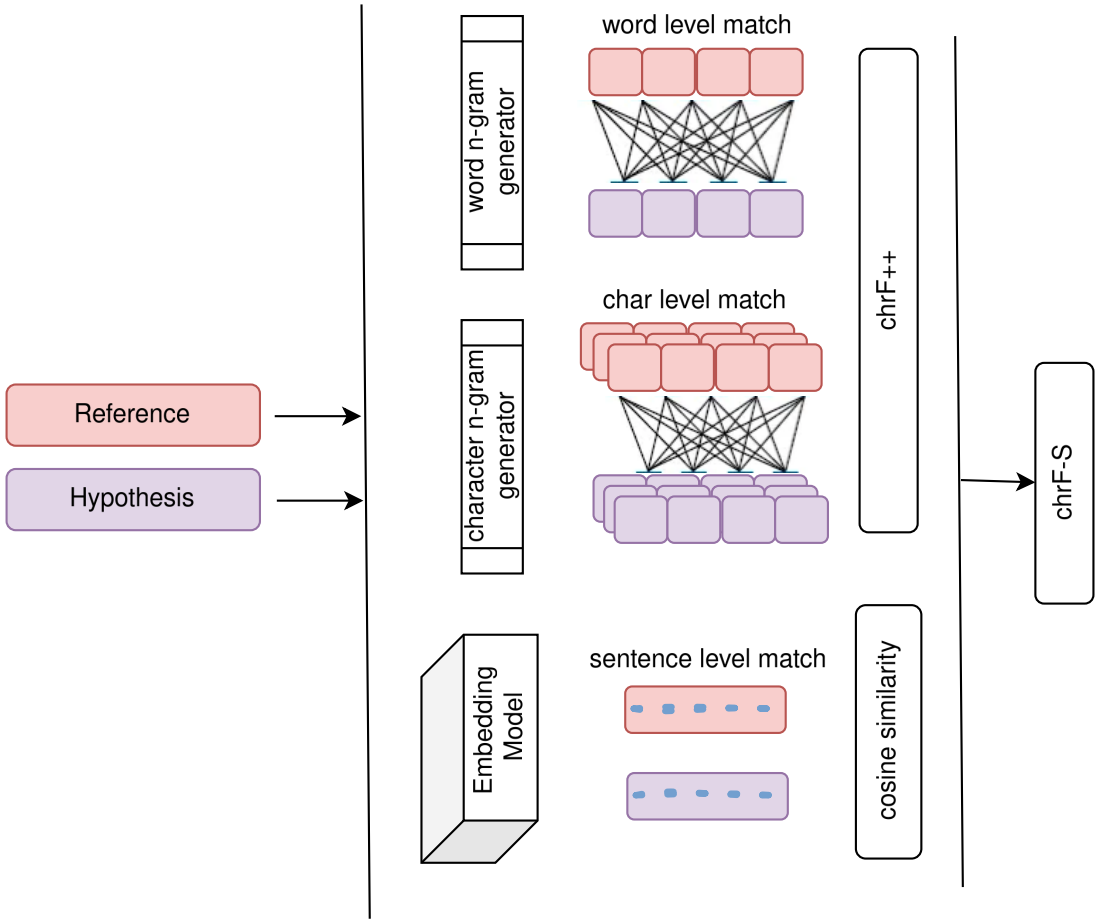

Figure 1: chrF-S Metric

We created five sub testsets[2] of different sizes having a fair distribution of sentences across all language pairs. We evaluated these five testsets (A, B, C, D, E) using **unsupervised reference based metrics:** BLEU, chrF++, BERTScore and chrF-S and further computed pearson (Kurtz and Mayo, 1979) correlation to compare the metrics in terms of their agreement with human judgements.

3.1 Evaluation

Table 2 reports the correlation scores of the metrics with MQM assessments on the testsets built from WMT-MQM (20-22) data. Similarly Table 3 displays the correlation scores of the metrics with direct assessments on the testsets created from WMT-DA (17-22) data. In both the tables, it is clearly evident that chrF-S has performed better. We notice that the correlation scores of chrF-S in WMT-DA testset is slightly less (<0.4), however when compared to other metrics it still stands as winner.

[2]code and testsets will be released

By incorporating sentence-level embeddings, chrF-S enhances its ability to evaluate the semantic closeness between the reference and translation, leading to better alignment with human judgments that prioritize meaning and context. This semantic dimension improves correlation scores with human assessments, making chrF-S a more accurate and reliable metric, especially when translations **differ lexically but are semantically equivalent**.

4 WMT24 Metrics Shared Task Participation

We have participated in the WMT24 Metric Shared Task by submitting the translation scores for official evaluation (en-de, en-es, ja-zh) and secondary evaluation (for all langauge pairs from the generalMT task). The test-set statistics are reported in Table 1.

The preliminary leaderboard for the official language pairs is released by the shared task is reported at Table 4, displaying the system-level Pearson correlations and segment-level Kendall Tau correlations of of en-de, en-es and ja-zh language

Rank	Participant	En-De sys-level Pearson	En-De seg-level Kendall	En-Es sys-level Pearson	En-Es seg-level Kendall	Ja-Zh sys-level Pearson	Ja-Zh seg-level Kendall
1	mengyao	1.0	0.85	1.0	0.82	1.0	0.98
2	jjuraska	1.0	0.57	0.99	0.59	0.99	0.55
3	gentaiscool	1.0	0.67	0.98	0.69	0.99	0.61
7	GEMBA-ESA	0.98	0.53	0.99	0.51	0.94	0.49
8	chrF-S	**0.97**	**0.51**	**0.99**	**0.5**	**0.97**	**0.56**
12	MetricsTaskBaseline	0.95	0.45	0.92	0.46	0.5	0.17

Table 4: WMT24 Prelimnary Leaderboard reporting system-level and segment-level correlations. Our metric correlations are highlighted in bold.

pairs. It is noteworthy that chrF-S has not only surpassed the baseline but also demonstrated performance on par with GEMBA, an LLM-based metric. When compared to other preceding supervised metrics, chrF-S, an unsupervised metric proves to be competitive, standing alongside other top performers in the field.

5 Conclusion

This paper contributes to the WMT24 metrics shared task by introducing chrF-S, an enhanced version of chrF++ that incorporates sentence-level semantics for more accurate MT evaluation. Our metric effectively captures both surface accuracy and deeper semantic meaning by integrating character-level, word-level, and sentence-level matching. The use of sentence embeddings enables chrF-S to better assess semantic closeness between translations and references, leading to improved correlation with human judgments such as MQM and direct assessments. Preliminary leaderboard results indicate that chrF-S is competitive with other leading metrics, underscoring its potential as a reliable and nuanced tool for evaluating translation quality.

Limitations

One significant limitation of this approach is its dependency on embedding models for sentence embeddings. The effectiveness of this method is restricted to languages for which appropriate sentence embedding models are available.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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. 2021b. 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.

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.

Albert K. Kurtz and Samuel T. Mayo. 1979. *Pearson Product Moment Coefficient of Correlation*, pages 192–277. Springer New York, New York, NY.

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*.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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*. Association for Computational Linguistics.