RESEARCH ARTICLE

signwriting-evaluation: Effective Sign Language Evaluation via SignWriting

Amit Moryossef[1], Rotem Zilberman[2], Ohad Langer[2]

1 University of Zürich, Switzerland
2 Bar-Ilan University, Israel

Abstract

The lack of automatic evaluation metrics tailored for SignWriting presents a significant obstacle in developing effective transcription and translation models for signed languages. This paper introduces a comprehensive suite of evaluation metrics specifically designed for SignWriting, including adaptations of standard metrics such as BLEU and chrF, the application of CLIPScore to SignWriting images, and a novel symbol distance metric unique to our approach. We address the distinct challenges of evaluating single signs versus continuous signing and provide qualitative demonstrations of metric efficacy through score distribution analyses and nearest-neighbor searches within the SignBank corpus. Our findings reveal the strengths and limitations of each metric, offering valuable insights for future advancements using SignWriting. This work contributes essential tools for evaluating SignWriting models, facilitating progress in the field of sign language processing. Our code is available at https://github.com/sign-language-processing/signwriting-evaluation.

Corresponding author: Amit Moryossef, amit@sign.mt

1. Introduction

Signed languages are the primary means of communication for deaf communities worldwide, possessing rich linguistic structures fundamentally different from spoken languages. SignWriting is a notation system developed to visually represent signed languages, capturing hand shapes, movements, facial expressions, and body positions[1]. Despite its potential to enhance sign language processing, the lack of specialized automatic evaluation metrics for SignWriting hinders the development of effective transcription and translation models[2].

Traditional evaluation metrics such as BLEU[3], chrF[4], and CLIPScore[5] have been extensively used in natural language processing tasks. However, these metrics are well-documented in their struggle to accurately assess the performance of models involving complex relations and meaningful nuances and therefore, they are not directly applicable to the visual-spatial SignWriting[6][7][5]. This discrepancy leads to inadequate assessments of model performance and often forces

researchers to develop ad-hoc evaluation metrics without proper validation[8].

To overcome these limitations, we present signwriting-evaluation , a comprehensive toolkit providing a suite of automatic evaluation metrics specifically tailored for SignWriting. Our toolkit adapts standard metrics such as BLEU and chrF for tokenized and untokenized Formal SignWriting (FSW) strings, respectively, enabling their application to sign language data. We also leverage CLIPScore by applying it to SignWriting images using the original CLIP model, bridging the gap between visual representations and textual evaluation. Moreover, we develop a novel symbol distance metric that accounts for the visual and spatial properties of SignWriting symbols, offering a better tailored assessment of transcription and translation models.

We demonstrate the efficacy of our evaluation metrics through qualitative analyses, including score distribution examinations and nearest-neighbor searches within the SignBank corpus, which contains around 230,000 unique single signs. Our findings highlight the strengths and limitations of each metric, providing valuable insights for future advancements in SignWriting processing.

By releasing signwriting-evaluation as an open-source toolkit, we aim to facilitate further research and development in the field of sign language processing. We believe that our work will help create more effective and accurate models for SignWriting transcription and translation.

2. Evaluation Metrics

To effectively evaluate SignWriting transcriptions and translations, we explore both traditional metrics (BLEU (§2.1), chrF (§2.2), and CLIPScore (§2.3)) adapted for SignWriting and introduce a new metric specifically designed to capture its unique characteristics (§2.4).

2.1. BLEU[3]

BLEU (Bilingual Evaluation Understudy) is a common metric for evaluating machine translation by comparing n-grams of the candidate translation with reference translations[9]. It measures the overlap of n-grams between the hypothesis and reference texts, providing a quantitative measure of their similarity.

To adapt BLEU for SignWriting, as a baseline metric, we tokenize the Formal SignWriting (FSW) strings into sequences of symbols. Each symbol in FSW represents a specific handshape, movement, facial expression, or other sign components, encoded as a combination of characters, which we treat as tokenized sequences. We use the BLEUimplementation from sacreBLEU[10].

However, applying BLEU directly to SignWriting presents challenges like requiring exact symbol position matches, where minor shifts cause mismatches and lower scores despite minimal visual differences[6]. It also ignores symbol similarity, treating tokens as discrete units. Moreover, the flexible symbol ordering in FSW can lead to lower scores even for the

same sign.

2.2. chrF[4]

chrF is a character n-gram F-score metric originally designed for machine translation evaluation. It computes the F-score based on the precision and recall of character n-grams between the hypothesis and reference texts. Since Formal SignWriting strings are composed of sequences of characters representing symbols, chrF can be directly applied without tokenization. We use the chrF implementation from sacreBLEU[10].

Using chrF to capture character-level similarities in SignWriting, can help assess minor symbol variations. However, like BLEU, it overlooks semantic meaning and visual representation of the symbols, potentially limiting its ability to correctly evaluate SignWriting outputs[7].

2.3. CLIPScore[5]

CLIPScore is an evaluation metric using the CLIP (Contrastive Language–Image Pre-training,[11]) model to compute the similarity between images and text. We adapt it for SignWriting by visualizing the script and computing the cosine similarity between image embeddings.

We use the original CLIP model without additional training, relying on its ability to capture visual similarities. Trained on a large image-text dataset, CLIP identifies visual similarities even when symbols differ, capturing aspects that token-based metrics may overlook. However, the train/test domain mismatch limits its ability to capture nuances, yielding similar embeddings for all SignWriting images with minor differentiation.

2.4. Symbol Distance Metric

To address the limitations of existing metrics above, we propose a novel metric specifically designed for SignWriting notation. This metric accounts for the visual and spatial properties of symbols, providing a more accurate evaluation of transcription and translation outputs.

Our metric assesses the similarity between two SignWriting signs by calculating distances between their constituent symbols based on attributes such as base shape, orientation, rotation, and position. We define a distance function that considers these attributes, weighted by their significance.

To compare two signs as sets of symbols, we use the Hungarian algorithm[12] to find the optimal matching that minimizes the total distance between symbols from the hypothesis and reference. The mean normalized distance $\bar{D}$ is calculated over the matched symbol pairs. We normalize the distances using a non-linear function to ensure they are in a consistent range, exponentially scaled by α which controls the normalization curve.

To account for differences in the number of symbols between signs, we compute a length adjustment factor, using the

number of symbols within each sign and a penalty severity control β:

$$L = \left| \frac{|S_{\text{hyp}}| - |S_{\text{ref}}||}{\max(|S_{\text{hyp}}|, |S_{\text{ref}}|) + 1} \right|^{\beta}$$

The final score between the hypothesis and reference signs is computed as $(1 - D)(1 - L)$, which is then normalized exponentially by parameter γ. This score ranges between 0 and 1, with higher values indicating greater similarity.

Our symbol distance metric provides a task-specific evaluation, capturing subtle differences in symbol attributes critical for conveying meaning in sign language. By incorporating both symbol-level distances and structural discrepancies, it offers a balanced and efficient assessment of model outputs.

3. Qualitative Evaluation

In this section, we analyze the effectiveness and limitations of the proposed metrics, focusing on both evaluation score distribution (§3.1) and performance in a nearest-neighbor search (§3.2).

3.1. Distribution of Scores

To understand how each metric behaves when comparing random pairs of signs, we computed the any-to-any scores for a random sample of 1,000 signs from the *SignBank*[1] corpus, which includes around 230,000 unique single signs. Intuitively, we expect random sign pairs to have low similarity scores due to their likely lack of relation. This should be reflected in a score distribution skewed towards lower values. Figure 1 shows the histograms of scores for each metric and reveals that:

- **BLEU** scores are concentrated near zero, due to random signs sharing few common tokens.
- **chrF** shows a wider spread, suggesting higher character-level overlap between random signs.
- **CLIPScore** yields scores clustered around a high value, due to the model's tendency to produce similar embeddings for different images[13].
- **Symbol Distance** results in a wide distribution of low scores, with a tail of higher scores, reflecting the low similarity between most random signs when considering symbol attributes and positions, and higher similarity between fewer signs.

These distributions help us understand the sensitivity and discriminative power of each metric in distinguishing between unrelated signs.
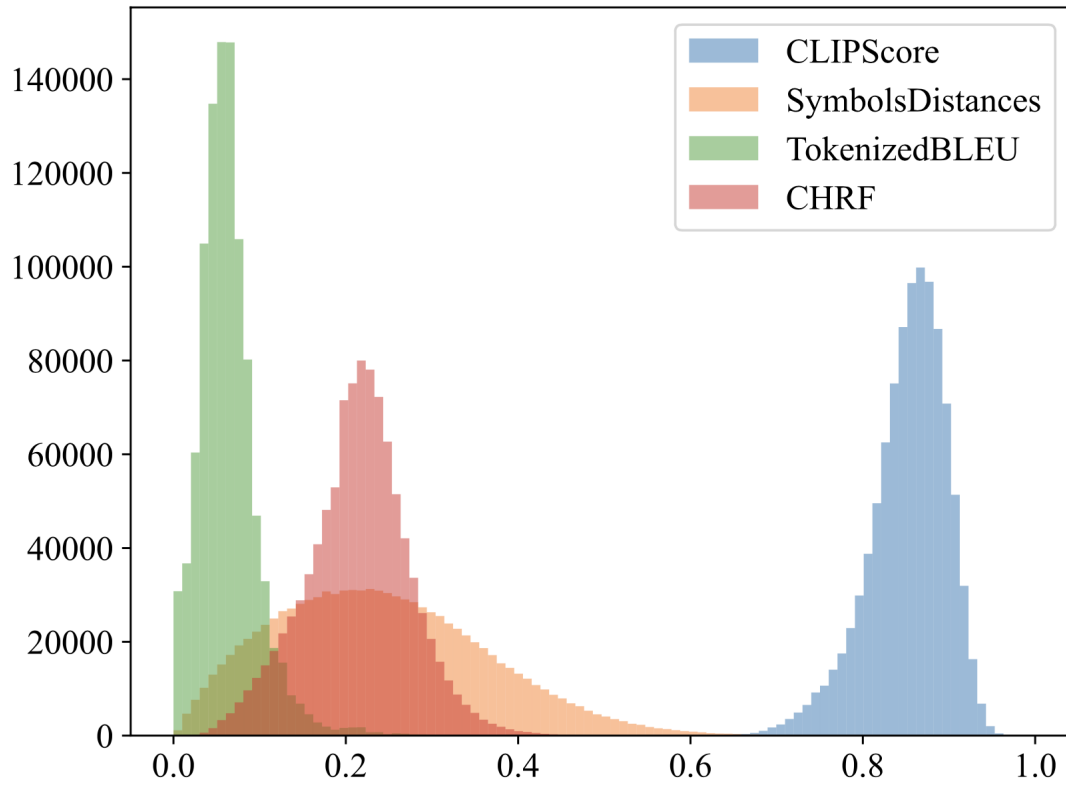
Figure 1. Distribution of scores for random pairs of signs using different evaluation metrics.

3.2. Nearest Neighbor Search

To further assess the practical effectiveness of the metrics, we conducted a nearest-neighbor search on the SignBank corpus. We selected multiple variations for the sign 'Hello' and retrievedtop 10 nearest-neighbors in the entire corpus for each metric. Table 1 shows the ranked results for the three reference signs across four metrics.

Table 1. Top 10 nearest neighbors for selected signs using different evaluation metrics. The reference signs are shown at the top, and the retrieved signs are displayed in order of decreasing similarity score from top to bottom.

- **BLEU** often misses the closest signs, due to its symbol order sensitivity and exact token match requirements.
- **chrF** performs slightly worse than modified BLEU, focusing on character-level overlaps but missing key patterns and visual similarities.
- **CLIPScore** often fails to differentiate subtle yet fundamental changes and assess closeness, likely due to the model's domain mismatch and tendency to produce similar embeddings.
- **Symbol Distance** consistently retrieves signs that are *more* visually and semantically similar to the reference, showing

its effectiveness in capturing SignWriting's nuances.

These qualitative results highlight the limitations of traditional metrics such as BLEU, chrF and CLIPScore for SignWriting, and demonstrate the advantage of utilizing a metric specifically designed to account for the properties of SignWriting.

4. Conclusions

In this paper, we introduced a comprehensive toolkit, *signwriting-evaluation*, providing specialized evaluation metrics for SignWriting strings. Our proposed Symbol Distance Metric is specifically designed to capture the unique characteristics of SignWriting, addressing the limitations of traditional metrics like BLEU, chrF, and CLIPScore when applied to sign language data.

Through qualitative evaluations, we demonstrated that our Symbol Distance Metric outperforms other metrics when assessing the similarity of SignWriting signs. It effectively accounts for symbol attributes, positions, and variations in symbol ordering, providing a more accurate and meaningful evaluation of model outputs.

We, therefore, encourage the research community to adopt our Symbol Distance Metric for evaluating SignWriting models, as it offers significant improvements over standard metrics not tailored for sign language processing. Using more appropriate tools will allow researchers to develop and assess models more effectively, ultimately advancing the field of sign language technology.

Our open-sourced toolkit is available at https://github.com/sign-language-processing/signwriting-evaluation. We welcome community contributions to enhance the toolkit, develop new metrics, and collaborate on advancing SignWriting evaluation, creating useful tools benefiting both researchers and the deaf community.

Limitations

We recognize that our current evaluation primarily emphasizes qualitative analysis. To further validate our metrics, a more rigorous quantitative analysis, aligned with human judgments, is necessary. This would provide deeper insights into how accurately the metric captures sign similarity and meaning.

There is also room for future enhancement. Incorporating machine learning techniques could optimize the weighting of symbols and better capture the subtle complexities of SignWriting.

Additionally, although our toolkit supports continuous signing by treating the sign sequence as a set and identifying the minimal match, expanding these metrics to better handle multi-sign sequences remains an important area for future research.

Footnotes

References

1. ^*Sutton V (1990). Lessons in sign writing. SignWriting.*

2. ^*Moryossef A, Jiang Z (2023). "SignBank+: Preparing a Multilingual Sign Language Dataset for Machine Translation Using Large Language Models". Published on Arxiv.*

3. a, b*Papineni K, Roukos S, Ward T, Zhu WJ (2002). "Bleu: a method for automatic evaluation of machine translation". Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. pp. 311–318. doi:10.3115/1073083.1073135.*

4. a, b*Popovi{\'c} M (2015). "chr{F}: character n-gram {F}-score for automatic {MT} evaluation". In: Bojar O, Chatterjee R, Federmann C, Haddow B, Hokamp C, Huck M, Logacheva V, Pecina P, editors. Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon, Portugal: Association for Computational Linguistics. p. 392–395. doi:10.18653/v1/W15-3049. Available from: https://aclanthology.org/W15-3049.*

5. a, b, c*Hessel J, Holtzman A, Forbes M, Le Bras R, Choi Y (2021). "CLIPScore: A reference-free evaluation metric for image captioning". Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing 7514–7528. doi:10.18653/v1/2021.emnlp-main.595.*

6. a, b*Reiter E (2018). "A structured review of the validity of BLEU".Computational Linguistics. **44** (3): 393–401. doi:10.1162/coli_a_00322.*

7. a, b*Kim JH, Huerta-Enochian M, Ko C, Lee DH (2024). "SignBLEU: Automatic evaluation of multi-channel sign language translation". Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pages 14796–14811, Torino, Italia. ELRA and ICCL.*

8. ^*Jiang Z, Moryossef A, Müller M, Ebling S (2023). "Machine translation between spoken languages and signed languages represented in SignWriting". Findings of the Association for Computational Linguistics: EACL 2023 pages 1661–1679, Dubrovnik, Croatia. Association for Computational Linguistics.*

9. ^*Blagec K, Dorffner G, Moradi M, Ott S, Samwald M (2022). "A global analysis of metrics used for measuring performance in natural language processing". Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP. Dublin, Ireland: Association for Computational Linguistics. pp. 52–63. doi:10.18653/v1/2022.nlppower-1.6.*

10. a, b*Post M (2018). "A call for clarity in reporting BLEU scores".Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels, Belgium: Association for Computational Linguistics. pp. 186–191. doi:10.18653/v1/W18-6319.*

11. ^*Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR; 2021. p. 8748-8763.*

12. ^*Kuhn HW (1955). "The Hungarian Method for the Assignment Problem".Naval Research Logistics Quarterly. **2** (1–2):*

83–97. doi:10.1002/nav.3800020109.

13. ^Ahmadi S, Agrawal A (2024). "An examination of the robustness of reference-free image captioning evaluation metrics". Findings of the Association for Computational Linguistics: EACL 2024 pages 196–208. https://aclanthology.org/2024.findings-eacl.14.