Qeios

Peer Review

# Review of: "signwriting-evaluation: Effective Sign Language Evaluation via SignWriting"

**Dimitar Shterionov**[1]

1. Cognitive Science and Artificial Intelligence, Tilburg University, Netherlands

General:

I find the idea of sign language-focused evaluation quite interesting and important for the SLMT and SLNLP community. Furthermore, assessing the use of established metrics like BLEU and chrF is even more beneficial - people would get a better idea of how these metrics behave when evaluating SL (represented as sign writing). And, third, the newly proposed metric has the potential to become a standard evaluation in SL (represented as sign writing, but even more, through the openly accessible repository, to be adopted to other formalisms).

However, the article lacks depth, and some parts of the methodology are not well explained or logical. These are listed below, along with some recommendations for improvement.

1. The motivation, which outlines the limitations and challenges of BLEU and chrF for SignWriting evaluation, lacks details. The presented limitations are not elaborated enough to motivate the work. Furthermore, some of these limitations originate from NLP for text / spoken language, and therefore it is difficult to draw parallels with the multi-linear nature of SLs. Perhaps the limitations in NLP for text can be mirrored as limitations in the evaluation of the serialised version of SignWriting (FSW).
2. It is not clear to me why it is mentioned that “researchers develop ad-hoc evaluation metrics without proper validation,” citing [8], as in [8] the BLEU, chrF2++, and MSE metrics are used. Perhaps an example should be given.
3. The introduction mentions adaptations of BLEU and chrF; however, what is noted in Section 2 is that the FSW is tokenized rather than modifying the BLEU and chrF scores. These are adaptations

 doi.org/10.32388/9JAB3P 

of the input sequence rather than of the metrics themselves.

4. The symbol distance metric needs more details in its formalisation. The formula or formulae presenting its calculation should all be present and not only the one on page 4.
5. The claim at the end of Section 2, “Our symbol distance metrics provides a task-specific evaluation, capturing subtle differences in symbol attributes critical for conveying meaning in sign language,” is not well supported. Except for the statement “This metric accounts for the visual and spatial properties of symbols...," there is nothing to support the aforementioned claim. Even more, such a statement should (at least) relate symbols to meaning to somehow make the claim valid.
6. Explanation about what the flexible symbol ordering of FSW is is missing. Perhaps add a footnote. Also, perhaps a footnote about how the symbol similarity is computed could be helpful.
7. Experiment 1: Distribution of Scores. While I do understand this experiment setup, it is difficult to put the results into a frame, as there is no indication about how similar or dissimilar these signs are in order to judge the validity of the scores. That is, even if the meaning of two signs is different, but the SignWriting symbols they can be represented with have an overlap, then their BLEU score (when treating the one as a source and the other as a target) will be high. For example, the sign for person and week in NGT are the same with the exception of the mouthing. Or the other way round – when signs have similar meanings but completely different signs, then their BLEU scores will be low; for example, the signs for being sick and for a lesson (again in NGT).
8. In Section 3.2, closeness is not defined.

Additional comments, grammar, and spelling:

- Introduction:
  - “... effective transcription and translation models[2].” → “...effective transcription and translation models using or based on SignWriting.” (in the sense that the work assesses SignWriting and it is difficult to claim wider impact without).
  - The sentence “However, these metrics are ... to the visual-spatial SignWriting [6][7][5].” is not clear.
  - Add a citation at the first mention of the CLIP model (...it to SignWriting images using the original CLIP model,...)
- The abbreviations MT and FSW are not used consistently.

- Evaluation metrics:
  - “...the candidate translation with reference translations[9].” → “...the candidate translation to reference translations[9].”
  - “...between symbols from the hypothesis and the reference.”
  - The bar on top of the D is too high, affecting the line space.
- Qualitative Evaluation:
  - “... subtle yet fundamental changes and assess...” → “... subtle yet fundamental changes and to assess...”
- Conclusions:
  - “...toolkit,signwriting-evaluation, providing specialized evaluation metrics” → “...toolkit, signwriting-evaluation, providing specialized evaluation metrics”

## Declarations

**Potential competing interests:** No potential competing interests to declare.