# Attention is All You Sign: Sign Language Translation with Transformers

Kayo Yin*[1,2] and Jesse Read[1]

[1] LIX, École Polytechnique, Institut Polytechnique de Paris, France
[2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
kayo@cmu.edu, jesse.read@polytechnique.edu

**Abstract.** This paper improves the translation system in Sign Language Translation (SLT) by using Transformers. We report a wide range of experimental results for various Transformer setups and introduce a novel end-to-end SLT system combining Spatial-Temporal Multi-Cue (STMC) and Transformer networks. Our methodology improves on the current state-of-the-art by over 5 and 7 BLEU respectively on ground truth (GT) glosses and predicted glosses of the PHOENIX-Weather 2014T dataset. On the ASLG-PC12 corpus, we report an improvement of over 16 BLEU. Our findings also reveal that end-to-end translation with predicted glosses outperforms translation on GT glosses. This shows the potential for further improvement in SLT by either jointly training the SLR and translation systems or by revising the gloss annotation scheme.

## 1 Introduction

In SLT, a tokenization system first generates glosses from sign language videos. Then, a translation system translates the recognized glosses into spoken language. Recent efforts work on the first step, but there has been none improving the translation system. This paper aims to fill this research gap by leveraging recent success in Neural Machine Translation (NMT), namely Transformers.

The contributions of this paper can be summarized as: (i) The first successful application of Transformers to SLT achieving state-of-the-art results; (ii) A novel STMC-Transformer model for end-to-end translation surpassing GT glosses translation contrary to previous assumptions; (iii) The first usage of weight tying, transfer learning, and ensemble learning in SLT.

## 2 Related Work

### 2.1 Sign Language Translation

[4] jointly use a 2D-CNN model to extract gloss-level features from video frames, and a sequence-to-sequence (seq2seq) model for German SLT. Subsequent works

*Work carried out while at École Polytechnique. The Titan X Pascal used for this research was donated by the NVIDIA Corporation.

all focus on improving tokenization rather than translation. A contemporaneous paper [5] jointly trains Transformers for both tokenization and translation. [8] estimate human keypoints to extract glosses, then use seq2seq models for Korean SLT. [1] use seq2seq models to translate ASL glosses [10].

### 2.2 Neural Machine Translation

Neural Machine Translation (NMT) employs neural networks for text translation. Recent methods typically use seq2seq models with an encoder and decoder.

Earlier approaches use recurrent [11] for the encoder and decoder. Recent works use attention mechanisms [2, 9] that calculates context-dependent alignment scores between encoder and decoder hidden states. [12] introduces the Transformer, a seq2seq model relying on self-attention instead of recurrence.

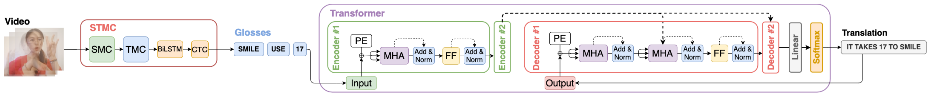

Fig. 1: STMC-Transformer network for end-to-end SLT

## 3 Model architecture

### 3.1 Spatial-Temporal Multi-Cue (STMC) Network

We use the STMC network [13] for tokenization. A spatial multi-cue (SMC) module decomposes the input video into spatial features of multiple visual cues. Then, a temporal multi-cue (TMC) module calculates temporal correlations within and between cues at different time steps. Obtained features are analyzed by Bi-directional Long Short-Term Memory (BiLSTM) [11] and Connectionist Temporal Classification (CTC) [7] units for sequence learning and inference.

### 3.2 Transformer

For translation, we train a two-layered Transformer [12] to maximize the log-likelihood $\sum_{(x_i,y_i)\in D} \log P(y_i|x_i,\theta)$, where $D$ contains gloss-text pairs.

## 4 Datasets

**PHOENIX-Weather 2014T** [4] is extracted from weather forecast airings of the German tv station PHOENIX. **ASLG-PC12** [10] is constructed from English data of Project Gutenberg that has been transformed into ASL glosses following a rule-based approach. We make our data and code publicly available[1].

[1] https://github.com/kayoyin/transformer-slt

# 5 Experiments and Discussions

We organize our experiments into two groups: (i) Gloss2Text in which we translate GT gloss annotations to simulate perfect tokenization; (ii) Sign2Gloss2Text where we perform end-to-end translation with the STMC-Transformer.

## 5.1 Gloss2Text (G2T)

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | METEOR |
|---|---|---|---|---|---|---|
| Raw data | 11.88 | 5.05 | 2.41 | 1.36 | 22.81 | 12.12 |
| RNN Seq2seq [4] | 44.13 | 31.47 | 23.89 | 19.26 | 45.45 | – |
| Transformer [5] | **48.90** | 36.88 | 29.45 | 24.54 | – | – |
| Transformer | 47.69 | 35.52 | 28.17 | 23.32 | 46.58 | 44.85 |
| Transformer Ens. | 48.40 | **36.90** | **29.70** | **24.90** | **48.51** | **46.24** |

(a) PHEONIX-WEATHER 2014T

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | METEOR |
|---|---|---|---|---|---|---|
| Raw data | 54.19 | 39.26 | 28.44 | 20.63 | 75.59 | 61.65 |
| Preprocessed data | 68.82 | 56.36 | 46.53 | 38.37 | 83.28 | 79.06 |
| Seq2seq [1] | 86.7 | 79.5 | 73.2 | 65.9 | – | – |
| Transformer | 92.98 | **89.09** | 85.63 | 82.41 | 95.87 | **96.46** |
| Transformer Ens. | **92.88** | 89.22 | **85.95** | **82.87** | **96.22** | 96.60 |

(b) ASLG-PC12

Table 1: G2T results

We find that smaller models are better suited for our data of limited size, and larger batch size gives better performance. On ASLG-PC12, transfer learning using English fastText [3] vectors improves performance, while weight-tying the decoder gives the best BLEU-4, likely as it provides regularization. In Table 1, our Transformer Ensemble improves the state-of-the-art on PHOENIX-Weather 2014T and ASLG-PC12 by 5 and 17 BLEU-4 respectively. [5] uses a similar architecture to our single Transformer and obtains comparable results.

## 5.2 German Sign2Gloss2Text (S2G2T)

| | Dev Set | | | | | | Test Set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | METEOR | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | METEOR |
| RNN G2T [4] | 44.40 | 31.93 | 24.61 | 20.16 | 46.02 | – | 44.13 | 31.47 | 23.89 | 19.26 | 45.45 | – |
| RNN S2G2T [4] | 42.88 | 30.30 | 23.03 | 18.40 | 44.14 | – | 43.29 | 30.39 | 22.82 | 18.13 | 43.80 | – |
| Transformer G2T | 48.85 | 36.62 | 29.23 | 24.38 | 49.01 | 46.96 | 48.40 | 36.90 | 29.70 | 24.90 | 48.51 | 46.24 |
| Transformer S2G2T [5] | 47.73 | 34.82 | 27.11 | 22.11 | – | – | 48.47 | 35.35 | 27.57 | 22.45 | – | – |
| STMC-Bahdanau RNN | 45.89 | 32.24 | 24.93 | 20.52 | 44.46 | 43.48 | 47.53 | 33.82 | 26.07 | 21.54 | 45.50 | 44.87 |
| STMC-Luong RNN | 45.61 | 32.54 | 26.33 | 21.00 | 46.19 | 44.93 | 47.08 | 33.93 | 26.31 | 21.75 | 45.66 | 44.84 |
| STMC-Transformer | 48.27 | 35.20 | 27.47 | 22.47 | 46.31 | 44.95 | 48.73 | 36.53 | 29.03 | 24.00 | 46.77 | 45.78 |
| STMC-Transformer Ens. | **50.31** | **37.60** | **29.81** | **24.68** | **48.70** | **47.45** | **50.63** | **38.36** | **30.58** | **25.40** | **48.78** | **47.60** |

Table 2: SLT performance using STMC for CSLR.

**Recurrent sequence-to-sequence networks** We first train seq2seq models with Gated Recurrent Units [6] and Luong [9] or Bahdanau [2] attention. Surprisingly, these outperform previous recurrent models that translate GT glosses.

**Transformer** We train Transformer models with the same architecture as in G2T. Parameter search yields an initial learning rate 1 with 3,000 warm-up steps and beam size 4. Finally, we use our 8 best models in ensemble decoding.

STMC-Transformer also outperforms Transformers that translate GT glosses. While STMC performs imperfect CSLR, its gloss predictions may be better processed by the Transformer. Again, glosses are merely a simplified intermediate representation of the actual sign language, so it is not entirely unexpected that translating GT glosses does not give the best performance. Moreover, Transformers outperform recurrent networks in this setup as well and STMC-Transformer improves the state-of-the-art for video-to-text translation by 7 BLEU-4.

## 6 Conclusions and Future Work

In this paper, we proposed Transformers for SLT, notably the STMC-Transformer. Our experiments using different setups demonstrate how Transformers obtain better SLT performance than previous RNN-based networks. We also achieve new state-of-the-art results on different translation tasks on various datasets.

We especially obtained better performance by using a STMC network for tokenization instead of translating GT glosses. As future work, we suggest either training a CSLR model to output glosses easily usable by an NMT model, or design a novel gloss annotation scheme that optimizes translation.

## References

1. Arvanitis, N., Constantinopoulos, C., Kosmopoulos, D.: Translation of sign language glosses to text using sequence-to-sequence attention models (2019)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. Camgoz, N., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation (2018). https://doi.org/10.1109/CVPR.2018.00812
5. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation (2020)
6. Chung, J., Çaglar Gülçehre, Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv **abs/1412.3555** (2014)
7. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks (2006). https://doi.org/10.1145/1143844.1143891
8. Ko, S., Kim, C., Jung, H., Cho, C.: Neural sign language translation based on human keypoint estimation (2019). https://doi.org/10.3390/app9132683
9. Luong, M.T., Pham, H., Manning, C.: Effective approaches to attention-based neural machine translation (08 2015). https://doi.org/10.18653/v1/D15-1166
10. Othman, A., Jemni, M.: English-asl gloss parallel corpus 2012: Aslg-pc12 (2012)
11. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems **4** (09 2014)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
13. Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for continuous sign language recognition. Proceedings of the AAAI Conference on Artificial Intelligence (07) (2020)