Experimental Machine Translation of the Swiss German Sign Language via 3D augmentation of body keypoints

Lorenz Hufe and Eleftherios Avramidis
German Research Center for Artificial Intelligence (DFKI Berlin)
Speech and Language Technology, Alt Moabit 91c, 10559 Berlin
lorenz.hufe@web.de, eleftherios.avramidis@dfki.de

Abstract

This paper describes the participation of DFKI-SLT at the Sign Language Translation Task of the Seventh Conference of Machine Translation (WMT22). The system focuses on the translation direction from the Swiss German Sign Language (DSGS) to written German. The original videos of the sign language were analyzed with computer vision models to provide 3D body keypoints. A deep-learning sequence-to-sequence model is trained on a parallel corpus of these body keypoints aligned to written German sentences. Geometric data augmentation occurs during the training process. The body keypoints are augmented by artificial rotation in the three dimensional space. The 3D-transformation is calculated with different angles on every batch of the training process.

1 Introduction

Despite the enormous progress of the Machine Translation (MT) of spoken (and written) languages, the MT of sign languages is in a very early stage (Yin et al., 2021; De Coster et al., 2022). Two major challenges are (a) the multimodal and multilateral nature of the sign languages and (b) the lack of data. On the one side, the multilateral and multimodal nature of the sign languages requires deep-learning topologies that differ substantially from the ones used in text-based MT. On the other side, the lack of data makes difficult the utilization of end-to-end deep learning algorithms, which usually require vast amounts of data. As a result, deep-learning experiments have been executed for very few sign languages (e.g. German Sign Language, DGS; American Sign Language, ASL) and narrow domains (e.g. weather forecasts), leaving open questions on the generalization of the methods to other sign languages and broader domains.

This year's Sign Language Translation (SLT) Task of the Seventh Conference of Machine Translation (WMT22) is contributing significant to this direction, by adding a new language pair (Swiss German Sign Language - DGSG - to German) and allowing extensive experimentation from several participants on the same dataset.

Our system uses computer vision models to analyze the sign language videos into body keypoints and uses these keypoints as the source-side input of the neural MT transformer, allowing to perform data augmentation via geometrical augmentations. Despite the difficulty of this shared task and the low results obtained, we publish this paper as a technical report, with the hope that it can contribute to the further research of this direction.

The rest of the paper is organized as following. Section 2 positions our contribution amidst related work. Section 3 describes the methods for training the system and Section 4 the technical set-up of the experiment. Section 5 provides and discusses some results, while Section 6 gives some conclusion and ideas for further work.

2 Related Work

Latest work on MT of sign languages has shown significant improvements using deep learning methods from the fields of computer vision and MT. State of the art work (Camgöz et al., 2018; Yin and Read, 2020; Camgöz et al., 2020; Zhou et al., 2022) employees transformers, which are given frame embeddings extracted from the videos of the signers.

Contrary to the use of pixel-based frame embeddings, Nunnari et al. (2021) suggests to use body keypoints from the hands, the skeleton and the face as input to the transformers. This requires to split the translation pipeline into a first phase, recognizing 3D keypoints from videos, and has the advantage that they can be augmented by applying transformation techniques. Our paper presents an implementation of that idea, applied to the case of DSGS.

The use of body keypoints has been considered

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 983 - 988

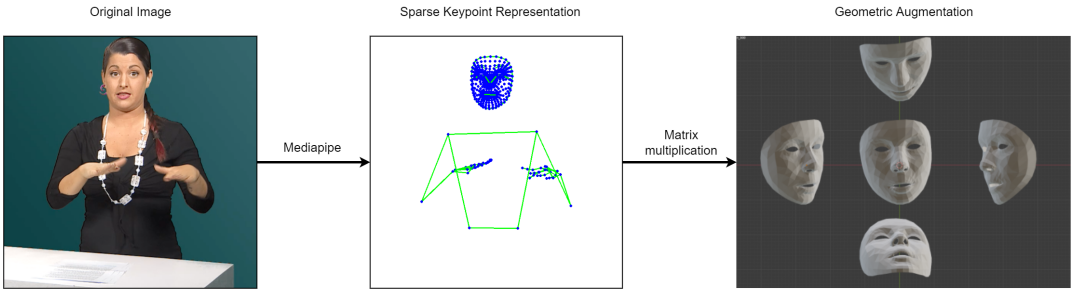

Figure 1: Mediapipe is used to extract sparse keypoint representations of the signer. The nature of the resulting 3D data allows for rotation, translation and shearing using matrix multiplication at virtually no cost.

by Gan et al. (2021), where skeleton pose information is processed together with the video frame input. Ko et al. (2018, 2019) use 2D coordinates of body keypoints to train the neural MT systems, but contrary to our work, they do not perform any geometrical transformations to the keypoints. Moryossef et al. (2021) analyze the applicability of the pose estimation systems to sign language recognition by evaluating the failure cases of the recognition models.

3 Method

Our system consists of three modules. The first module converts images of the signer into intermediate keypoint representations. The second module employs data augmentation to increase sample efficiency and decrease the effect of spurious feature correlations. Spurious data correlation in high dimensional spaces can lead to Clever Hans effects (Kauffmann et al.). The last module is the trainable transformer that translates from keypoint representation to German text, while interacting with the augmentation module.

3.1 Keypoint extraction

There are multiple reasons to believe that keypoint representations could prove beneficial in SLT. Only few and small datasets are available for SLT. That is because firstly there are only few known data sources for SL. Secondly the data transcription for SL needs expert knowledge which is costly and hard to find. Thirdly SL data inherently needs video footage of signing human, which makes anonymisation near impossible thus leads to privacy problems when detecting new potential data sources.

A end-to-end SLT pipeline needs to make sense of the movement of the human signer and translate these motions into written language. Practically speaking this means the pipeline internally needs to learn two tasks on limited data. However only the translation task depends on the costly and limited SLT datasets, while the task of detecting the motion could be eased by employing pose estimation which is not specific to SLT and therefore is more explored and cheaper in terms of data acquisition.

The extraction of the keypoints was done by using the computer vision models of MediaPipe Holistic (Grishchenko and Bazarevsky, 2020) which combines three pre-trained computer vision pipelines that detect the hand keypoints (MediaPipe Hands; Zhang et al., 2020), the keypoints of the body pose (BlazePose; Bazarevsky et al., 2020), and a keypoint mesh for the face (BlazeFace; Bazarevsky et al., 2019).

When data points were missing, the values were substituted by zero values.

3.2 Geometrical transformation

The geometrical transformation is applied during the training process of the transformer model. For every iteration of the training process, the 3D keypoints are given to the geometrical transformation module. This returns the co-ordinates of the original keypoint mesh after being rotated. The 3D keypoints get rotated around the x, y and z axis by some angle R_x, R_y and R_z respectively, using rotation matrices. First, the rotation around the x-axis takes place, followed by y and then the z axis.

The rotation angle is drawn at random at every training iteration, such that every batch is rotated to a different setting. The angle of the rotation is limited to a particular range, which makes sense for the particular axis. R_x is drawn from [-60°, +60°] while R_y and R_z are drawn from [-10°, +10°].

3.3 Sequence-to-sequence model

The sequence-to-sequence model is based on a NMT transformer model similar to (Camgöz et al.,

2018). We provide the network the keypoint representation, by concatenating all mediapipe keypoints and then flattening them into a 708 dimensional vector. The target language is the Swiss German text.

4 Experiment setup

The experiment took place using only the corpora FocusNews, as provided by the shared task organizers, including keypoints precomputed with MediaPipe. Due to time restrictions, the SRF corpus was not used, since it did not provide any keypoints. The training set had 10,136 sentences, the validation set 420 sentences and the test set 488 sentences. Due to problems with the keypoint-subtitle alignment only 393 of the 420 sentences of the validation set were used.

For training the model we modified the NMT toolkit JoeyNMT[1] (Kreutzer et al., 2019), extending the SLT branch created by Camgöz et al. (2020). We followed the text pre-processing of the previous implementation, which included text lowercasing. The geometrical transformations were done with array computations using NumPy (Harris et al., 2020). The automatic evaluation metrics were computed using SacreBLEU (Post, 2018).

In order to optimize the system we ran several experimental rounds. The training parameters for all rounds can be seen in Table 2. The experimental rounds were run by modifying the following parameters:

- **max. rotation**: The maximum angle for the random rotation that took place for every iteration. A max. rotation of 10° here means that for every iteration batch, a random degree value within [-10°, +10°] was drawn.
- **patience**: The learning rate scheduler stops when no significant progress is measured with the evaluation metric, after a number of epochs. This parameter defines how patient the scheduler is in that regards.
- **LR scheduler metric**: The metric used for measuring the progress on the validation set.
- **layers**: The number of layers for the encoder and the decoder of the transformer.

5 Results

As part of our parameter we ran 5 experimental rounds which are shown in Table 1. Due to time limitations it was not possible to experiment with the full spectrum of parameters, including ablation tests which would indicate the contribution of possible parameter values. Even in that case, the very low metric scores would not lead to more significant conclusions.

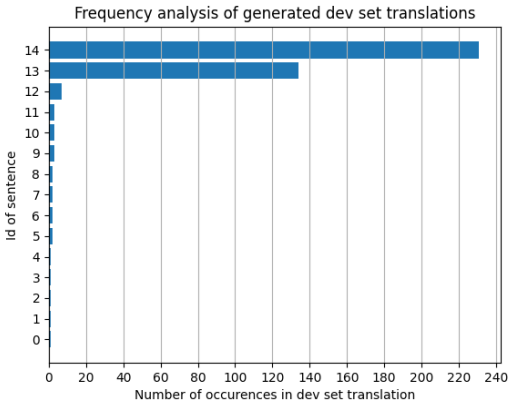

Figure 2: Overview over the translation sentence frequencies over the dev set

From the first experiments it was obvious that the use of BLEU-4 as a validation metric could not contribute to the optimization, because its values are always zero and also the training time was very short. For this reason we chose ChrF as validation metric for our last two experiments. Increasing the patience deemed necessary, so that the training mechanism can get enough random samples from the augmentation process. For our best iteration we experimented with both 3 and 4 layers, resulting into slightly better performance with the 4 layer setting.

In overall, the results of our experiments, as measured by automatic metrics, showed very low performance. No version of our pipeline could achieve non-zero BLEU-4 score on the provided development set, meaning that no n-gram of order 4 was correctly matched between the hypothesis and the reference. The experiments measured with BLEU-3 and ChrF indicate as better run the configuration with 60 degrees rotation range at the X axis, 10 degrees on the other axes, and a patience of 500. When analyzing the output on the validation set we found that for the 393 different sentences of the validation set, only 15 different translations were repeatedly produced as highlighted in figure 2 and listed in Appendix A. The two most common translations make up for 92% of the cases. This behaviour suggests that the model learned two main

[1]Our code is available at https://github.com/DFKI-SignLanguage/slt under Apache 2.0 License

max rotation +/- (°)			LR scheduler			scores			
x	y	z	patience	metric	layers	BLEU-3	BLEU-4	ChrF	runtime (h)
10	10	10	25	BLEU	4	0,28	0,00	15,36	00:21
10	10	10	50	BLEU	4	0,28	0,00	15,36	00:31
60	10	10	50	BLEU	4	0,00	0,00	17,58	00:24
60	10	10	500	ChrF	3	0,310	0,00	16,08	07:44
60	10	10	500	ChrF	4	0,314	0,00	16,43	04:14

Table 1: Overview over the results on the validation set when employing different settings.

parameter	value
feature size	708
max sentence len.	400
dropout	0,1
FF size	2048
heads	8
embeddings dim.	512
hidden size	512
optimizer	adam
batch size	32
random seed	42
weight decay	0,001
learning rate	0,001
validation freq.	100
beam size	1
beam alpha	-1
translation max len.	30

Table 2: Training parameters

prototype translations and is not sensitive to the input when translating.

6 Conclusion and Further Work

Due to the poor results, very little can be concluded about the effect of the proposed geometric augmentation strategy. As suggested by the preliminary results of the shared task (Müller et al., 2022) no group was able to achieve good results on the task. Unfortunately, due to the strict workshop timeline we could not perform further experiments to empirically prove the causes of this low performance. We are planning to do this in future work, including an ablation study of the different modules and a comparison with the state-of-the-art on other datasets. Further research should be invested in exploring the possible use cases for geometric data augmentation in MT of SL.

Acknowledgements

This work has been funded by the German Ministry of Education and Research through the project SocialWear (01IW20002). Many thanks to our colleagues Cristina España i Bonet, Yasser Hamidullah and Fabrizio Nunnari for providing ideas and support for this experiment.

References

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. *CoRR*, abs/2006.1.

Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. *CoRR*, abs/1907.0.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, San Francisco, CA, USA. IEEE.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10020–10030. Institute of Electrical and Electronics Engineers (IEEE).

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2022. Machine Translation from Signed to Spoken Languages: State of the Art and Challenges.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. 2021. Skeleton-Aware Neural Sign Language Translation. *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4353–4361.

Ivan Grishchenko and Valentin Bazarevsky. 2020. MediaPipe Holistic — Simultaneous Face, Hand and Pose Prediction, on Device.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Jacob Kauffmann, Lukas Ruff, Grégoire Montavon, and Klaus-Robert Müller. The clever hans effect in anomaly detection.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683.

Sang-Ki Ko, Jae Gi Son, and Hyedong Jung. 2018. Sign language recognition with recurrent neural network using human keypoint detection. In *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems*, RACS '18, page 326–328, New York, NY, USA. Association for Computing Machinery.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Amit Moryossef, Ioannis Tsochantaridis, Joe DInn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. 2021. Evaluating the Immediate Applicability of Pose Estimation for Sign Language Recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 3429–3435, Nashville, TN, USA. IEEE Computer Society.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regua Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2022. Findings of the WMT 2022 shared task on sign language translation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabrizio Nunnari, Cristina España-Bonet, and Eleftherios Avramidis. 2021. A data augmentation approach for sign-language-to-text translation in-the-wild. In *Proceedings of the 3rd Conference on Language, Data and Knowledge*, volume 93 of *OpenAccess Series in Informatics, OASIcs*, Zaragoza, Spain. Dagstuhl publishing.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. MediaPipe Hands: On-device Real-time Hand Tracking. *CoRR*, abs/2006.1.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2022. Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation. *IEEE Transactions on Multimedia*, 24:768–779.

Appendix

A Translations

0. **Empty**
1. die eltern sind sehr engagiert und kämpfen für die pille verbieten.
2. das ziel der konferenz sind vorträge von swisscom zu zeigen, dass diese kinder noch nicht gebärdensprache.
3. das ziel der konferenz sind vorträge von swisscom zu zeigen, dass diese kinder noch nicht zugänglich.
4. das ziel der konferenz sind vorträge von swisscom zu zeigen, dass sie sich nicht mit einer behinderung einsetzen.
5. die postverteilungs-firma
6. bis zum nächsten mal.
7. die eltern sind sehr engagiert und kämpfen für die gebärdensprache, ihre tochter hat.
8. das ziel der swisscom ist eine optimale beratung und einen guten service anzubieten.
9. die gehörlosen kinder freuten sich sehr, da sie alles verstanden und somit integriert geschult integriert geschult werden schulen sollen.
10. die eltern sind gehörlos.
11. die eltern sind sehr engagiert und kämpfen für die hochschule.
12. die forscher meinen, dass kinder mit cochlea-implantate über eine genauso gute lebensqualität wie hörende kinder verfügen, ohne psychosoziale folgen.

13. die eltern sind sehr engagiert und kämpfen für die gebärdensprache, ihre kultur und ihre rechte.
14. die voraussetzungen für diese stelle sind ein kürzlich abgeschlossenes hochschulstudium sowie die bereitschaft, arbeiten im sinne der gleichstellung zu schreiben.