- ORIGINAL ARTICLE -

A study on pose-based deep learning models for gloss-free Sign Language Translation

Estudio sobre modelos de aprendizaje profundo basados en poses para Traducción de Lengua de Señas sin glosas

Pedro Dal Bianco[1,3], Gastón Ríos[1,3], Waldo Hasperué[1,2], Oscar Stanchi[1,4], Facundo Quiroga[1,2], and Franco Ronchetti[1,2]

[1]*Instituto de Investigación en Informática LIDI, Universidad Nacional de La Plata, Argentina*
{pdalbianco, grios, whasperue, ostanchi, fquiroga, fronchetti}@lidi.info.unlp.edu.ar
[2]*Comisión de Investigaciones Científicas de la Pcia. de Bs. As. (CIC-PBA), Argentina*
[3]*Becario Doctoral UNLP*
[2]*Becario Doctoral CONICET*

Abstract

Sign Language Translation (SLT) is a challenging task due to its cross-domain nature, different grammars and lack of data. Currently, many SLT models rely on intermediate gloss annotations as outputs or latent priors. Glosses can help models to correctly segment and align signs to better understand the video. However, the use of glosses comes with significant limitations, since obtaining annotations is quite difficult. Therefore, scaling gloss-based models to millions of samples remains impractical, specially considering the scarcity of sign language datasets. In a similar fashion, many models use video data that requires larger models which typically only work on high end GPUs, and are less invariant to signers appearance and context. In this work we propose a gloss-free pose-based SLT model. Using the extracted pose as feature allow for a sign significant reduction in the dimensionality of the data and the size of the model. We evaluate the state of the art, compare available models and develop a keypoint-based Transformer model for gloss-free SLT, trained on RWTH-Phoenix, a standard dataset for benchmarking SLT models alongside GSL, a simpler laboratory-made Greek Sign Language dataset.

Keywords: Deep Learning, Gloss-free, Pose Estimation, Sign Language Datasets, Sign Language Translation.

Resumen

La Traducción de Lenguaje de Señas es una tarea desafiante ya que atraviesa múltiples dominios, diferentes gramáticas y falta de datos. Actualmente, muchos modelos de SLT dependen de glosas como anotaciones intermedias o salidas. Estas pueden ayudar a los modelos a segmentar y alinear correctamente las señas para comprender mejor el video. Sin embargo, su uso conlleva limitaciones significativas, ya que obtenerlas es bastante difícil. Por lo tanto, escalar modelos basados en glosas a millones de muestras sigue siendo impráctico, especialmente considerando la escasez de bases de datos de lengua de señas. De igual forma, muchos modelos utilizan videos como entrada, lo que requiere de modelos más grandes que típicamente solo funcionan en GPUs de alta gama y son menos invariantes a la apariencia y el contexto de los señantes. En este trabajo proponemos un modelo de SLT basado en poses y sin glosas. Usar la pose extraída como entrada permite una reducción significativa en la dimensionalidad de los datos y en el tamaño del modelo. Evaluamos el estado del arte, comparamos modelos disponibles y desarrollamos un modelo Transformer basado en *keypoints* para SLT sin glosas, entrenado sobre RWTH-Phoenix, un conjunto de datos estándar para la evaluación de modelos SLT, y sobre GSL, un conjunto de datos de Lengua de Señas Griega hecho en un laboratorio.

Palabras claves: Bases de Datos de Lenguaje de Señas, Estimación de Poses, Lenguaje de Señas, Libre de Glosas, Traducción de Lenguaje de Señas.

1 Introduction

Artificial Intelligence models based on Deep Neural Networks are the key technology behind many new applications such as autonomous driving, automatic monitoring of traffic cameras, text translation, speech recognition, and others [1]. Particularly, the fields of Computer Vision (CV) and Natural Language Processing (NLP) have been, thus far, the most successful in leveraging these models to enhance their performance and enable the development of new systems [2].

A significant application field of these techniques that combines CV and NLP is Sign Language Recognition (SLR). SLR seeks to develop systems capable of understanding the individual signs performed in a video. Sign Language Translation (SLT) goes a

step further and also requires the ability to translate a message in sign language into a written language, to facilitate communication between deaf communities and speakers [3, 4]. Due to its nature, SLT usually employs hybrid models, with CV to capture visual patterns and convert them into an internal representation, and NLP to generate the translation based on that representation [3, 5, 6]. Sign languages have their own grammar which in most cases greatly differs from their written counterpart. This makes SLT a significantly more complex problem than SLR.

While there have been advancements in this area recently, primarily driven by the development of deep neural models, we are still far from building accurate and robust applications [3, 4]. Although the models have made significant progress, the most substantial bottleneck is the lack of training data, a deficiency that varies for each sign language depending on the region [3, 4]. Actually, for most sign languages across the world, the amount of labelled data is very low and hence they can be considered low-resource languages [7].

Datasets for SLT are typically composed of videos and their corresponding translations into a written language. Relying on extra elements, like wearable bracelets, gloves or 3D cameras, can limit even more the amount of available resources. Also, systems that use smart gloves, wristbands or other wearables are considered intrusive and not accepted by sign language communities [8]. Also, nowadays, pose detection models that can extract pose and depth information from an RGB video are available and have been used as feature extractors for SLT models. The usage of pose features instead of the full video comes with several advantages such as a significant reduction of dimensionality of the input data and the removal of noise such as the background, lightning and clothing of the signer. This makes it the most viable approach for running SLT models on low power devices such as mobile devices.

Another feature sometimes included in SLT datasets is an intermediate representation called *glosses*. A sign language gloss is a written representation of a sign in one or more words of a spoken language, commonly the majority language of the region [9]. Translating from sign language (SL) videos into glosses results in an easier task than full SLT as there is a one-to-one relation between signs and glosses and both follow the same order. As such, gloss-based methods have significantly improved the SLT performance compared to end-to-end gloss-free approaches [10]. However, glosses do not accurately represent the meaning of signs in all cases and glossing has several limitations and problems [11]: (i) they are inherently sequential, whereas signs often exhibit simultaneity [12]; (ii) as glosses are based on spoken languages, there may be an implicit infuence of the spoken language projected onto the sign language [13, 11]; (iii) there is no universal standard on how glosses should be constructed: this leads to diferences between corpora of diferent sign languages, or even between several sign language annotators working on the same corpus [14]. Finally, annotating glosses is a labor intensive task, which requires fine-grained alignment and labeled by specialists, significantly constraining the scalability of gloss-based SLT methods.

In this work, we explore how Transformer based models perform in gloss-free SLT using only pose information as input. The transformer architecture is nowadays the state of the art for most NLP tasks, so this is intended to work as a baseline for future experiments featuring more complex model alongside pretraining techniques.

2 Related Work

2.1 Datasets

Currently, the most relevant SLT dataset is RWTH-Phoenix-Weather 2014 T [15] (RWTH for shortness). It contains videos of German Sign Language (GSL) extracted from German public TV weather forecasts. This dataset is used today as the main benchmark for SLT and, having a vocabulary of over 1000 signs, it was considered until recently, the only resource for large-scale continuous sign language worldwide [4]. RWTH has been recorded under real-life conditions, which may result in a more challenging dataset than those that were laboratory-made. Besides presenting more diverse scenarios, lightning conditions and signers, real-life generated datasets typically present a significant amount of sentences and tokens that appear really few times or only once (known as singletons) across the whole dataset.

This can be clearly seen in Table 1, that shows a comparison between RWTH and GSL, a laboratory-made dataset composed of common phrases in Greek Sign Language, repeated many times.

Table 1: Comparison between RWTH and GSL.

Dataset	RWTH	GSL
Language	German	Greek
Sign language	GSL	Greek SL
Real life	Yes	No
Signers	9	7
Duration [h]	10.71	9.51
# Samples	7096	10,295
# Unique sentences	5672	331
% Unique sentences	79.93%	3.21%
Vocab. size (w)	2887	473
# Singletons (w)	1077	0
% Singletons (w)	37.3%	0%

2.2 Pose-based gloss-free SLT

Even though SLR and SLT are not novel fields of study, gloss-free SLT is rather recent, as the first works following this approach date from not more than two years ago. Gloss-based approaches for SLT still achieve the best results: [16] represents SoTA for gloss-based SLT in RWTH with a BLEU score of 28.95, while [17], the SoTA for gloss-free SLT scores 23.09. However, gloss-free models obtain competitive results without having to deal with all the limitations mentioned before.

As for gloss-free SLT, most works combine the usage of a visual encoder with a pretrained Large Language Model (LLM) model. In [17], the authors propose a method called Factorized Learning assisted with Large Language Model, where they first train only the visual encoder with a simple transformer network for decoding and then use the output of the visual encoder to train an LLM (MBart [18]), already pretrained on multilingual corpora. In [10], a similar standard Transformer model pretrained on specific tasks designed to reduce the semantic gap between visual and textual representations and it achieves a BLEU of 21.74. In [19], the authors performed an analysis of existing models to confirm how gloss annotations make SLT easier and confirmed that it can help the model implicitly learn the location of semantic boundaries in continuous sign language videos. To achieve this in a gloss-free SLT Transformer model, they modified the attention mechanism to ensure similar values between subsequent frames of the video. Following this approach they achieved a BLEU score of 15.74.

Models that use only positional information for SLR have been successfully developed achieving competitive results against video-based models. An example of these can be found at [7]. However, to the best of our knowledge, the only work that approached gloss-free SLT using only pose information is [20], where the authors train an encoder-decoder GRU model only on positional information trying different normalization and data augmentation methods. They primarily trained the model over the KETI database, obtaining a BLEU score of 84.39. Following the same approach over RWTH they obtained a BLEU score of 13.31. The difference in BLEU is explained by the difference in the complexity of the databases mainly due to the fact that KETI is laboratory-made and containing a simple and reduced set of sentences.

3 Experiments

For this work, a transformer model was developed following the standard architecture presented in [21] with some modifications to adapt it to an SLT task.

First, it uses a pose encoder module composed of 3 stacked convolutional layers that run a 1D convolution across the temporal dimension with a kernel size of 1. The goal of this encoder is to embed the pose into a meaningful vector of the pose. As for the decoder, it uses a standard embedding layer. Then, both the representations of the pose and the word embeddings are concatenated with their respective positional encodings before being used as input for the Transformer as shown in figure 1.

Two sets of experiments were carried on, one over RWTH and another over GSL. Smaller versions of the model proved to work better over GSL: best accuracy was obtained when using a hidden dimension of 16 for the Transformer model, the number of encoder layers was set to 1, the number of decoder layers to 4 and the dropout to 0.2. When training over RWTH, the best performing model had a hidden dimension of 64, the number of encoder layers was set to 2, the number of decoder layers to 6 and the dropout to 0.1.

The model was trained with RWTH poses generated by Mediapipe [22]. The poses are encoded through 543 pose keypoints: 33 pose landmarks, 468 face landmarks, and 21 hand landmarks per hand. Pose information was accessed through the Sign Language Datasets library [23]. On RWTH, the model achieved a per word accuracy of 41% while on GSL it obtained 93%.

Once the model was trained, two methods were tested for generating the output text: greedy decoding and beam decoding, with a beam size of 32. In both experiments the beam decoding slightly surpassed the greedy decoding obtaining a BLEU-4 score of over RWTH and of 44 over GSL. Complete results can be observed at table 2.

Finally, it's important to highlight that the resulting base model consists of 3.9 million parameters. A small size compared to video-based SoTA models like [17], which consists of 25.61 million.

4 Conclusions and Future Work

In this article, we presented an initial study on gloss-free, transformed-based SLT models. We experimented with RWTH-Phoenix the most well-known SLT dataset, and identified key issues blocking effective use of this type of models for SLT and with GSL, a simpler laboratory-made dataset.

Although the presented model under performs other SoTA models in the same tasks, we intend the model to be used as baseline for future experiments as gloss free SLT is still an emerging field and there are almost no works that only rely on pose information.

Transformers are renown for requiring larger amounts of data than other models in various domains. In this work, we have established this issue, and highlighted its importance as SLT is a low resource field, with reduced availability and quality of datasets.

Currently, our SLT model's performance is limited by model size and computational requirements. In the future, we plan to train larger versions of the model alongside more complex data augmentation methods to prevent overfitting. Additionally, we intend to train

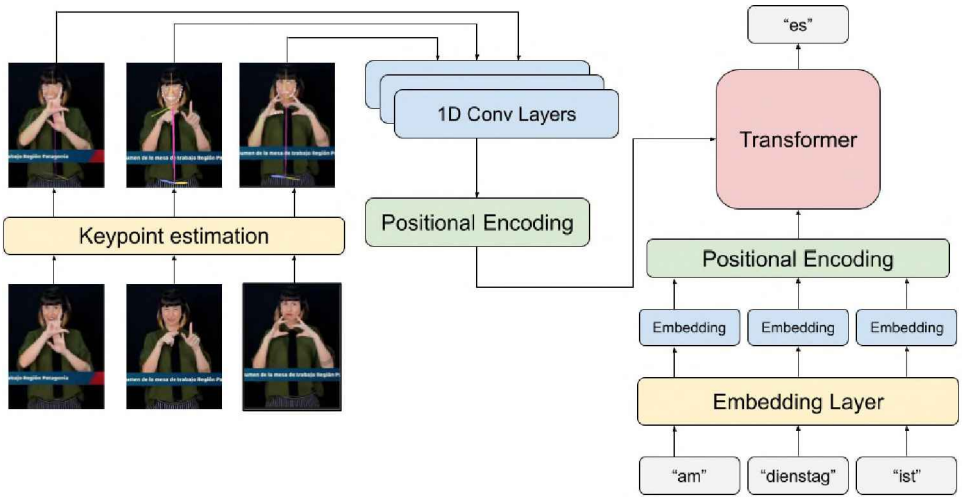

Figure 1: Scheme of the described model.

Table 2: Experiments results.

Dataset	Method	Accuracy	BLEU-4	BLEU-3	BLEU-2	BLEU-1
RWTH	Greedy	41.7%	5.7	6.85	9.98	15.87
RWTH	Beam	41.7%	5.9	6.85	10	18.87
GSL	Greedy	93.4%	43.06	54.55	63.45	75.46
GSL	Beam	93.4%	43.74	55.15	63.2	75.78

and evaluate the model on other datasets in order to have a more general baseline.

Finally, we will perform experiments pre-training the encoder on multiple sign language databases, an interesting and under-explored line of research. In this fashion, the encoder can effectively learn to extract relevant more general representations for poses, to later match it with language specific decoders.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

The authors confirm contribution to the paper as follows: PD: Conceptualization, Data curation, Software, Investigation, Writing – original draft preparation; GR: Data curation; Software WH: Supervision, Validation; OS: Data curation; Software; FQ: Supervision, Writing-Reviewing and Editing; FR: Supervision, Writing-Reviewing and Editing. All authors reviewed the results and approved the final version of the manuscript.

Funding

This work has been possible thanks to the support of the program Stic-AmSud framed in the project Stic-AmSud 23-STIC-06.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] Y. Bengio, Y. Lecun, and G. Hinton, "Deep learning for ai," *Communications of the ACM*, vol. 64, no. 7, pp. 58–65, 2021.

[3] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, ..., and M. Ringel Morris, "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, October 2019, pp. 16–31.

[4] O. Koller, "Quantitative survey of the state of the art in sign language recognition," *arXiv preprint arXiv:2008.09918*, 2020.

[5] I. Papastratis, C. Chatzikonstantinou, D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Artificial intelligence technologies for sign language," *Sensors*, vol. 21, no. 17, p. 5843, 2021.

[6] J. Zheng, Y. Wang, C. Tan, S. Li, G. Wang, J. Xia, ..., and S. Z. Li, "Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 141–23 150.

[7] P. Selvaraj, G. Nc, P. Kumar, and M. Khapra, "Openhands: Making sign language recognition accessible with pose-based pretrained models across languages," *arXiv preprint arXiv:2110.05877*, 2021.

[8] M. Erard. (2017) Why sign language gloves don't help deaf people. [Online]. Available: https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/

[9] M. De Coster, D. Shterionov, M. Van Herreweghe, and J. Dambre, "Machine translation from signed to spoken languages: State of the art and challenges," *Universal Access in the Information Society*, pp. 1–27, 2023.

[10] B. Zhou, Z. Chen, A. Clapés, J. Wan, Y. Liang, S. Escalera, Z. Lei, and D. Zhang, "Gloss-free sign language translation: Improving from visual-language pretraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 871–20 881.

[11] N. Frishberg, N. Hoiting, and D. I. Slobin, "Transcription," in *Sign Language*. Berlin: De Gruyter Mouton, 2012, pp. 1045–1075. [Online]. Available: https://doi.org/10.1515/9783110261325.1045

[12] M. Vermeerbergen, L. Leeson, and O. A. Crasborn, *Simultaneity in Signed Languages: Form and Function*. Amsterdam: John Benjamins Publishing, 2007, vol. 281. [Online]. Available: https://doi.org/10.1075/cilt.281

[13] M. Vermeerbergen, "Past and current trends in sign language research," *Language & Communication*, vol. 26, no. 2, pp. 168–192, 2006. [Online]. Available: https://doi.org/10.1016/j.langcom.2005.10.004

[14] M. De Sisto, V. Vandeghinste, S. E. Gómez, M. De Coster, D. Shterionov, and H. Seggion, "Challenges with sign language datasets for sign language recognition and translation," in *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*. Marseille, France: European Language Resources Association (ELRA), 2022, pp. 2478–2487.

[15] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7784–7793.

[16] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 043–17 056, 2022.

[17] Z. Chen, B. Zhou, J. Li, J. Wan, Z. Lei, N. Jiang, Q. Lu, and G. Zhao, "Factorized learning assisted with large language model for gloss-free sign language translation," *arXiv preprint arXiv:2403.12556*, 2024.

[18] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[19] A. Yin, T. Zhong, L. Tang, W. Jin, T. Jin, and Z. Zhao, "Gloss attention for gloss-free sign language translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2551–2562.

[20] Y. Kim, M. Kwak, D. Lee, Y. Kim, and H. Baek, "Keypoint based sign language translation without glosses," *arXiv preprint arXiv:2204.10511*, 2022.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, ..., and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[23] A. Moryossef and M. Müller, "Sign language datasets," https://github.com/sign-language-processing/datasets, 2021.

Citation: P. Dal Bianco, G. Ríos, W. Hasperué, O. Stanchi, F. Quiroga and F. Ronchetti *A study on pose-based deep learning models for gloss-free Sign Language Translation*. Journal of Computer Science & Technology, vol. 24, no. 2, pp. 99-103, 2024.
DOI: 10.24215/16666038.24.e09
Received: April 15, 2024 **Accepted:** September 2, 2024.