Gloss-free Argentinian Sign Language Translation with pose-based deep learning models

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

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

Abstract. The main challenge of automatic Sign Language Translation (SLT) is obtaining data to train models. For Argentinian Sign Language (LSA), the only dataset available for SLT is LSA-T, which contains extracts of a news channel in LSA and the corresponding Spanish subtitles provided by the authors. LSA-T contains a wide variety of signers, scenarios, and lightnings that could bias a model trained on it. We propose a model for Argentinian gloss-free SLT, since LSA-T does not contain gloss representations of the signs. The model is also pose-based to improve performance on low resource devices. Different versions of the model are also tested in two other well-known datasets to compare the results: GSL and RWTH Phoenix Weather 2014T. Our model stablished the new SoTA over LSA-T, which proved to be the most challenging due to the variety of topics covered that result in a vast vocabulary with many words appearing few times.

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

1 Introduction

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 [1, 11]. 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 [1, 15, 21]. 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 [1, 11]. 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 [1, 11]. 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 [16].

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 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 [6]. 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 [22]. However, glosses do not accurately represent the meaning of signs in all cases and glossing has several limitations and problems [9]: (i) they are inherently sequential, whereas signs often exhibit simultaneity [19]; (ii) as glosses are based on spoken languages, there may be an implicit infuence of the spoken language projected onto the sign language [18, 9]; (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 [7]. 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 performs 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, LSA-T [5] is the only dataset available for Argentinian SLT. LSA-T was built from videos from the YouTube channel CN Sordos, a news channel created in 2020 by deaf people and deaf people's relatives. The hosts use LSA to communicate the news with Spanish subtitles provided by the authors. There is gender parity among the signers, and videos contain different locations, backgrounds, and lighting conditions. It features a wide variety of topics, which results in a significant amount of sentences and tokens that appear a very few times or only once (known as singletons) across the whole dataset. This results in the dataset being more challenging for SLT compared to laboratory-made datasets.

As a benchmark, the most relevant SLT dataset today is RWTH-Phoenix-Weather 2014 T [2] (RWTH for shortness). It contains videos of German Sign Language (GSL) extracted from German public TV weather forecasts. This dataset is used as the main benchmark for SLT and, having a vocabulary of more than 1000 signs, it was until recently considered the only resource for large-scale continuous sign language worldwide [11]. The aforementioned problem of low-frequency words persists, but is not as present as in LSA-T. This can be clearly seen in Table 1, that shows a comparison between RWTH, LSA-T and also GSL, a laboratory-made dataset composed of common phrases in Greek Sign Language, repeated many times.

Table 1. Statistics of the three datasets used in this paper: LSA-T, RWTH and GSL

Dataset	RWTH	LSA-T	GSL
Language	German	Spanish	Greek
Sign language	GSL	LSA	GSL
Real life	**Yes**	**Yes**	No
Signers	9	**103**	7
Duration [h]	10.71	21.78	9.51
# Samples	7096	14,880	10,295
# Unique sentences	5672	**14,254**	331
% Unique sentences	79.93%	**95.79%**	3.21%
Vocab. size (w)	2887	**14,239**	310
# Singletons (w)	1077	**7150**	0
% Singletons (w)	37.3%	**50.21%**	0%
Resolution	210x260	**1920x1080**	848x480
FPS	25	**30**	**30**

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 first appeared in 2022. Gloss-based approaches for SLT still achieve the best results: [3] represents SoTA for gloss-based SLT in RWTH with a BLEU score of 28.95, while [4], 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 [4], 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 [12]), already pretrained on multilingual corpora. In [22], 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 [20], 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 [16]. However, to the best of our knowledge, the only work that approached gloss-free SLT using only pose information is [10], 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 since KETI is laboratory-made and contains a simple and reduced set of sentences.

3 Experiments

For this work, a transformer model was developed following the standard architecture presented in [17] 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.

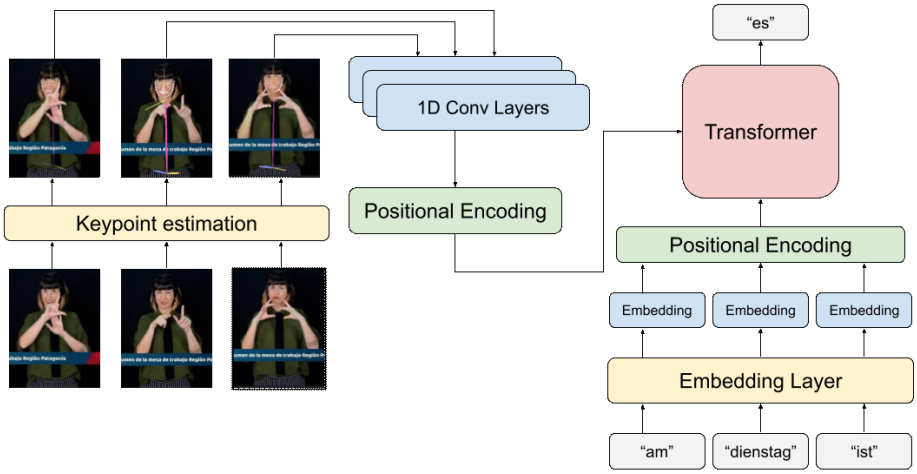

Fig. 1. Scheme of the described model.

The model was trained and evaluated on LSA-T, and smaller versions were trained on RWTH and GSL. As expected, LSA-T was more complex to train and required larger versions of the model, as it can be seen in table 2.

Table 2. Hyperparameters of the best performing models for each dataset.

Dataset	LSA-T	RWTH	GSL
Hidden dimension size	256	64	16
# Encoder layers	2	2	1
# Decoder layers	6	4	2
# Dropout	0.2	0.2	0.1

The model was trained with poses generated by Mediapipe [13]. 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 [14]. 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. The complexity of the dataset in terms of the vocabulary size and the number of singletons was reflected in the training results, as can be seen in the table 3.

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 [4], which consists of 25.61 million.

Table 3. Comparison of Greedy and Beam Search Methods.

Dataset	Method	Accuracy	BLEU-4	BLEU-3	BLEU-2	BLEU-1
LSA-T	Greedy	16.7%	0.2	0.3	5	6.4
LSA-T	Beam	16.7%	0.05	0.1	5	6.7
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

4 Conclusions and Future Work

In this article, we presented an gloss-free, pose-based SLT model for translating Argentinian Sign Language. The results were significantly lower than those obtained on RWTH-Phoenix and GSL datasets, but this difference is explained by the characteristics of the dataset.

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

Transformers are renowed for requiring larger amounts of data than other models in various domains. In this work, we have confirmed this issue for the SLT domain, 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 and evaluate the model on other datasets in order to have a more general baseline.

Finally, we will perform experiments pretraining 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.

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. Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., ..., Ringel Morris, M.: Sign language recognition, generation, and translation: An interdisciplinary perspective. In: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility. pp. 16–31 (October 2019)
2. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7784–7793 (2018)

3. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. Advances in Neural Information Processing Systems **35**, 17043–17056 (2022)
4. Chen, Z., Zhou, B., Li, J., Wan, J., Lei, Z., Jiang, N., Lu, Q., Zhao, G.: Factorized learning assisted with large language model for gloss-free sign language translation. arXiv preprint arXiv:2403.12556 (2024)
5. Dal Bianco, P., Ríos, G., Ronchetti, F., Quiroga, F., Stanchi, O., Hasperué, W., Rosete, A.: LSA-T: The first continuous argentinian sign language dataset for sign language translation. In: Ibero-American Conference on Artificial Intelligence. pp. 293–304. Springer International Publishing (November 2022)
6. De Coster, M., Shterionov, D., Van Herreweghe, M., Dambre, J.: Machine translation from signed to spoken languages: State of the art and challenges. Universal Access in the Information Society pp. 1–27 (2023)
7. De Sisto, M., Vandeghinste, V., Gómez, S.E., De Coster, M., Shterionov, D., Seggion, H.: 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). pp. 2478–2487. European Language Resources Association (ELRA), Marseille, France (2022)
8. Erard, M.: Why sign language gloves don't help deaf people (2017), https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/
9. Frishberg, N., Hoiting, N., Slobin, D.I.: Transcription. In: Sign Language, pp. 1045–1075. De Gruyter Mouton, Berlin (2012). https://doi.org/10.1515/9783110261325.1045, https://doi.org/10.1515/9783110261325.1045
10. Kim, Y., Kwak, M., Lee, D., Kim, Y., Baek, H.: Keypoint based sign language translation without glosses. arXiv preprint arXiv:2204.10511 (2022)
11. Koller, O.: Quantitative survey of the state of the art in sign language recognition. arXiv preprint arXiv:2008.09918 (2020)
12. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
13. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ..., Grundmann, M.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
14. Moryossef, A., Müller, M.: Sign language datasets. https://github.com/sign-language-processing/datasets (2021)
15. Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., Daras, P.: Artificial intelligence technologies for sign language. Sensors **21**(17), 5843 (2021)
16. Selvaraj, P., Nc, G., Kumar, P., Khapra, M.: Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. arXiv preprint arXiv:2110.05877 (2021)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
18. Vermeerbergen, M.: Past and current trends in sign language research. Language & Communication **26**(2), 168–192 (2006). https://doi.org/10.1016/j.langcom.2005.10.004, https://doi.org/10.1016/j.langcom.2005.10.004

19. Vermeerbergen, M., Leeson, L., Crasborn, O.A.: Simultaneity in Signed Languages: Form and Function, vol. 281. John Benjamins Publishing, Amsterdam (2007). https://doi.org/10.1075/cilt.281, https://doi.org/10.1075/cilt.281
20. Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., Zhao, Z.: Gloss attention for gloss-free sign language translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2551–2562 (2023)
21. Zheng, J., Wang, Y., Tan, C., Li, S., Wang, G., Xia, J., ..., Li, S.Z.: 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. pp. 23141–23150 (2023)
22. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20871–20881 (2023)