Transformers in Sign Language Translation: An In-Depth Overview

Varsha Ghodase[1], Divya Chaudhari[2], Shubhada Aher[3], Ratan Ahire[4], Komal Pathare[5]

Professor Rahul Kumar[6]

[1,2,3,4,5]B.E Student, Department of Computer Engineering, GS Moze College of Engineering, Pune (India)
[6]Assistant Professor, Computer Engineering Department, GS Moze College of Engineering, Pune (India)

ABSTRACT:

In recent years, Sign Language Translation has gained significant attention which aims to reduce the communication barrier between deaf and hearing communities. In this paper, we briefly outline Sign Language recognition and Sign Language Translation and provide a detailed analysis of the Transformer model. In this survey paper, we examine the core issues in Sign Language Translation and put forth potential directions for its development.

Keywords: Sign Language translation, Transformers

1 Introduction

Sign language is the first language for those who were born deaf or lost their hearing in early childhood [3]. Unlike spoken language, sign language relies on gestures rather than words. Sign languages have their own grammatical structures and vocabulary, and each country often has its own sign language. Some common examples include American Sign Language (ASL), British Sign Language (BSL), and Auslan in Australia. They have their own lexicons and grammatical constructs, thus converting between sign and spoken language is a translation problem [5].

Sign Language Recognition (SLR) is a technology that focuses on understanding and translating sign language gestures into written or spoken language. It enables computers to interpret the complex hand movements, facial expressions, and body postures used in sign language. It is used in various applications, from real-time sign language interpretation to sign language-driven computer interfaces. Despite the numerous advances in SLR [15] and even the move to the challenging Continuous SLR (CSLR) [33, 36] problem, do not allow us to provide meaningful interpretations.[10]

Sign Language Translation (SLT), on the other hand, goes beyond mere recognition by not only understanding sign language but also translating it into written or spoken language. Sign Language Translation (SLT) takes the concept of SLR a step further. SLT not only recognizes sign language gestures but also translates them into written or spoken language. Several SLT systems have been proposed in the past, including rule-based systems (Zhao et al.,2000) and statistical methods (Bungeroth and Ney, 2004).[8] There has been a shift in SLT from rule based systems to more advanced methods, including statistical techniques and deep learning approaches. These improvements primarily relate to feature extraction, while the text emphasizes an alternative focus on enhancing the translation model by creating a more robust encoder-decoder architecture.

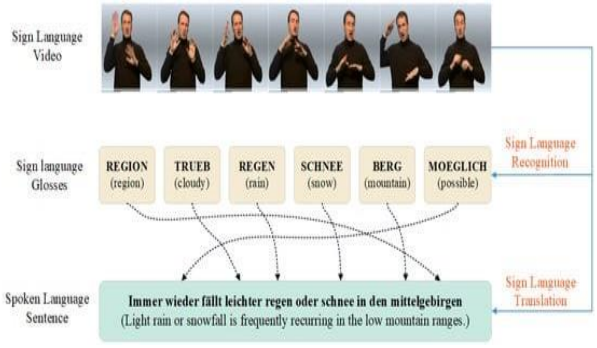

Sign Language Video
Sign language Glosses
REGION (region)
TRUEB (cloudy)
REGEN (rain)
SCHNEE (snow)
BERG (mountain)
MOEGLICH (possible)
Sign Language Recognition
Spoken Language Sentence
Immer wieder fällt leichter regen oder schnee in den mittelgebirgen
(Light rain or snowfall is frequently recurring in the low mountain ranges.)
Sign Language Translation

Fig. 1: The difference between SLR and SLT

2 Sign Language Translation (STL)

Earlier approaches focused on translating from sign language glosses to written language text, it is crucial to acknowledge that glosses may not comprehensively convey the meaning of signed utterances and can be influenced by written language. Gloss-free SLT refers to the absence of gloss supervision throughout the training and testing stages, including pretraining and fine-tuning [2]. Consequently, there has been a shift towards end-to-end SLT, which involves direct translation from sign language videos to written language text. Researchers initially approached STL as a Neural Machine Translation (NMT) challenge and developed the first end-toend SLT model by combining convolutional neural networks (CNNs) with attention-based encoder-decoder networks using recurrent neural networks (RNNs). A significant advancement in NMT was the introduction of the Transformer network, a sequence transduction model based on attention mechanisms. Transformer architecture has gained popularity in the field of SLT in recent years. Some studies adopt a two-stage approach, first recognizing glosses from sign videos (Sign2Gloss) and then mapping glosses into spoken language sentences (Gloss2Text). In contrast, other studies pursue an end-to-end approach, predicting spoken language sentences directly from sign video inputs.

3 Transformers

Transformers are a type of deep neural network architecture that consists of an encoder and decoder pair. They serve as the core technology for improving the accuracy and efficiency of SLT systems. Transformers work by handling input data all at once and using the self-attention mechanism to figure out how words or signs in a sequence relate to one another. This parallelization enhances computational efficiency, making transformers highly scalable. The architecture consists of an encoder-decoder structure, each comprising multiple layers of self-attention mechanisms and feedforward networks.

4 Improving Sign Language Translation Using Transformers

Transformer models, notably GPT and CLIP, witnessed a notable increase in size and scale. These larger models showcased enhanced performance across various natural language processing (NLP) tasks, particularly capturing attention for their prowess in handling both textual and visual data. The rise of multimodal transformers, exemplified by models like CLIP, sparked a growing interest in their applicability to diverse tasks involving different data types.

As these transformer models expanded, addressing computational and memory demands became imperative. Techniques like model pruning, quantization, and distillation were deployed to streamline these models. Despite these challenges, pretrained transformer models like BERT and GPT retained their pivotal role in numerous NLP applications. Finetuning these models on specific tasks emerged as a successful strategy, consistently delivering state-of-the-art performance across diverse domains

In the context of Sign Language Translation (SLT), the transformer architecture underwent evolution from a baseline model with three layers in both encoder and decoder components. The SLT domain confronted significant challenges, primarily stemming from the scarcity of labeled data. To navigate this hurdle, researchers turned to transfer learning techniques, incorporating pretrained language models such as BERT and mBART-50 into their translation frameworks. To combat overfitting, the ”frozen pretrained transformer” technique came into play, strategically freezing parameters during training and leveraging pretrained models like BERT as a foundational starting point for the SLT model. The application of Frozen Pretrained Transformers (FTPs), specifically utilizing BERT-base pretrained on an English text corpus, demonstrated the effective transfer of selfattention patterns to the SLT task, spanning new modalities, tasks, and languages. The notable improvements in SLT model performance, as indicated by increased BLEU4, ROUGE-L, and CHRF scores, emphasized the efficacy of this approach. In scenarios with gloss-level annotations, the improvement averaged around 1 BLEU-4 compared to a baseline trained from scratch. In situations without gloss annotations, this gain widened, reaching an average increase of 1.95 BLEU-4. The study not only sheds light on the challenges inherent in SLT but also underscores the paramount importance of prioritizing data quality over quantity. It suggests promising avenues for future research, advocating for meticulous data cleaning and the exploration of advanced feature extraction techniques from the Sign Language Recognition domain.

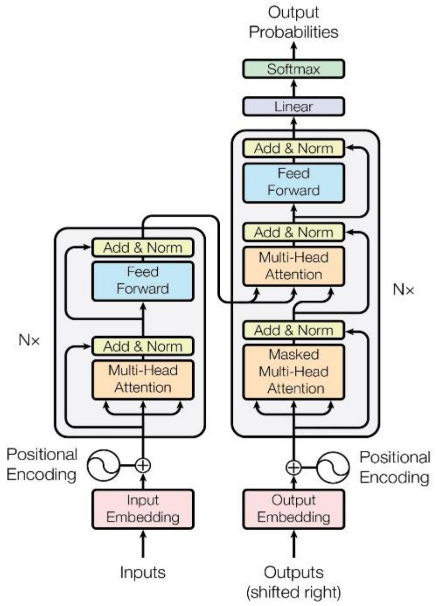


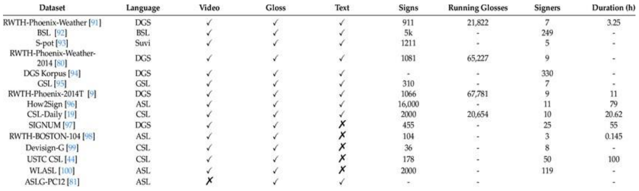
Fig. 2: Transformer Architecture

5 Datasets

Sign Language Translation (SLT) is still in its early research stages, with a limited number of studies, mainly due to the scarcity of datasets, which is a significant hindrance to its development. The PHOENIX corpus, especially PHOENIX14, stands out as a foundational dataset, laying the groundwork for Sign Language Recognition (SLR) and Translation (SLT) investigations. However, the shortage of high-quality, diverse sign language data is evident in the RWTH-PHOENIXWeather-2014T dataset, comprising fewer than 9000 samples. The RWTH-PHOENIX-Weather 2014T corpus is a significant stride, presenting an extensive vocabulary, a continuous SLT dataset with diverse sign language videos, gloss annotations, and German translations. As an extension of PHOENIX14, this dataset aims to overcome previous limitations and foster unconstrained sign language research. Ongoing efforts to refine segmentation boundaries and enhance normalization schemes emphasize the commitment to advancing SLT research.

To address the lack of parallel corpus pairs, Coster et al. introduced a frozen pretrained transformer (FPT) model, and Zhao et al. focused on learning linguistic characteristics of spoken language to enhance translation performance. Fu et al. proposed a contrastive learning model, and Mocialov et al. incorporated transfer learning from a large corpus like Penn Treebank. Chen et al. devised a progressive pretraining approach using external data, and Cao et al. introduced a task-aware instruction network (TIN) for improved translation. Addressing the gloss-text task, Moryossef et al. proposed rule-based augmentation strategies.

This collective effort showcases a dynamic response to the challenges posed by the limited availability of sign language datasets, with researchers actively working to refine segmentation boundaries and improve normalization schemes. However, the field remains a work in progress, and continued collaborative efforts are essential for the further advancement of Sign Language Translation research.



Dataset	Language	Video	Gloss	Text	Signs	Running Glosses	Signers	Duration (h)
RWTH-Phoenix-Weather [91]	DGS	✓	✓	✓	911	21,822	7	3.25
BSL [92]	BSL	✓	✓	✓	5k	-	249	-
S-pot [93]	Suvi	✓	✓	✓	1211	-	5	-
RWTH-Phoenix-Weather-2014 [80]	DGS	✓	✓	✓	1081	65,227	9	-
DGS Korpus [94]	DGS	✓	✓	✓	-	-	330	-
GSL [95]	GSL	✓	✓	✓	310	-	7	-
RWTH-Phoenix-2014T [9]	DGS	✓	✓	✓	1066	67,781	9	11
How2Sign [96]	ASL	✓	✓	✓	16,000	-	11	79
CSL-Daily [19]	CSL	✓	✓	✓	2000	20,654	10	20.62
SIGNUM [97]	DGS	✓	✓	✗	455	-	25	55
RWTH-BOSTON-104 [98]	ASL	✓	✓	✗	104	-	3	0.145
Devisign-G [99]	CSL	✓	✓	✗	36	-	8	-
USTC CSL [44]	CSL	✓	✓	✗	178	-	50	100
WLASL [100]	ASL	✓	✓	✗	2000	-	119	-
ASLG-PC12 [81]	ASL	✗	✓	✓	-	-	-	-

Fig. 3: Datasets in SLR and SLT

The only used resource is a corpus of 70M tweets randomly extracted from Twitter, a collection of about 800M tweets, for building the word embeddings.

6 Quantifying Translation Quality: Metrics

The evaluation of Sign Language Translation (SLT) systems involves the application of various metrics to assess their performance and efficacy. Past research in this field has extensively explored different matrices to measure the quality and accuracy of SLT models. Widely used metrics in SLT research include BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and CHRF (Character n-gram F-score). BLEU evaluates translation precision by comparing n-grams in reference and candidate translations, offering a numerical measure of similarity. ROUGE assesses content overlap, prioritizing recall and proving useful for evaluating the informativeness of translations. CHRF, on the other hand, considers character level n-grams, offering a detailed analysis of translation quality.

The BLEU metric is a fundamental tool in Sign Language Translation (SLT), crucial for evaluating how accurately translations represent sign languages. It focuses on the precision of n-grams, essential for capturing nuanced meanings in diverse sign languages. Being a well-established benchmark in machine translation, BLEU ensures consistency and comparability when assessing SLT systems.

7 Conclusion

In conclusion, this survey navigates through the progress in Sign Language Translation (SLT), spotlighting the influence of transformer models such as GPT and CLIP. The paper emphasizes the importance of these models in addressing the distinctive challenges posed by sign languages, especially in scenarios involving multiple modes of communication. Beyond chronicling current accomplishments, this survey establishes a platform for forthcoming investigations, encouraging researchers to delve deeper into feature extraction methodologies and collaborative endeavors in SLT research. In essence, the survey provides nuanced insights into the current state of the field while instilling optimism for future strides in Sign Language Translation.

References

[1] Liang, Z., Li, H., and Chai, J., 2023. “Sign language translation: A survey of approaches and techniques”. *Electronics,* 12(12), p. 2678.

[2] Zhou, B., Chen, Z., Clapes, A., Wan, J., Liang,´ Y., Escalera, S., Lei, Z., and Zhang, D., 2023. “Gloss-free sign language translation: Improving from visual-language pretraining”. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20871–20881.

[3] Miyazaki, T., Morita, Y., and Sano, M., 2020. “Machine translation from spoken language to sign language using pre-trained language model as encoder”. In Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: sign language resources in the service of the language community, technological challenges and application perspectives, pp. 139–144.

[4] De Coster, M., and Dambre, J., 2022. “Leveraging frozen pretrained written language models for neural sign language translation”. *Information,* 13(5), p. 220.

[5] Sincan, O. M., Camgoz, N. C., and Bowden, R., 2023. “Is context all you need? scaling neural sign language translation to large domains of discourse”. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1955–1965.

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

[7] Cabot Alvarez, P., 2022.´ “Sign language translation with pseudo-glosses”. B.S. thesis, Universitat Politecnica de Catalunya.`

[8] De Coster, M., D’Oosterlinck, K., Pizurica, M., Rabaey, P., Verlinden, S., Van Herreweghe, M., and Dambre, J., 2021. “Frozen pretrained transformers for neural sign language translation”. In 18th Biennial Machine Translation Summit (MT Summit 2021), Association for Machine Translation in the Americas, pp. 88– 97.

[9] Yin, K., and Read, J., 2020. “Better sign language translation with stmc-transformer”. *arXiv preprint arXiv:2004.00588*.

[10] Walsh, H., Saunders, B., and Bowden, R., 2022. “Changing the representation: Examining language representation for neural sign language production”. *arXiv preprint arXiv:2210.06312*.