Overview of Sign Language Translation Based on Natural Language Processing

Hanmo Wang

Software Department, Henan University, 475004, Kaifeng, China

Abstract. This paper explores the progress, challenges, and future directions in Sign Language Translation (SLT) within the broader field of Sign Language Processing (SLP), which combines Computer Vision (CV) and Natural Language Processing (NLP) to translate sign language videos into spoken language texts. The study begins by examining various sign language representation methods, such as video, gesture, symbol systems, and annotation, analyzing their strengths and weaknesses. It highlights the critical need for high-quality, large-scale datasets to advance SLT research, while acknowledging challenges like data scarcity, annotation inconsistencies, and ethical concerns. The paper then reviews recent SLT research, identifying key challenges and proposing solutions, such as expanding datasets through collaboration with the deaf and hard-of-hearing community, and employing advanced data collection techniques. Additionally, it suggests applying NLP methods like transfer learning and large language models to address specific challenges. Finally, the paper advocates for stronger interdisciplinary collaboration between CV and NLP to develop models and algorithms that are better suited to the unique aspects of sign languages.

1 Introduction

According to the World Health Organization, hearing loss afflicts over 5% of the global populace, translating to approximately 430 million individuals. By 2050, projections indicate that the number will surge to over 700 million, comprising roughly one-tenth of the world's population, with disabling hearing impairment. Data from the China Disabled Person's Federation reveals that as of 2020, China is home to about 3.2 million with hearing impairments and 618,000 individuals who are deaf-mute. Sign language, a fundamental mode of communication for the deaf and mute, has long been established as the linchpin of their community. Reflecting an ancient adage, "The deaf employ their eyes as ears, and the mute use their hands as mouths," this method of interaction has been refined over centuries among those with hearing loss. In the contemporary world, the push for equality and respect for the disabled has emerged as a global movement, leading to wider societal embrace and acknowledgment of this distinct demographic. Consequently, sign language, a critical instrument for their interaction, merits the requisite recognition and integration. With

Corresponding author: 2225050185@henu.edu.cn

advancements in machine translation, NLP has revolutionized technological engagement. Nevertheless, the predominantly voice or text-based reliance of NLP models inadvertently excludes those who communicate via sign language. Therefore, including SLP in NLP research assumes profound importance [1].

SLP stands out as a discipline at the confluence of CV and NLP. Traditionally, the domain has exhibited a tilt towards CV, primarily due to the paucity of tools for the efficient manipulation of sign language videos, with a concentration on the comprehension and analysis of sign languages through image recognition technologies. However, with the advent of advanced technologies, NLP theories and tools have come to the fore in SLP research. The visionary approach advocated by researchers such as Kayo Yin and Amit Moryossef, integrating SLP with NLP, demonstrates remarkable foresight [1]. This amalgamation not only encourages NLP scholars to delve into SLP but also leverages NLP technologies like semantic understanding and machine translation to bolster the progression of SLP, thus facilitating seamless communication between sign language and natural language. It concurrently addresses impending challenges, including the development of proficient segmentation techniques, the aggregation of sign language data, the formulation of models based on linguistic theories, and the active involvement of the deaf community in research endeavors. These pursuits are instrumental in enhancing the deaf community's quality of life and promoting societal inclusiveness.

Pioneering scholars such as Razieh Rastgoo and Kourosh Kiani have delineated the dual components of SLP: the visual and linguistic modalities, with the former encompassing image/video data and the latter incorporating natural language text [2]. Hence, both CV and NLP technologies are indispensable in managing these input modalities, with deep learning methodologies enhancing model efficacy in both domains. Zuo Ronglai, Fangyun Wei, and associates have introduced the Natural Language-Assisted Sign Language Recognition Framework (NLA-SLR), harnessing the semantic information in annotations to bolster sign language recognition, and demonstrating state-of-the-art performance on Sign Language Recognition (SLR) datasets [3]. This underscores the vitality of interdisciplinary collaboration in SLP. CV provides the technical framework for image capture and motion analysis, while NLP offers the theoretical substratum for semantic comprehension and linguistic generation. This synergy demands the innovation of algorithms that cater to the unique spatiotemporal attributes of sign languages, as well as a focus on linguistic facets such as the syntax and composition of gestures. Viewing SLP as a pivotal branch of NLP not only propels technological progress but also embodies the ethos of social inclusivity and diversity.

The collaborative efforts between disciplines, coupled with community participation, are poised to promote further advancements in SLP, significantly aiding and improving the lives of the deaf. This paper aims to encapsulate recent breakthroughs in SLT within the field. By examining and synthesizing the research contributions of various scholars, the paper aspires to delineate promising trajectories for future inquiry, providing motivation and direction for continued exploration in the domain.

2 Representations and data

2.1 Representations

Prior to conducting a synopsis, this paper delineates the contemporary manifestations of sign language, as exemplified in Fig. 1. These depictions function as conduits for linguistic data, highlighting the necessity to identify suitable forms of articulation.

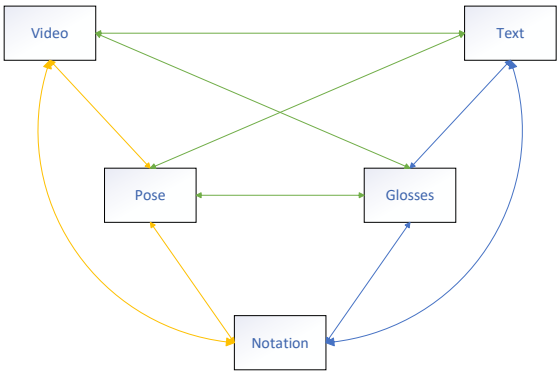

Fig. 1. The yellow and blue arrows represent the tasks to be undertaken by CV and NLP, respectively, whereas the green arrow indicates the collaborative efforts required from both CV and NLP. （Photo/Picture credit : Original）

Video: The video medium represents an intuitive and immediate approach, capable of visually encapsulating the complete array of information expressed by a sign language user. Its strength lies in the thorough retention of such information. Nonetheless, a significant shortcoming is the inclusion of an abundance of extraneous data not pertinent to sign language, which leads to elevated expenses in storage, processing, and transmission.

Pose: Derived from the abstraction of video, the pose method distills the essential elements of sign language into a streamlined skeletal animation format. Established techniques are available that facilitate the conversion of video to poses [4], effectively retaining semantic content while discarding superfluous material. However, pose representation maintains a high-dimensional continuous nature, rendering it less amenable to immediate processing as a natural language.

Notation: The notation system further refines features from poses, encapsulating sign language in symbolic form. Within the realms of linguistics and semiotics, a notation system encompasses graphical elements, symbols, characters, and abbreviations invested with particular meanings. A multitude of systems, such as HamNoSys [5], Stokoe notation [6], and Sutton SignWriting [7], have emerged as indispensable educational resources for the deaf and hard of hearing. Mastery of a script that directly corresponds to their native language is more intuitive and accessible than learning a spoken language, effectively bridging the divide between sign and written languages, and enhancing literacy and communicative proficiency. Moreover, the possession of a written form of their language instills within the deaf community a sense of cultural dignity and identity, affirming the legitimacy and intricacy of their communication mode. For NLP research, notation systems, being more discrete and lower-dimensional, akin to text, are better suited for existing NLP methodologies, offering a more holistic retention of sign language information with minimal loss of semantic integrity.

Gloss: Glossing involves the transcription of individual sign symbols, assigning a distinct semantic tag to each. Current glossing methodologies, however, have not fully encapsulated the breadth of a user's expression, including aspects such as body posture, eye contact, and spatial relationships [8], necessitating further development for precise information and semantic representation.

This paper concentrates primarily on SLT, encompassing the translation between notation systems and text, as well as between glosses and text.

2.2 Sign language datasets

Deep learning neural networks heavily depend on substantial datasets, but sign language data present unique challenges that demand comprehensive solutions. Table 1 present some datasets.

Issues of dataset size and diversity emerge due to the constraints of limited resources and time, resulting in small sample sizes that often fail to capture the wide array of sign language variations and complexities. For example, a dataset confined to a single environmental setting or a narrow demographic of signers may hinder the model's ability to adapt to diverse real-world situations and communities.

The integrity of annotation quality and consistency is paramount for training robust models. The nuanced nature of sign language, compounded by potential variations in annotator interpretations, poses considerable obstacles to achieving annotation uniformity and accuracy [9].

Privacy and ethical considerations are also non-negotiable aspects of collecting and sharing sign language data. Respecting participant privacy and consent is essential, as is ensuring the ethical construction and use of datasets to prevent discrimination and misconceptions [10].

The challenges associated with gathering sign language data have resulted in a scarcity of available corpora. To address this, researchers must engage in collaborations with the deaf and hard-of-hearing community and explore advanced methods for the collection and refinement of sign language data.

Table 1. This table give a simple presentation about sign language datasets.

Datasets	Languages	Description
RWTH-PHOENIX-Weather 2014	German Sign Language (DGS)	This dataset contains continuous sign language video recordings from weather forecasts. It includes both the videos and corresponding gloss-level annotations.
ASLLVD	American Sign Language (ASL)	A large-scale dataset containing videos of ASLLVD annotations such as handshapes, locations, and movements.
BSL Corpus	British Sign Language (BSL)	BSL Corpus Project provides a collection of BSL videos with annotations, focusing on linguistic and sociolinguistic variation.
CSL-500 Dataset	Chinese Sign Language(CSL)	The CSL-500 is a large-scale dataset for CSL. It includes 500 commonly used sign language words, each with 1,000 video performances. This dataset is extremely valuable for the research and development of Chinese SLR systems
SignBank	Multiple sign languages, including	SignBank is a lexical database system that contains detailed information about individual

	ASL, BSL, and others.	signs, including their phonological features, glosses, translations, and example videos. It is used primarily for research, teaching, and developing sign language dictionaries.

3 SLT works overview

Necati Cihan Camgöz and Simon Hadfield, along with other scholars [11], have pioneeringly shifted the paradigm in SLR to focus on SLT. This innovative approach involves converting sign language from video content into spoken translations. Their research addresses the variances in word order and grammatical structure between sign and spoken languages, leveraging the Neural Machine Translation (NMT) architecture. This architecture integrates end-to-end learning and pre-training to simultaneously grasp the spatial characteristics of sign language, the underlying linguistic model, and the correspondence between sign and spoken languages. The researchers have curated the first publicly accessible continuous SLT dataset, RWTH-PHOENIX-Weather 2014T (PHOENIX14T), which is derived from the PHOENIX14 dataset and provides spoken language translations along with lexical annotations for German sign language videos. With over 950,000 frames, 67,000 sign language words, and 99,000 German words, this dataset offers a robust empirical foundation. Their study presents quantitative and qualitative results across diverse SLT configurations, marking a significant advancement in this emergent domain. The peak translation performance yields a BLEU-4 [12] score of 19.26, and the frame-level and lexical-level segmentation networks achieved BLEU-4 scores of 9.58 and 18.13, respectively, within the end-to-end framework.

Yin Kayo and Jesse Read [8] proposed the STMC-Transformer encoder-decoder framework as a replacement for RNN in the field of SLT. This model, grounded on the robust Transformer architecture, integrates a Spatial Multi-cue (SMC) module designed to extract spatial characteristics from a variety of visual cues such as facial expressions, hand gestures, full frames, and postures. It further employs a series of Temporal Multi-cue (TMC) modules and temporal pooling layers to capture temporal dependencies between inter-cue and intra-cue features across different time intervals, simultaneously maintaining the distinctiveness of each cue and exploring their interactions. Bidirectional LSTM (BiLSTM) and Connectionist Temporal Classification (CTC) units facilitate sequential learning and reasoning for both inter-cue and intra-cue feature dynamics. The model was trained on the PHOENIX14T dataset, and it achieved a 5-point and 7-point BLEU score enhancement over the previous state-of-the-art in Gloss-to-Text and Video-to-Text translation tasks, respectively, validating the effectiveness of the STMC-Transformer. The model's superior performance in video-to-text translation as opposed to GT gloss translation indicates that glosses may not be the most efficient representation for sign language, thus justifying the exploration of more proficient representation methods.

Amit Moryossef and Kayo Yin, along with other researchers [13], have underscored the paucity of sign language corpora。 This limitation that hampers the performance of models based on the Transformer architecture. To address this issue, they proposed two rule-based heuristic methods for generating pseudo-parallel sentence pairs from monolingual spoken language data, thereby augmenting machine translation capabilities in resource-constrained environments. Their experimental work involved modifying the Transformer Gloss-to-Text model, originally created by Kayo Yin and Jesse Read, by replacing traditional word segmentation with Byte-Pair Encoding (BPE) to improve handling of unknown words and overall efficiency. Pre-training on synthetic data yielded a notable enhancement in translation

performance, with BLEU score improvements of 3.14 for American Sign Language to English and 2.20 for German Sign Language to German translation tasks.

Harry Walsh and Ben Saunders, along with other researchers [14], have conducted an in-depth analysis of the process by which spoken sentences are converted into generated sign language videos. This process has been delineated into two primary stages: the initial conversion of spoken sentences into a sequence of vocabulary, followed by the subsequent translation of this sequence into sign language videos. The current study is primarily concerned with the first phase, focusing on the transition from spoken sentences to vocabulary sequences. This research marks a significant progression from the traditional text-to-gloss approach to a more sophisticated text-to-notation system, with HamNoSys serving as the notational framework. The experimental methodology incorporates advanced techniques such as tokenization, word embedding, and the application of Transformer models for training purposes. A variety of tokenization strategies, including Word, Character, BPE, and WordPiece, have been implemented to evaluate their influence on translation efficacy. For word embedding, cutting-edge language models such as BERT and Word2Vec are employed to produce enhanced sentence-level embeddings. The training phase incorporates supervised learning to predict hand shapes and HamNoSys annotations, thereby improving the quality of translation. In terms of dataset selection, the study leverages the PHOENIX14T dataset alongside the MineDGS datas. The model has achieved a BLEU-4 score of 26.99 on the MineDGS dataset and a BLEU score of 25.09 on the PHOENIX14T dataset. Both of them represent new benchmarks in the field of SLT These results validate the effectiveness and innovation of the proposed method.

Zifan Jiang, Amit Moryossef, and their colleagues [15] engaged in the interpolation of SignWriting and spoken language within advanced Machine Translation frameworks. The choice to focus on SignWriting was informed by its extensive benefits, such as its linguistic universality, ease of understanding, comprehensive documentation, and robust computational backing, including full compatibility with Unicode and ASCII standards. Although it presents a pictographic interface, SignWriting is a systematically organized script, where each sign is depicted through a combination of box marks, letters, and punctuation arranged in a two-dimensional space. Utilizing the largest available dataset, SignBank from SignPuddles, the research team developed cutting-edge techniques to parse, factorize, decode, and assess SignWriting, drawing on the principles of neural factorized machine translation. Their method demonstrated a BLEU score over 30 when translating from American Sign Language to English and BLEU scores surpassing 20 in bilingual sign-to-speech translations. Central to this study is the integration of a symbolic intermediary text representation, as opposed to the direct end-to-end translation between sign language videos/gestures and spoken text.

4 Challenges and Proposals

Nonetheless, SLT continues to confront a multiplicity of unresolved challenges, including: 1. The paucity of extensive parallel corpora of sign language videos and their spoken language translations, which hampers model training and evaluation. 2. The multifaceted and variable nature of sign language representation, with regional idiosyncrasies, and the inclusion of facial expressions and body language, presenting additional complexities for SLT systems. 3. The prevalent use of an intermediary representation, such as gloss, which can be imprecise, while the exploration of more robust notational systems remains limited. 4. The absence of sign language expressions for certain proper nouns, with users often resorting to fingerspelling, a practice that current SLT approaches struggle to accurately discern. 5. The technical limitations of existing mobile devices, which impede the real-time, portable application of SLT technology in everyday settings.

To navigate these impediments, the following strategies are proposed: 1. Intensify collaboration with the deaf community to not only expand the breadth of sign language datasets but also to ensure that SLP technologies are aligned with their communicative requirements and perspectives. 2. Given the resource-constrained nature of sign languages, the adoption of transfer learning, utilizing pre-trained models on rich, general datasets, as a means to reduce dependency on extensive annotated data. 3. The integration of large language models(LLMs), which are gaining prominence in deep learning for their proficiency in multilingual tasks, can significantly bolster SLT by harnessing their robustness in comprehension, reasoning, and generalization, and expediting application deployment through cloud computing platforms. 4. SLP necessitates a fusion of CV and NLP expertise, calling for researchers to cultivate collaborative abilities and interdisciplinary knowledge to advance both technological and theoretical integration.

5 Conclusion

SLT represents a cornerstone domain within the broader spectrum of SLP. Its primary objective is to bridge the gap between sign language videos and spoken language texts, a pursuit that spans the disciplines of CV and NLP. This intersectionality is underpinned by technologies such as image recognition, semantic understanding, and machine translation. This Paper encapsulates the seminal advancements in the SLT field from recent years, while also casting a critical eye towards potential research vectors and the inherent challenges that accompany them.

Central to the SLT research paradigm is the accurate representation of sign language, which is currently approached through diverse methodologies including video capture, pose estimation, notation systems, and annotation techniques. While video capture provides an holistic record of sign language expressions, it is not without its drawbacks such as high processing costs and significant redundancy. Pose representation, on the other hand, distills the essence of signs through animated skeletal models, albeit in a high-dimensional and continuous format. Sign systems, by contrast, offer a discrete and reduced-dimensional framework that aligns more closely with NLP methodologies. Annotation, crucial for the indexing and interpretation of sign language, still necessitates advancements in precision.

The progression of SLT is contingent upon the availability of large-scale, high-quality datasets of sign language, which currently confront challenges related to size, diversity, consistency in annotation, as well as privacy and ethical considerations. Collaborative efforts with the deaf community, alongside innovative approaches to data collection, are thus indispensable in mitigating the scarcity of these resources.

This paper surveys a spectrum of SLT models and frameworks that have emerged within the field., which are reviewed in this paper. Building upon these works, the paper analyzes the challenges faced by SLT, approaches including dataset development, representation techniques, and application scenarios. Finally, feasible proposals are offered: enhancing community collaborations, leveraging techniques such transfer learning, LLMs, and cloud as well as bolster researchers' knowledge base and capabilities.

References

1. K. Yin, A. Moryossef, J. Hochgesang, Y. Goldberg, and M. Alikhani, Including Signed Languages in Natural Language Processing. Association for Computational Linguistics. 1, 7347–7360, (2021)
2. R. Rastgoo, K. Kiani, S. Escalera, and M. Sabokrou, Sign Language Production: A Review. Computing Research Repository. 2103.15910, 3451-3461, (2021)

3. R. Zuo, F. Wei, and B. Mak, Natural Language-Assisted Sign Language Recognition. CVPR 2023, 2303.12080, 14890-14900, (2023):
4. Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 43.1, 172-186 (2021)
5. H. Thomas, HamNoSys—Representing sign language data in language resources and language processing contexts (2004)
6. W. Stokoe, Sign language structure: An outline of the visual communication systems of the American deaf. Journal of deaf studies and deaf education. 10.1, 3-37 (2005)
7. V. Sutton, https://www.signwriting.org/
8. K. Yin and J. Read, Better Sign Language Translation with STMC-Transformer. International Committee on Computational Linguistics, 5975–5989 (2020)
9. S. Hassan, M. Seita, L. Berke, Y. Tian, E. Gale, S. Lee, and M. Huenerfauth, ASL-Homework-RGBD Dataset: An annotated dataset of 45 fluent and non-fluent signers performing American Sign Language homeworks. In Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, 67–72 (2022)
10. A. Voskou, K. P. Panousis, H. Partaourides, K. Tolias, and S. Chatzis. A New Dataset for End-to-End Sign Language Translation: The Greek Elementary School Dataset. 2023 IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS, ICCVW 2310.04753, 1958-1967 (2023)
11. N. Cihan Camgöz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, Neural Sign Language Translation, Computer Vision and Pattern Recognition, 7784-7793 (2018)
12. K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318 (2002)
13. A. Moryossef, K. Yin, G. Neubig, and Y. Goldberg, Data Augmentation for Sign Language Gloss Translation.", AT4SSL@MTSummit 2105.07476, 1-11 (2021)
14. H. Walsh, B. Saunders, and R. Bowden, Changing the Representation: Examining Language Representation for Neural Sign Language Production. CoRR (2022)
15. Z. Jiang, A. Moryossef, M. Mueller, and S. Ebling, Machine Translation between Spoken Languages and Signed Languages Represented in SignWriting. 17TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, EACL 2023, 1706-1724 (2023)