

A Survey of Sign Language Translation

Jiwon Shin

University of California, Berkeley

INFO 159: Natural Language Processing

May 9, 2022

Abstract

Machine translation (MT), the task of automatically translating one language to another, has been around for a while. While spoken languages have been the focus of MT, sign languages have also started getting attention in the MT and NLP space. This paper presents an overview of the relatively new and major breakthrough in Sign Language Translation (SLT).

1 Introduction

Sign languages are the primary means of communication for many of the hearing-impaired communities worldwide. Just as each country or region has its own spoken language, there are numerous sign languages from around the world, i.e. there isn't a single universal sign language. Like spoken languages, sign languages have their own specific linguistic rules (Camgöz et al., 2018) and are visual-based natural languages (Zheng et al., 2020).

However, sign languages and spoken languages differ in dimensionality. Sign language is a multi-dimensional form of communication (Yin and Read, 2020b) that simultaneously relies on both manual features, such as hand shape and pose, as well as non-manual features, such as facial expression and movement of the head and body (Camgöz et al., 2020b). Sign languages are not directly translated to spoken languages word by word (Camgöz et al., 2018), whereas spoken languages are translated in a linear pattern (Yin and Read, 2020b). The complex dimensionality and high-density information of sign languages make research challenging in this field of study.

Sign Language Translation (SLT), the task of translating sign languages to spoken languages, is one of the most important and under-examined tasks of Sign Language Processing (SLP). SLT approaches involve two

steps: *video-to-gloss* recognition and *gloss-to-text* translation (Moryossef et al., 2021). There has been extensive research done on the first step of recognition using Computer Vision (CV) but limited research to improve the translation model in a Natural Language Processing (NLP) space.

Research in SLT can benefit the hearing-impaired community. In a predominantly hearing society, it can be challenging for hearing-impaired individuals to effectively communicate with non-signers (Moryossef et al., 2021). SLT can become an important application to bridge the gap between signers and non-signers while allowing each party to use their preferred language (Moryossef et al., 2021; Zheng et al., 2020).

Our goal in this paper is to provide an overview of the sign language translation research and a sense of the current state of the art in this subfield in 2022. In Section 2, we discuss the translation tasks, different approaches, and evaluation metrics in the field of sign language translation by surveying major historical papers. In Section 3, we discuss the challenges and the current state of the art in SLT in 2022. Finally, we conclude the paper by discussing future directions in this subfield of NLP.

2 Systems

Next, we characterize existing SLT systems along three dimensions: the *translation task* a system was developed for as well as the *approach* and the *evaluation metrics* it used.

2.1 Sign Language Translation Tasks

According to Yin et al. (2021b), there are six common SLP tasks: detection, identification, segmentation, recognition, translation, and production. For the scope of this paper, we will focus on translation.

SLT is the task of translating sign languages to spoken languages. There are two main methods of performing translation: an end-to-end translation and gloss-to-text translation.

First, an end-to-end translation, often referred to as Sign2Text or Sign2Gloss2Text, translates sign videos to spoken languages. This method is often decomposed into two steps: video-to-gloss recognition and gloss-to-text translation (Moryossef et al., 2021). First, video-to-gloss recognition, also known as Sign Language Recognition (SLR), is a tokenization system that generates glosses from sign language videos (Yin and Read, 2020b). Gloss is a sequence of transcribed, isolated words in spoken language that are yet in a comprehensible sentence (Moryossef et al., 2021). Then, gloss-to-text translation is a translation system that translates the recognized glosses (Yin and Read, 2020b) into spoken language.

Gloss-to-text translation, often referred as Gloss2Text, deals solely with the latter part of the end-to-end transition. As mentioned above, gloss-to-text translation translates glosses to comprehensive spoken language sentences.

We will see how both methods are used throughout the history of SLT research.

2.2 Approaches

In this section, we will be exploring major historical papers that have defined the area of SLT.

Historically, there has been more research done in the SLR area compared to SLT. Here are some of the early works of SLR: Zhao et al. (2000) proposed a rule-based system; Stein et al. (2006) proposed a morpho-syntax based SMT system; Lee and Xu (1996) proposed a Hidden Markov Model for SLR (Parton 2006); Example-Based Machine Translation has also been explored for SLR systems (Yang et al., 2001; Morrissey and Way, 2005); Lichtenauer et al. (2008) proposed a statistical Dynamic Time Warping (SDTW) system.

More recently, with the development of deep learning, Neural Machine Translation (NMT) has arisen as the most powerful algorithm to perform this task (Lanners, 2019). In the following section, we will be taking a deeper look into NMT models developed for SLT.

Neural Sign Language Translation/ Neural Machine Translation (NMT)

The standard algorithm for NMT is the encoder-decoder network, also called the sequence to sequence network, an architecture that can be implemented with RNNs or with Transformers (Jurafsky and Martin, 2022).

2.2.1 Encoder-Decoder Model (Seq2Seq)

Depending on the translation task, an encoder-decoder network uses an encoder to encode the input (whether that be a text or video) and a decoder to decode the encoded input to output a comprehensive spoken language text or gloss.

2.2.2 Using Attention

One problem with the encoder-decoder approach is the bottleneck problem in which the neural network takes in a fixed-length vector as its input, which makes translating long sentences difficult.

To address this issue, Bahdanau et al. (2015) proposes an attention mechanism to compute the context vector by letting the model search for relevant parts from the hidden states of the encoder (Ko et al., 2018). In this paper, they create a new model architecture with an alignment function that computes how well the encoder and decoder hidden states match.

Later, Luong et al. (2015) has further improved this approach by adding two effective attention mechanisms for NMT: the global approach which always looks at all source positions and the local one that only attends to a subset of source positions at a time.

Furthermore, Camgöz et al. (2018) utilize both Bahdanau and Luong attention to propose the very first end-to-end translation network (Camgöz et al., 2020b). They propose a network with 2D-CNN based spatial embedding, various tokenization methods including RNN-HMM hybrids, and attention based encoder-decoder networks to perform German sign language translation from sign video frames. The paper also introduces the RWTH-PHOENIX-Weather 2014T which is the first publicly available and most used Continuous SLT dataset. Using the PHOENIX-Weather 2014T, the model yields a BLEU-4 score of 19.26 for Gloss2Text and 18.13 for Sign2Gloss2Text.

Following this, while the previous encoder-decoder architectures are based on RNN

cells, Ko et al. (2018) explores a new multi-head attention network architecture proposed by Vaswani et al. (2017) which is based solely on attention mechanisms without any recurrence and convolution. The multi-head attention is used in three different ways: encoder-decoder attention, encoder self-attention, and decoder self-attention. As well as introducing the new Korea Electronics Technology Institute (KETI) sign language dataset, they specifically concentrate on estimating human keypoints to extract glosses, then explore 4 different attention types in their research for the translation of Korean sign language.

In terms of Gloss2Text translation, Arvanitis et al. (2019) applies a Seq2Seq model with attention mechanisms to translate ASL glosses of the ASLG-PC12 dataset.

2.2.3 Encoder-Decoder with Transformer

The encoder-decoder architecture can also be implemented using transformers (rather than RNN/LSTMs) as the component modules (Jurafsky and Martin, 2022).

Camgoz et al. (2020b) proposes an end-to-end translation system with a multi-task transformer that jointly performs CSLR (tokenization) and SLT (translation). To help the achievement of CSLR, they propose a Sign Language Recognition Transformer (SLRT), an encoder transformer model trained using a Connectionist Temporal Classification (CTC) loss, to predict sign gloss sequences. The SLT task proposes training an autoregressive transformer decoder model, named Sign Language Translation Transformer (SLTT), which exploits the spatio-temporal representations learned by the SLRT. SLTT is trained to predict one word at a time to generate the corresponding spoken language sentence. This method improves the BLEU-4 score to 21.32 using the PHOENIX-2014T dataset.

Furthermore, Yin and Read (2020a, 2020b) are the first ones to propose the STMC-Transformer for SLT by performing both Gloss2Text and Sign2Gloss2Text experiments. Rather than only using full frame information for CSLR, this method is unique in that it uses multiple cues such as face, hand, full frame, pose information. They tokenize continuous signed language videos to glosses by using the STMC-Transformer network, which involves a spatial multi-cue (SMC) module, temporal multi-cue

(TMC) module, and Bi-directional Long Short-Term Memory (BiLSM) and CTC units for sequence learning. Glosses from the STMC-Transformer are then inputted into a two-layered transformer. Using the PHOENIX-Weather 2014T, the model yields a BLEU-4 score of 24 on the test set and 25.40 using an ensemble of five models. This experiment demonstrates how Transformer obtains better SLT performance than previous RNN-based networks.

2.3 Evaluation

In this section, we discuss the metrics used to evaluate the SLT systems.

The most widely used evaluation metric is BLEU (for BiLingual Evaluation Understudy), an automatic machine translation evaluation that ranges from 0 to 100 in percentage. BLEU indicates how similar the machine translation correlates to human translation (numerical “translation closeness” metric) (Papineni et al., 2002), with higher values representing more similarity between the two texts. Specifically, BLEU-4 scores are widely used in the SLT research, and they indicate a 4-gram overlap between machine translation output and reference/human translation.

Other machine translation metrics include Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) and Metric for Evaluation of Translation with Explicit ORDERing (METEOR) (Dorr et al., 2010) to name a few. They both automatically determine the quality of texts and similarity between machine translations and human translation.

3 The State of the Art

In this section, we discuss challenges in sign language transition and the current state of the art in this subfield of NLP in 2022.

Working on SLT to improve the translation model has been a challenging task due to their visual-gestural modality, spatial-temporal aspect, lack of written form (Yin et al., 2021b), and data scarcity. One of the major challenges is the lack of parallel corpora (Bragg et al., 2019; De Coster et al., 2021). In order to conduct SLT research, there needs to be parallel text or glosses with sign videos. Most of the SLR research has relied on weakly annotated datasets. In an effort to solve this problem, Forster et al. (2012, 2014) released the

RWTH-PHOENIX-Weather 2012 dataset and its extended version RWTH-PHOENIX-Weather 2014 dataset. Later, Camgöz et al. (2018) published an annotated dataset called PHOENIX-Weather 2014T, which constitutes a parallel corpus including sign language videos, sign-gloss annotations and also German translations from new anchors in the weather forecasting domain, which are all segmented into parallel sentences. However, because these videos have been filmed in a studio environment and are specific to the weather forecast, it fails to represent the real-world. Despite recent promising results on this dataset, it is still considered a tiny dataset from the perspective of Neural Machine Translation (NMT) (De Coster et al., 2021). To reiterate Yin et al. (2021b), we need more annotated data to train SLT models for real-world applications. By doing so, we hope to learn more about the linguistic rules of sign languages and be more inclusive of the hearing-impaired communities in NLP.

With the development of deep learning and neural machine translation, sign language translation systems have slowly been getting attention. So far, most of the machine translation studies have been conducted on spoken languages, but it is time to shine light on sign languages as well. Within the field of sign language processing, there has been relatively more work done on sign language recognition compared to sign language translation. It was only recently that Camgöz et al. (2018) proposed the very first end-to-end neural translation model. Since then, NMT architectures with attention mechanisms and transformers have been developed. Yet, the field of sign language translation is very new and open. As of now, the best performing model was proposed by Yin and Read in 2020b which uses the STMC-Transformer to yield a BLEU-4 score of 24 on the 2014T test set. Even then, De Coster et al. (2021) argues that the improvements from Yin and Read (2020b) are related to feature extraction rather than network architecture, calling attention to the need of improving the translation model by creating a more powerful encoder-decoder model.

Although there are limitations and challenges in SLT, there is so much work that can be done in this space of NLP. Whether it be building NLP pipelines, such as tokenization, syntactic analysis, named entity recognition (NER), and coreference resolution, or collecting

real-world data (Yin et al., 2021b), the future for sign language processing is vast and open.

4 Conclusion

In this paper, we provided an overview of the sign language translation research, discussing major papers and their tasks, approaches, and evaluation metrics. We also discussed the current state-of-the-art of SLT and the future improvements in this space. We hope more research on SLT can bring light to the hearing-impaired communities, help us learn more about the linguistic insights to build models, and bridge the gap between signers and non-signers.

References

- Nikolaos Arvanitis, Constantinos Constantinopoulos, and Dimitris Kosmopoulos. 2019. Translation of Sign Language Glosses to Text Using Sequence-to-Sequence Attention Models. In the *15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 296-302, doi: 10.1109/SITIS.2019.00056.
- Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 16–31. <https://doi.org/10.1145/3308561.3353774>
- Necati Cihan Camgöz., Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.
- Necati Cihan Camgöz., Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.
- *Mathieu De Coster, Karel D’Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden,

- Mieke Van Herreweghe, and Joni Dambre. 2021. [Frozen Pretrained Transformers for Neural Sign Language Translation](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 88–97, Virtual. Association for Machine Translation in the Americas.
- Bonnie Dorr, Matt Snover, and Nitin Madnani. 2010. Part 5: Machine Translation Evaluation Chapter 5.1 Introduction.
- *Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. *LREC*.
- *Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. [Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. 2022.
- Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choong Sang Cho. 2018. Neural Sign Language Translation based on Human Keypoint Estimation. <http://arxiv.org/abs/1811.11436>
- Quinn Lanners. 2019. Neural Machine Translation. *Medium*, Towards Data Science, <https://towardsdatascience.com/neural-machine-translation-15ecf6b0b>.
- Christopher Lee and Yangsheng Xu. 1996. Online, interactive learning of gestures for human/robot interfaces. In *Proceedings of IEEE International Conference on Robotics and Automation 4* (1996): 2982-2987 vol.4.
- Jeroen Lichtenauer, Emile Hendriks, and Marcel Reinders. 2008. Sign language recognition by combining statistical dtw and independent classification. *IEEE transactions on pattern analysis and machine intelligence*, 30:2040–6, 12.
- Chin-Yew Lin. 2004. [ROUGE: A Package for Automatic Evaluation of Summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- *Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. [Effective Approaches to Attention-based Neural Machine Translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- *Sara Morrissey and Andy Way. 2005. [An Example-Based Approach to Translating Sign Language](#). In *Workshop on example-based machine translation*, pages 109–116, Phuket, Thailand.
- *Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data Augmentation for Sign Language Gloss Translation. <https://arxiv.org/abs/2105.07476>
- *Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a Method for Automatic Evaluation of Machine Translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Becky Sue Parton. Sign Language Recognition and Translation: A Multidisciplinary Approach From the Field of Artificial Intelligence. *Journal of deaf studies and deaf education*. 11. 94-101. 10.1093/deafed/enj003.
- *Daniel Stein, Jan Bungeo, and Hermann Ney. 2006. [Morpho-Syntax Based Statistical Methods for Automatic Sign Language Translation](#). In *Proceedings of the 11th Annual conference of the European Association for Machine Translation*, Oslo, Norway. European Association for Machine Translation.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NIPS 2017)*, pages 6000–6010, 2017. 6, 7, 8, 10 [54] R. Vedantam, C. L. Zitnick, and D. Parikh. C
- *Jie Yang, Jiang Gao, Ying Zhang, and Alex Waibel. 2001. Towards automatic sign translation. In *Proceedings of the first international conference on Human language technology research (HLT '01)*. Association for Computational Linguistics, USA, 1–6. <https://doi.org/10.3115/1072133.1072223>
- Kayo Yin and Jesse Read. 2020a. Attention is all you sign: sign language translation with transformers. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, volume 4.
- *Kayo Yin and Jesse Read. 2020b. 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.

*Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021b. Including signed languages in natural language processing. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Virtual.

Liwei Zhao, Karin Kipper, William Schuler, Christian Vogler, Norman Badler, and Martha Palmer. 2000. *A machine translation system from English to American Sign Language*. In *Proceedings of the Fourth Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 54–67, Cuernavaca, Mexico. Springer.

Jiangbin Zheng and Zheng Zhao and Min Chen and Jing Chen and Chong Wu and Yidong Chen and Xiaodong Shi, and Yiqi Tong. 2020. An Improved Sign Language Translation Model with Explainable Adaptations for Processing Long Sign Sentences. <https://doi.org/10.1155/2020/8816125>

*10 required venue papers