

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 singers 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

250
251 cells, Ko et al. (2018) explores a new multi-head
252 attention network architecture proposed by
253 Vaswani et al. (2017) which is based solely on
254 attention mechanisms without any recurrence and
255 convolution. The multi-head attention is used in
256 three different ways: encoder-decoder attention,
257 encoder self-attention, and decoder self-attention.
258 As well as introducing the new Korea Electronics
259 Technology Institute (KETI) sign language
260 dataset, they specifically concentrate on
261 estimating human keypoints to extract glosses,
262 then explore 4 different attention types in their
263 research for the translation of Korean sign
264 language.

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

2.2.3 Encoder-Decoder with Transformer

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

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

290 Furthermore, Yin and Read (2020a,
291 2020b) are the first ones to propose the STMC-
292 Transformer for SLT by performing both
293 Gloss2Text and Sign2Gloss2Text experiments.
294 Rather than only using full frame information for
295 CSLR, this method is unique in that it uses multiple
296 cues such as face, hand, full frame, pose
297 information. They tokenize continuous signed
298 language videos to glosses by using the STMC-
299 Transformer network, which involves a spatial
multi-cue (SMC) module, temporal multi-cue

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

2.3 Evaluation

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

312 The most widely used evaluation metric
313 is BLEU (for BiLingual Evaluation Understudy),
314 an automatic machine translation evaluation that
315 ranges from 0 to 100 in percentage. BLEU
316 indicates how similar the machine translation
317 correlates to human translation (numerical
318 “translation closeness” metric) (Papineni et al.,
319 2002), with higher values representing more
320 similarity between the two texts. Specifically,
321 BLEU-4 scores are widely used in the SLT
322 research, and they indicate a 4-gram overlap
323 between machine translation output and
324 reference/human translation.

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

3 The State of the Art

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

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

- 450 Mieke Van Herreweghe, and Joni Dambre. 2021. 500
 451 **Frozen Pretrained Transformers for Neural Sign 501**
 452 **Language Translation.** In *Proceedings of the 1st 502*
 453 *International Workshop on Automatic Translation 503*
 454 *for Signed and Spoken Languages (AT4SSL)*, pages 504
 455 88–97, Virtual. Association for Machine 505
 456 Translation in the Americas.
- 457 Bonnie Dorr, Matt Snover, and Nitin Madnani. 2010. 506
 458 Part 5: Machine Translation Evaluation Chapter 5.1 507
 459 Introduction.
- 460 *Jens Forster, Christoph Schmidt , Thomas Hoyoux , 508
 461 Oscar Koller , Uwe Zelle , Justus Piater , and 509
 462 Hermann Ney. 2012. RWTH-PHOENIX-Weather: 510
 463 A Large Vocabulary Sign Language Recognition 511
 464 and Translation Corpus. *LREC*.
- 465 *Jens Forster, Christoph Schmidt, Oscar Koller, Martin 512
 466 Bellgardt, and Hermann Ney. 2014. **Extensions of 513**
 467 the Sign Language Recognition and Translation 514
 468 Corpus RWTH-PHOENIX-Weather.
- In
- Proceedings of the Ninth International Conference 515*
-
- on Language Resources and Evaluation (LREC'14)*
- ,
-
- pages 1911–1916, Reykjavik, Iceland. European
-
- Language Resources Association (ELRA).
- 469 Daniel Jurafsky and James H. Martin. Speech and 516
 470 Language Processing. 2022.
- 471 Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and 517
 472 Choong Sang Cho. 2018. Neural Sign Language 518
 473 Translation based on Human Keypoint Estimation. 519
<http://arxiv.org/abs/1811.11436>
- 474 Quinn Lanners. 2019. Neural Machine Translation. 520
Medium, Towards Data Science,
<https://towardsdatascience.com/neural-machine-translation-15ecf6b0b>.
- 475 Christopher Lee and Yangsheng Xu. 1996. Online, 521
 interactive learning of gestures for human/robot 522
 interfaces. In *Proceedings of IEEE International 523*
Conference on Robotics and Automation 4 (1996): 524
 2982–2987 vol.4.
- 476 Jeroen Lichtenauer, Emile Hendriks, and Marcel 525
 Reinders. 2008. Sign language recognition by 526
 combining statistical dtw and independent 527
 classification. *IEEE transactions on pattern analysis 528*
and machine intelligence, 30:2040–6, 12.
- 477 Chin-Yew Lin. 2004. **ROUGE: A Package for 529**
Automatic Evaluation of Summaries. In *Text 530*
Summarization Branches Out, pages 74–81, 531
 Barcelona, Spain. Association for Computational 532
 Linguistics.
- 478 *Thang Luong, Hieu Pham, and Christopher D. 533
 Manning. 2015. **Effective Approaches to Attention- 534**
based Neural Machine Translation. In *Proceedings 535*
of the 2015 Conference on Empirical Methods in 536
Natural Language Processing, pages 1412–1421,
- 479 Lisbon, Portugal. Association for Computational 537
 Linguistics.
- 480 *Sara Morrissey and Andy Way. 2005. **An Example- 538**
Based Approach to Translating Sign Language. In 539
Workshop on example-based machine translation, 540
 pages 109–116, Phuket, Thailand.
- 481 *Amit Moryossef, Kayo Yin, Graham Neubig, and 541
 Yoav Goldberg. 2021. Data Augmentation for Sign 542
 Language Gloss Translation. 543
<https://arxiv.org/abs/2105.07476>
- 482 *Kishore Papineni, Salim Roukos, Todd Ward, and 544
 Wei-Jing Zhu. 2002. **Bleu: a Method for Automatic 545**
Evaluation of Machine Translation. In *Proceedings 546*
of the 40th Annual Meeting of the Association for 547
Computational Linguistics, pages 311–318, 548
 Philadelphia, Pennsylvania, USA. Association for 549
 Computational Linguistics.
- 483 Becky Sue Parton. Sign Language Recognition and 550
 Translation: A Multidisciplined Approach From the 551
 Field of Artificial Intelligence. *Journal of deaf 552*
 studies and deaf education. 11. 94–101.
 10.1093/deafed/enj003.
- 484 *Daniel Stein, Jan Bungeroth, and Hermann Ney. 553
 2006. **Morpho-Syntax Based Statistical Methods for 554**
Automatic Sign Language Translation. In *Proceedings 555*
of the 11th Annual conference of the European 556
Association for Machine Translation, Oslo, Norway. 557
 European Association for Machine 558
 Translation.
- 485 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob 559
 Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz 560
 Kaiser, and Illia Polosukhin. 2017. Attention Is All 561
 You Need. In *Advances in Neural Information 562*
Processing Systems 30: Annual Conference on 563
Neural Information Processing Systems 2017 (NIPS 564
 2017), pages 6000–6010, 2017. 6, 7, 8, 10 [54] R. 565
 Vedantam, C. L. Zitnick, and D. Parikh. C
- 486 *Jie Yang, Jiang Gao, Ying Zhang, and Alex Waibel. 566
 2001. Towards automatic sign translation. In *Proceedings 567*
of the first international conference on Human 568
language technology research (HLT '01). Association 569
 for Computational Linguistics, USA, 570
 1–6. <https://doi.org/10.3115/1072133.1072223>
- 487 Kayo Yin and Jesse Read. 2020a. Attention is all you 571
 sign: sign language translation with transformers. In 572
Sign Language Recognition, Translation and 573
Production (SLRTP) Workshop-Extended Abstracts, 574
 volume 4.
- 488 *Kayo Yin and Jesse Read. 2020b. Better sign 575
 language translation with STMC-transformer. In 576
Proceedings of the 28th International Conference 577
on Computational Linguistics, pages 5975–5989, 578
 Barcelona, Spain (Online). International Committee 579
 on Computational Linguistics.

550 *Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav
 551 Goldberg, and Malihe Alikhani. 2021b. Including
 552 signed languages in natural language processing. In
 553 *Joint Conference of the 59th Annual Meeting of the*
 554 *Association for Computational Linguistics and the*
 555 *11th International Joint Conference on Natural*
 556 *Language Processing (ACL-IJCNLP)*, Virtual.

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

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

568
 569 *10 required venue papers
 570

571
 572
 573
 574
 575
 576
 577
 578
 579
 580
 581
 582
 583
 584
 585
 586
 587
 588
 589
 590
 591
 592
 593
 594
 595
 596
 597
 598
 599

600
 601
 602
 603
 604
 605
 606
 607
 608
 609
 610
 611
 612
 613
 614
 615
 616
 617
 618
 619
 620
 621
 622
 623
 624
 625
 626
 627
 628
 629
 630
 631
 632
 633
 634
 635
 636
 637
 638
 639
 640
 641
 642
 643
 644
 645
 646
 647
 648
 649