Semi-Supervised Spoken Language Glossification

Huijie Yao[1] **Wengang Zhou**[1,*] **Hao Zhou**[2] **Houqiang Li**[1,*]
[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China [2]Baidu Inc.
{yaohuijie, zhouh156}@mail.ustc.edu.cn, {zhwg, lihq}@ustc.edu.cn

Abstract

Spoken language glossification (SLG) aims to translate the spoken language text into the sign language gloss, *i.e.*, a written record of sign language. In this work, we present a framework named Semi-Supervised Spoken Language Glossification (S^3LG) for SLG. To tackle the bottleneck of limited parallel data in SLG, our S^3LG incorporates large-scale monolingual spoken language text into SLG training. The proposed framework follows the self-training structure that iteratively annotates and learns from pseudo labels. Considering the lexical similarity and syntactic difference between sign language and spoken language, our S^3LG adopts both the rule-based heuristic and model-based approach for auto-annotation. During training, we randomly mix these complementary synthetic datasets and mark their differences with a special token. As the synthetic data may be less quality, the S^3LG further leverages consistency regularization to reduce the negative impact of noise in the synthetic data. Extensive experiments are conducted on public benchmarks to demonstrate the effectiveness of the S^3LG. Our code is available at https://github.com/yaohj11/S3LG.

1 Introduction

Sign Language is the most primary means of communication for the deaf. Translating between sign and spoken language is an important research topic, which facilitates the communication between the deaf and the hearing (Bragg et al., 2019; Yin et al., 2021). To support the development of applications, sign language gloss has been widely used as an intermediate step for generating sign language video from spoken language text (Saunders et al., 2020, 2022) or the inverse direction (Pu et al., 2020; Chen et al., 2022). The sign language gloss is the written representation of the signs. As a generally adopted way for sign language transcription, gloss is sufficient to convey most of the key information in sign language. In this work, we focus on the first step of the former task named spoken language glossification (SLG), which aims to translate the spoken language text into the sign language gloss.

SLG is typically viewed as a low-resource sequence-to-sequence mapping problem. The previous methods (Zhu et al., 2023; Walsh et al., 2022; Egea Gómez et al., 2021, 2022) rely on the encoder-decoder architectures (Luong et al., 2015; Sutskever et al., 2014) to jointly align the embedding space of both languages in a data-driven manner. Since the data collection and annotation of sign language requires specialized knowledge, obtaining a large-scale text-gloss dataset is time-consuming and expensive (De Coster et al., 2023). As a result, the performance of SLG models is limited by the quantity of parallel data (Camgoz et al., 2018; Zhou et al., 2021). Witnessing the success of introducing monolingual data to enhance low-resource translation quality (Cheng et al., 2016; Sennrich et al., 2016; Pan and Yang, 2009; Zoph et al., 2016) in neural machine translation (NMT), we are motivated to explore the accessible unlabeled spoken language texts to improve SLG.

In this work, we present a framework named Semi-Supervised Spoken Language Glossification (S^3LG) to boost SLG, which iteratively annotates and learns from pseudo labels. To implement the above idea, the proposed S^3LG adopts both the rule-based heuristic and model-based approach to generate pseudo glosses for unlabeled texts. The rule-based synthetic data has high semantic accuracy, however, the fixed rules make it difficult to cover complex expression scenarios. The model-based approach on the other hand is more flexible for learning the correspondence between sign language and spoken language and generates pseudo gloss with higher synthetic diversity. These complementary synthetic datasets are randomly mixed as a strong supplement for the training of the SLG

[*]Corresponding authors: Wengang Zhou and Houqiang Li

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9300–9312
August 11-16, 2024 ©2024 Association for Computational Linguistics

model. Besides, the model-based synthetic data is generated by the SLG model, which sets a good stage for iteratively re-training the SLG model.

In addition, S^3LG introduces a simple yet efficient design from three aspects. Firstly, in each iteration, the training process is separated into two stages, *i.e.*, pre-training and fine-tuning for domain adaptation. Secondly, to encourage the model to learn from the noisy pseudo labels, we apply the consistency regularization term to the training optimization and gradually increase the weight of the consistency regularization in the training curriculum. It enforces the consistency of the predictions with network perturbations (Gao et al., 2021) based on the manifold assumption (Oliver et al., 2018). Thirdly, to encourage the SLG model to learn complementary knowledge from different types of synthetic data, a special token is added at the beginning of input sentences to inform the SLG model which data is generated by the rule-based or model-based approach (Caswell et al., 2019). Through end-to-end optimization, our S^3LG achieves significant performance improvement over the baseline. Surprisingly, the experiments show that the translation accuracy on low-frequency glosses is promisingly improved. We conjecture that the SLG model acts differently in annotating the high-frequency and low-frequency glosses, and such bias is mitigated by the rule-based synthetic data.

In summary, our contributions are three-fold:

- We propose a novel framework S^3LG for SLG (namely, text-to-gloss translation), which iteratively annotates and learns from the synthetic data. It adopts two complementary methods, *i.e.*, the rule-based heuristic and model-based approach for auto-annotation.
- We further leverage consistency regularization to reduce the negative impact of noise in the synthetic data. The biases of the SLG model on low-frequency glosses are mitigated by incorporating the rule-based synthetic data.
- We conduct extensive experiments to validate our approach, which shows encouraging performance improvement on the two public benchmarks, *i.e.*, CSL-Daily (Zhou et al., 2021) and PHOENIX14T (Camgoz et al., 2018).

2 Related Work

In this section, we briefly review the related works on spoken language glossification and semi-supervised learning.

Spoken language glossification. Camgoz et al. (2018) publish the first sign language dataset PHOENIX14T and pioneer the linguistic research for sign language (De Coster et al., 2021; Cao et al., 2022). With the advance of NMT, the previous methods (Stoll et al., 2020; Saunders et al., 2020) adopt the encoder-decoder paradigm, which can be specialized using different types of neural networks, *i.e.*, RNNs (Yu et al., 2019), CNNs (Gehring et al., 2017). Considering gloss as a text simplification, Li et al. (2022) propose a novel editing agent. Instead of directly generating the sign language gloss, the agent predicts and executes the editing program for the input sentence to obtain the output gloss. By leveraging the linguistic feature embedding, Egea Gómez et al. (2021) achieve remarkable performance improvement. Egea Gómez et al. (2022) further apply the transfer learning strategy result in continues performance increasing. Recently, Zhu et al. (2023) first introduce effective neural machine translation techniques to SLG with outstanding performance improvements, which lays a good foundation for further research.

Semi-supervised learning. Generating pseudo labels for the unlabeled data is a widely adopted semi-supervised learning algorithm in low-resource NMT, known as back-translation (Gulcehre et al., 2015) and self-training (Zhang and Zong, 2016), respectively. With the target-side monolingual data, back-translation obtains pseudo parallel data by translating the target-side sentences into the source-side sequences. As an effective data augmentation method, it is widely adopted in the inverse task of SLG, *i.e.*, gloss-to-text translation (Moryossef et al., 2021; Zhang and Duh, 2021; Angelova et al., 2022; Chiruzzo et al., 2022). Due to the lack of the sign language gloss corpus, it is hard to incorporate large-scale monolingual data in the training process of SLG with back-translation (Rasooli and Tetreault, 2015). In contrast, the self-training requires source-side monolingual data to generate pseudo parallel data based on a functional source-to-target translation system. Since it is hard to optimize a neural translation system with an extremely limited amount of parallel data (Moryossef et al., 2021; Zhang and Duh, 2021; Xie et al., 2020), this motivates us to go along this direction and design more effective algorithms.

Different from the aforementioned methods, we focus on iteratively annotating and learning from

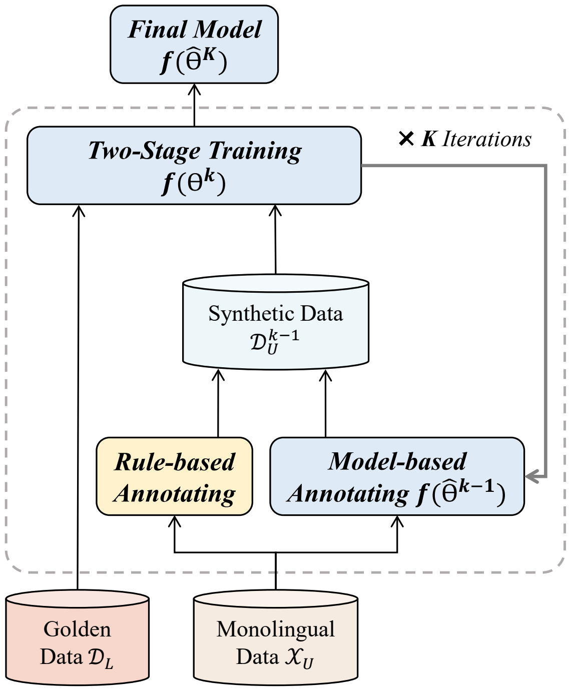

Figure 1: Overview of the proposed S^3LG. It iteratively annotates and learns from pseudo labels to obtain the final SLG model $f(\hat{\theta}^K)$ after total K iterations. For the k-th iteration, the synthetic data $\mathcal{D}_U^{k-1}$ is conducted by randomly mixing the two complementary pseudo glosses generated by the fixed rules and the previously obtained SLG model $f(\hat{\theta}^{k-1})$, respectively. Note that at the first iteration, only the rule-based synthetic data is available.

pseudo labels. Considering the lexical similarity and syntactic difference between sign language and spoken language, we adopt two complementary approaches (*i.e.*, rule-based heuristic and model-based approach) to generate synthetic data. Moreover, we put forward the consistency regularization and tagging strategy to reduce the negative impact of noisy synthetic data.

3 Methodology

In this section, we first introduce the overview of our S^3LG in Sec. 3.1; then we elaborate on the annotating methods for monolingual data in Sec. 3.2; and finally, we detail the training strategy in Sec. 3.3.

3.1 Overview

The primary objective of the SLG model is to acquire knowledge about the mapping $f(\theta) : \mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the collection of spoken language text and sign language gloss associated with the vocabulary $\mathcal{V}$, respectively. θ is the parameters of the SLG model. Most SLG model adopts the encoder-decoder architecture, where the input $\boldsymbol{x} \in \mathcal{X}$ is first encoded to devise a high-level context representation. It is then passed to the decoder to generate the output $\boldsymbol{y} \in \mathcal{Y}$. The encoder and decoder can be specialized using different types of neural networks. Given a set $\mathcal{D}_L = \{(\boldsymbol{x}_L^i, \boldsymbol{y}_L^i)\}_{i=1}^M$ of M labeled samples and a set $\mathcal{X}_U = \{\boldsymbol{x}_U^i\}_{i=1}^N$ of N unlabeled data, we aim to design a semi-supervised framework for SLG to improve text-to-gloss translation by exploring both the labeled and unlabeled data. To this end, we propose S^3LG, which iteratively annotates and learns from two complementary synthetic data generated by the rule-based heuristic and model-based approach, respectively.

Fig. 1 provides an overview of the S^3LG approach, which consists of three main steps, namely rule-based annotating, model-based annotating, and two-stage training. At the k-th iteration, the synthetic data $\mathcal{D}_U^{k-1} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_U^{i,k-1})\}_{i=1}^N$ is composed of two parts, *i.e.*, rule-based $\mathcal{D}_{U,r} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_{U,r}^i)\}_{i=1}^N$ and model-based synthetic data $\mathcal{D}_{U,m}^{k-1} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_{U,m}^{i,k-1})\}_{i=1}^N$. Based on the monolingual data $\mathcal{X}_U = \{\boldsymbol{x}_U^i\}_{i=1}^N$, the rule-based and model-based synthetic data are generated by the fixed rules and the functional SLG model $f(\hat{\theta}^{k-1})$ obtained from the previous iterations, respectively. We randomly mix the two complementary synthetic data and add a special token at the beginning of the input sentences. The synthetic data $\mathcal{D}_U^{k-1} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_U^{i,k-1})\}_{i=1}^N$ is then concatenated with the original golden data $\mathcal{D}_L = \{(\boldsymbol{x}_L^i, \boldsymbol{y}_L^i)\}_{i=1}^M$ as a strong supplement for training an SLG model $f(\theta^k)$, where $N \gg M$. After repeating K times, we obtain the final SLG $f(\hat{\theta}^K)$ model. Notably, in the first iteration, only the rule-based heuristic is available to generate the pseudo gloss sequences for the monolingual data, as in, $\mathcal{D}_U^0 = \mathcal{D}_{U,r}$.

3.2 Annotating Monolingual Data

Compared with the limited size of text-gloss pairs, the unlabeled spoken language sentences are easy to reach. To leverage both the labeled and unlabeled data to enhance the SLG performance, we employ the rule-based heuristic and model-based approach to generate the pseudo parallel data and use it to enrich the original golden data for training.

Rule-based annotating. Given that sign language gloss is annotated based on the lexical elements from the corresponding spoken language, a naive rule is to copy the unlabeled texts as gloss (Zhu et al., 2023; Moryossef et al., 2021). Then, we further apply language-specific rules to different sign languages, respectively. A Chinese spoken language text is first separated at the word level and then one-on-one mapped into the closet glosses based on the lexical similarity. As German sign language texts often include affixes and markers, thus, we perform lemmatization on each word in the text. We leverage the open-source spaCy[1] (Honnibal and Montani, 2017) to obtain the linguistics information. Using the above rule-based annotating system, the monolingual data $\mathcal{X}_U = \{\boldsymbol{x}_U^i\}_{i=1}^N$ is mapped as rule-based synthetic data $\mathcal{D}_{U,r} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_{U,r}^i)\}_{i=1}^N$. We provide a detailed list of rules in Appendix A.

Model-based annotating. While the rule-based heuristic allows high lexical similarity between text and gloss, it cannot capture complicated syntactic divergence between two languages. Therefore, following the self-training structure, we further employ a functional SLG model to predict the pseudo glosses for monolingual data, based on more flexible correspondence learned from training data. Because there is a mutually reinforcing relationship between the translation model and the data it generates. As the model-based synthetic data is generated by the SLG model, it is possible to improve performance by iteratively re-training the SLG model. At the k-th iteration ($k > 1$), based on the best SLG model $f(\hat{\theta}^{k-1})$ in $k-1$ iterations, the monolingual data $\mathcal{X}_U = \{\boldsymbol{x}_U^i\}_{i=1}^N$ is annotated as model-based synthetic data $\mathcal{D}_{U,m}^{k-1} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_{U,m}^{i,k-1})\}_{i=1}^N$.

3.3 Two-Stage Training

The proposed S^3LG iteratively annotates and learns from the synthetic data. As S^3LG is a data-centric framework, we keep the SLG model simple but competitive, which is the vanilla Transformer model (Vaswani et al., 2017). Without loss of generality, we take the k-th iteration as an example to introduce the two-stage training strategy, as shown in Fig. 2.

At the beginning of the k-th iteration, we re-initialize a new SLG model $f(\theta^k)$, where the input text $\boldsymbol{x} = \{x_t\}_{t=1}^{T_x}$ with T_x words is first encoded into a context representation. The decoder generates the target sequence $\boldsymbol{y} = \{y_t\}_{t=1}^{T_y}$ with T_y glosses based on the conditional probability $p(\boldsymbol{y}|\boldsymbol{x};\theta^k)$. Specifically, the conditional probability is formulated as:

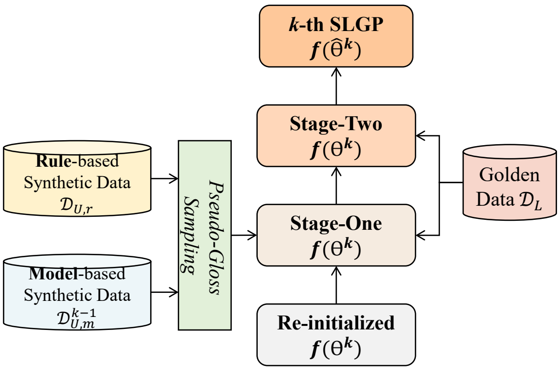

Figure 2: Illustration of the training process in the k-th iteration.

$$p(\boldsymbol{y}|\boldsymbol{x};\theta^k) = \prod_{t=1}^{T_y} p(\boldsymbol{y}_t|\boldsymbol{x}, \boldsymbol{y}_{0:t-1};\theta^k), \qquad (1)$$

where $\boldsymbol{y}_{0:t-1} = \{y_0, \ldots, y_{t-1}\}$ denotes the previous output sub-sequence at the t-th step. The initial token $\boldsymbol{y}_0$ represents the beginning of a sequence.

3.3.1 Pseudo-Glosses Sampling

Once the rule-based $\mathcal{D}_{U,r} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_{U,r}^i)\}_{i=1}^N$ and model-based synthetic data $\mathcal{D}_{U,m}^{k-1} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_{U,m}^{i,k-1})\}_{i=1}^N$ are obtained, we integrate the annotations at the data level to leverage the two complementary synthetic data. Aiming at informing the SLG model that the two auto-annotation methods complement each other, a method-specific token is added at the beginning of the spoken language text. For each unlabeled text $\boldsymbol{x}_U^i$, we randomly select a pseudo gloss $\boldsymbol{y}_U^{i,k-1}$ from the two pseudo glosses $\boldsymbol{y}_{U,r}^i$ and $\boldsymbol{y}_{U,m}^{i,k-1}$ with equal probability. Based on the previous $k-1$ iterations, we obtain the synthetic data $\mathcal{D}_U^{k-1} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_U^{i,k-1})\}_{i=1}^N$ for enlarging the training samples to re-train the SLG model. At the initial iteration, the synthetic data is formulated as $\mathcal{D}_U^0 = \mathcal{D}_{U,r} = \{(\boldsymbol{x}_U^i, \boldsymbol{y}_{U,r}^i)\}_{i=1}^N$.

3.3.2 Training Objective

For optimizing the SLG model $f(\theta^k)$ with both the synthetic data $\mathcal{D}_U^{k-1}$ and golden data $\mathcal{D}_L$, we introduce two kinds of training objective, *i.e.*, cross-entropy loss, and consistency regularization.

Cross-entropy loss. As shown in Equ. 1, the SLG generates the target translation based on the conditional probability provided by the decoder. The

[1] https://spacy.io/

cross-entropy loss computed between the annotation and the output of the decoder, which is formulated as:

$$L_{CE}(\boldsymbol{x}, \boldsymbol{y}, \theta^k) = -log\, p(\boldsymbol{y}|\boldsymbol{x}; \theta^k), \qquad (2)$$

where the $\boldsymbol{y}$ denotes the gloss annotation.

Consistency regularization. The data distribution should be under the manifold assumption, which reflects the local smoothness of the decision boundary (Belkin and Niyogi, 2001). The consistency regularization is computed between two predictions with various perturbations to conform to the manifold assumption. We apply the network dropout as the perturbation. With the dropout strategy, the activated parts of the same model are different during training. The consistency regularization is formulated as:

$$\begin{aligned} L_{CR}(\boldsymbol{x}, \theta^k) =& KL(f(\boldsymbol{x}; \theta_1), f(\boldsymbol{x}; \theta_2)) \\ &+ KL(f(\boldsymbol{x}; \theta_2), f(\boldsymbol{x}; \theta_1)), \end{aligned} \qquad (3)$$

where θ_1 and θ_2 denotes the different sub-models of the SLG $f(\theta^k)$ with dropout during training. $f(\boldsymbol{x}; \theta)$ denotes the predictions given by the SLG model $f(\theta)$. $KL(teacher, student)$ denotes the KL (Kullback-Leibler) divergence loss that aligns the student's network to the teacher's network.

Overall, the loss function of the proposed S^3LG is formulated as:

$$L(\boldsymbol{x}, \boldsymbol{y}, \theta^k) = L_{CE}(\boldsymbol{x}, \boldsymbol{y}, \theta^k) + w \cdot L_{CR}(\boldsymbol{x}, \theta^k), \qquad (4)$$

where the weight w balances the effect of two parts of restraints.

3.3.3 Stage-One and Stage-Two

To alleviate the domain mismatching between the monolingual data $\mathcal{X}_U$ and golden data $\mathcal{D}_L$, the training process is separated into two stages, *i.e.*, pre-training and fine-tuning, which is a conventional way for domain adaption. The SLG model is first trained on the concatenation of large-scale synthetic data $\mathcal{D}_U^{k-1}$ and golden data $\mathcal{D}_L$ with the pre-training epochs T. To amplify the impact of synthetic data, the pre-training epochs gradually increase as the iteration grows. Thus the training objective of stage one is formulated as:

$$\underset{\theta^k}{min} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}_L \cup \mathcal{D}_U^{k-1}} L(\boldsymbol{x}, \boldsymbol{y}, \theta^k). \qquad (5)$$

Subsequently, this model is fine-tuned only on the low-resource in-domain golden data $\mathcal{D}_L$ until convergence. The training objective of stage two is formulated as:

$$\underset{\theta^k}{min} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}_L} L(\boldsymbol{x}, \boldsymbol{y}, \theta^k). \qquad (6)$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our approach on two public sign language translation datasets, *i.e.*, PHOENIX14T (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021). Both datasets provide the original sign language video, sign language gloss, and spoken language text annotated by human sign language translators. We focus on annotated text-gloss parallel data in this work. The PHOENIX14T dataset is collected from the German weather forecasting news on a TV station. The CSL-Daily dataset is a Chinese sign language dataset and covers a wide range of topics in daily conversation.

Monolingual data. Following the previous work (Zhou et al., 2021), we obtain the monolingual spoken language sentences close to the topic of golden data. For the PHOENIX14T and CSL-Daily dataset, we collect $566,682$ and $212,247$ sentences. The statistics of the data mentioned above are shown in Appendix B and C.

Evaluation metrics. Referring to the previous works (Zhou et al., 2021; Li et al., 2022; Zhu et al., 2023), we quantify the performance of the generated gloss in terms of accuracy and consistency based on the BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), respectively. The BLEU-N (N ranges from 1 to 4) is widely used in NMT to show the matching degree of N units between two sequences. Besides, the ROUGE cares more about the fluency degree of generated sequences. Both the evaluation metrics indicate attributes of generated gloss, noting that the higher values demonstrate better translation performance.

Training settings. We implement the proposed approach on Pytorch (Paszke et al., 2017). For PHOENIX14T and CSL-Daily, the SLG model consists of 5 and 3 layers, respectively. To mitigate the overfitting problem, we apply the common strategy such as dropout and label smoothing. For the network setting, the dimensions of the embedding layer and the feed-forward network are 512 and 2048, respectively. The number of the attention head is 8. For the optimization setting, we leverage

	Dev					Test				
	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Stoll et al. (2020)	48.42	50.15	32.47	22.30	16.34	48.10	50.67	32.25	21.54	15.26
Saunders et al. (2020)	55.41	55.65	38.21	27.36	20.23	54.55	55.18	37.10	26.24	19.10
Amin et al. (2021)	-	-	-	-	-	42.96	43.90	26.33	16.16	10.42
Egea Gómez et al. (2021)	-	-	-	-	-	-	-	-	-	13.13
Zhang and Duh (2021)	-	-	-	-	-	-	-	-	-	16.43
Li et al. (2022)	-	-	-	-	-	49.91	-	-	25.51	18.89
Saunders et al. (2022)	57.25	-	-	-	21.93	56.63	-	-	-	20.08
Egea Gómez et al. (2022)	-	-	-	-	-	-	-	-	-	20.57
Walsh et al. (2022)	58.82	60.04	42.85	32.18	25.09	56.55	58.74	40.86	30.24	23.19
Zhu et al. (2023)	-	-	-	-	27.62	-	-	-	-	24.89
Baseline	57.06	58.84	41.07	29.76	22.29	55.28	57.36	38.90	27.80	20.22
S^3LG	**61.60**	**62.36**	**46.30**	**35.63**	**28.24**	**59.62**	**60.67**	**43.69**	**32.91**	**25.70**

Table 1: Performance comparison of our proposed S^3LG with methods for SLG on PHOENIX14T.

	Dev					Test				
	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Li et al. (2022)	-	-	-	-	-	52.78	-	-	29.70	21.30
Baseline	50.26	52.90	32.79	20.94	14.05	50.75	53.23	33.24	21.21	13.92
S^3LG	**61.52**	**65.88**	**47.67**	**36.05**	**27.95**	**61.75**	**65.88**	**47.90**	**36.06**	**27.74**

Table 2: Performance comparison of our proposed S^3LG with methods for SLG on CSL-Daily.

the Adam (Kingma and Ba, 2014). During training, the learning rate and batch size are fixed to 5×10^{-5} and 32, respectively. As at the beginning, the predictions of the SLG model might be unreliable, the consistency regularization weight w ramps up, starting from zero, along a linear curve until reaching the w (Laine and Aila, 2016). Following the previous setting (Li et al., 2019), we randomly shuffle and drop words of the spoken language sentences as data augmentation.

Inference details. In inference, we use the beam search strategy (Wu et al., 2016) to increase the decoding accuracy. For both the CSL-Daily and the PHOENIX14T dataset, the search width and length penalty are set to 3 and 1.0, respectively. In the process of generating the pseudo glosses for monolingual data, we simply set the search width to 1 for efficiency. The experiment is run on an NVIDIA GeForce RTX 3090 with approximately 40 hours of computational time.

4.2 Comparison with State-of-the-Art Methods

We compare the proposed S^3LG with the previous text-to-gloss systems on two public benchmarks, *i.e.*, PHOENIX14T (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021). The performances are shown in Tab. 1 and Tab. 2, respectively.

As our goal is to explore how to incorporate monolingual data for SLG, our baseline adopts the vallia Transformer as the SLG model which only learns from the original golden data. By combining all proposed components, our S^3LG achieves substantial improvements against the baseline across all evaluation metrics. The S^3LG achieves 28.24 and 27.95 BLEU-4 on the DEV set of the PHOENIX14T and CSL-Daily dataset, which surpasses the baseline by 5.95 and 13.9, respectively. The quantitative results demonstrate the effectiveness of utilizing the complementary synthetic data and designs in our S^3LG. For the PHOENIX14T and CSL-Daily datasets, we evaluate the performance using the gloss-level tokenizer as the original annotations, respectively.

Zhu et al. (2023) provides translation performance in different settings, including semi-supervised, transfer learning, and multilingual. For a fair comparison, we cite their best performance in the monolingual and bilingual settings. The results prove the advantage of our novel designs, which distinguishes our approach from previous SLG systems. The previous works mainly tested on the PHOENIX14T datasets, while CSL-Daily is also an important benchmark for different sign language tasks. To attract more research attention on Chinese Sign Language, we report our performance on this dataset.

Setting	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Baseline	57.06	58.84	41.07	29.76	22.29
+ Self-training	58.31	60.90	42.39	30.75	22.97
+ Consistency	59.74	60.14	43.54	32.64	25.25
+ Rule-based	59.86	60.01	44.65	34.25	27.12
+ Aug.+Tag.	**61.60**	**62.39**	**46.30**	**35.63**	**28.24**

Table 3: Effect of our proposed components. 'Self-training' represents directly applying the iterative self-training strategy to the baseline. 'Consistency' denotes adding the additional restraint consistency regularization for the training objective. 'Rule-based' and 'Aug.+Tag.' denote combining the model-based synthetic data with the rule-based synthetic data and marking them with the tagging process, respectively.

w	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	60.22	62.99	45.40	34.34	26.82
5	61.58	**64.10**	**47.08**	35.60	27.75
20	**61.60**	62.36	46.30	**35.63**	**28.24**
40	59.04	57.31	42.43	32.31	25.36

Table 4: Impact of the consistency regularization weight w.

4.3 Ablation Study

To validate the effectiveness of each component proposed in our S^3LG framework, unless otherwise specified, we put forward several ablation studies on the DEV set of the PHOENIX14T dataset.

Impact of proposed components. The main difference between our proposed method and the existing works is to leverage the complementary synthetic as a supplement for the training of the SLG model. To evaluate the effectiveness of each proposed component, we gradually add them to the baseline SLG system. Directly applying the iterative self-training process to the baseline delivers a performance gain of 0.68 BLEU-4, which motivates us to design a more effective algorithm. We further apply consistency regularization to enforce the predictions of synthetic data under the manifold assumption, which achieves a gain of 2.28 BLEU-4. Subsequently, combining the rule-based and model-based synthetic data can be helpful with 1.87 BLEU-4 improvements. Besides, the result shows that tagging and applying data augmentation to different types of synthetic data is also a useful strategy, which provides a further gain of 1.12 BLEU-4. The results are shown in Tab. 3.

Impact of w. In our experiments, the consistency regularization weight w is set to 20. This hyper-parameter determines the importance of the consistency regularization compared with the cross-entropy loss. In Tab. 4, we examine the effect of consistency regularization weight with a set of different values. As the w is set to be 20, S^3LG achieves its best performance.

K	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
0	57.06	58.84	41.07	29.76	22.29
1	60.66	60.94	45.02	34.37	27.12
2	61.16	62.21	45.79	34.95	27.80
3	61.24	62.30	46.19	35.42	28.06
4	**61.60**	**62.36**	**46.30**	**35.63**	**28.24**
5	61.60	62.36	46.30	35.63	28.24

Table 5: Impact of iteration number K. '0' and '1' denote the baseline and only applying the rule-based synthetic data to enlarge the training data, respectively.

Scale	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
0×	57.06	58.84	41.07	29.76	22.29
5×	60.96	61.00	44.90	34.18	26.91
15×	60.94	61.63	45.88	35.27	27.98
>30×	**61.60**	**62.36**	**46.30**	**35.63**	**28.24**

Table 6: Scale of synthetic data. '0×' and '>30×' represent utilizing non-monolingual data and all the monolingual data for the training of the SLG model, respectively.

Quantity	Method	ROUGE	BLEU-2	BLEU-3	BLEU-4
1%	Baseline	17.63	8.31	4.86	3.20
	Ours	**33.44**	**17.29**	**9.05**	**4.69**
5%	Baseline	34.20	19.44	11.16	7.24
	Ours	**45.14**	**26.93**	**17.56**	**12.03**
25%	Baseline	49.44	33.89	22.87	16.47
	Ours	**55.44**	**40.01**	**29.44**	**22.33**

Table 7: Quantity of annotated data.

Impact of iteration number K. The iteration number K is an important hyper-parameter and fixed to 4 in the previous experiments. To explore the effect of K, we conduct experiments with different iteration numbers. Tab. 5 shows the best performance of the K iterations, where $K = 1$ (*i.e.*, the initial iteration), the training data is composed of the rule-based synthetic data and golden data. By combining both the rule-based and model-based synthetic data, the performance of the SLG model is converged at the 4-th iteration.

Scale of synthetic samples. As mentioned above, the collected monolingual data outnumbers the annotated data over 30 times. In the previous experiments, all the monolingual data is incorporated with the golden data to enhance the training pro-

T	Growing	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
0	✗	57.06	58.84	41.07	29.76	22.29
1	✗	59.20	58.73	42.96	32.44	25.32
5	✗	60.52	59.88	44.22	33.80	26.72
10	✗	60.94	**61.18**	45.08	34.44	27.17
15	✗	**60.97**	61.05	**45.32**	**34.77**	27.37
30	✗	60.56	60.59	44.98	34.63	**27.69**
15	✓	61.60	62.36	46.30	35.63	28.24

Table 8: Effect of pre-training epochs T. 'Growing' represents increasing the pre-training epochs between different iterations.

Setting	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Original-Baseline	57.06	58.84	41.07	29.76	22.29
+S^3LP	61.60	62.36	46.30	35.63	28.24
BERT-Baseline	58.26	59.84	42.37	31.21	23.62
+S^3LP	**62.00**	**63.72**	**47.27**	**36.14**	**28.48**

Table 9: Impact of leveraging synthetic data. 'BERT-Baseline' denotes a stronger baseline by enhancing the original baseline with the pre-trained language model, BERT.

cess of the SLG model. In this ablation study, we investigate the scale of synthetic data. As shown in Tab. 6, the performance improves approximately along a log function curve according to the synthetic data volume. We speculate that pre-training on too much noisy out-of-domain synthetic data may drown the impact of golden data, which finally causes limited performance improvements.

Quantity of annotated golden data. As the purpose of our proposed approach is to learn from the monolingual with pseudo glosses, we provide experiments regarding various quantities of synthetic data, in Tab. 7. Under all settings, S^3LG outperforms the baseline. However, the experiment suggests that our proposed approach is more suitable for the scene when the baseline SLG performance is in the range of $7 - 20$ BLEU-4.

Effect of pre-training epochs T. In Tab. 8, we evaluate the impact of pre-training with synthetic data with different epochs. To simplify the hyperparameter search process, we first conduct the experiments with fixed pre-training epochs T between different iterations and further apply the pre-training epochs growing strategy to it. The S^3LG achieves the best performance under the setting of that the pre-training epoch is 15 for the first iteration and then gradually increases by 10 between iterations.

Appearance	≤ 3	≤ 6	≤ 8	≤ 10	≤ 15
Amount	42	49	56	57	66
Baseline	2.38	1.75	1.44	2.24	8.46
Model-Based	2.38	1.75	1.44	3.37	10.00
Model & Rule-based	2.38	**3.50**	**4.34**	**6.74**	**15.38**

Table 10: Translation accuracy of low-frequency glosses. 'Appearance' represents how many times the glosses appear in the TRAIN set of golden parallel data. 'Amount' denotes how many samples contain the low-frequency glosses in the DEV set.

Impact of leveraging synthetic data. To verify whether our performance improvement mainly comes from the synthetic data rather than simply enhancing the encoder with the monolingual data, we conduct experiments by enhancing the original baseline with the pre-trained language model, BERT (Devlin et al., 2019). As the results presented in Tab. 9, we observe that leveraging the pre-trained language model improves the translation quality, while our proposed approach achieves larger performance gains. We further combine our approach with the pre-trained language model. The experimental results demonstrate that the performance improvement of our approach stems from two aspects: better comprehension of the encoder and better generation capability of the decoder.

Results of CHRF metric. To provide more information, following the previous work (Müller et al., 2023), we evaluate the performance of our proposed approach based on CHRF (Popović, 2016) metric. The S^3LP achieves 56.02 and 54.84 CHRF on the DEV and TEST set of PHOENIX14T, which surpasses the baseline (52.00 and 51.32) by 4.02 and 3.52, respectively.

Translation accuracy of low-frequency glosses. As the SLG model tends to predict the glosses with high frequency in training data, we believe that utilizing the model-free annotating approach can mitigate the model-based annotating bias. To verify this, we put forward the experiments for different synthetic data settings with the translation accuracy metric under different low-frequency standards, namely, a gloss appears less than how many times are considered as low frequency, as shown in Tab. 10. The translation accuracy of low-frequency glosses is formulated as $accuracy = N_{pred}/N_{all}$, where N_{pred} and N_{all} denote the number of samples that are predicted with the correct low-frequency glosses and samples that contain the low-frequency glosses, respectively. We can see that the translation accuracy of the SLG model leveraging both the model-based and rule-based synthetic data achieves promising improvements against the model-based one.

5 Conclusion

In this work, we present a semi-supervised framework named S^3LG for translating the spoken language text to sign language gloss. With the goal of incorporating large-scale monolingual spoken language texts into SLG training, we propose the S^3LG approach to iteratively annotate and learn from pseudo glosses. Through a total of K iterations, the final SLG model achieves significant performance improvement against the baseline. During each iteration, the two complementary synthetic data generated from the rule-based and model-based approaches are randomly mixed and marked with a special token. We introduce a consistency regularization to enforce the consistency of the predictions with network perturbations. Extensive ablation studies demonstrate the effectiveness of the above designs. Besides, the translation accuracy on low-frequency glosses is improved.

Limitations

In the hope of attracting more research attention for SLG in the future, we provide the detailed limitations in the next. On the one hand, the results (see Tab. 6 and Tab. 7) of the experiments suggest that in the extremely low-resource scenarios, our performance improvements might be less significant as a large amount of monolingual data is available. We conclude that with the very limited golden data as anchors, it is hard to learn from the large-scale synthetic data. Although utilizing synthetic data as a strong supplement for training data can achieve promising performance improvements, it is hard to achieve the equal impact of enlarging the training data with annotated data from human interpreters. To promote the development of sign language research, the fundamental way might be continuously collecting large amounts of annotated data.

On the other hand, we realize that sign language glosses do not properly represent sign languages (Müller et al., 2023). However, as explained in the introduction section, sign language glosses are sufficient to convey most of the key information in sign language. Given the resource limitations and the current technological capabilities, we believe the two-stage way (*i.e.*, text-to-gloss and gloss-to-gesture) is more achievable and practical to support the development of the application for converting spoken language into sign language. There are many solutions for animating a 3D avatar and making gloss-indexed gestures smoothly and naturally. Correspondingly, research on the former stage lacks enough attention. We think improving the SLG model's performance can be a promising way to implement better sign language production systems.

Ethical Considerations

Even with extensive advances in the development of neural machine translation methods in the spoken language area, the study of sign language is still in its infancy. At the same time, the development of different sign languages is very uneven. According to the existing approaches, the current researches mainly focus on DGS, leaving other sign languages unexplored. As studied in (Bragg et al., 2019; Yin et al., 2021), there is limited research to reveal the sign language linguistics character, which also limits the utility of prior knowledge.

As most sign language researchers are hearing people, the provided sign language system might not meet the actual needs of the deaf. Therefore, to bridge the commutation between the two communities, cooperation could be mutual. By consulting with native signers, we will proactively seek to design the translation system to be inclusive and user-centered in the future. We also encourage people of both communities to try out the existing systems and point out their disadvantages to guide the promising direction and accelerate the study of sign language.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Contract U20A20183 & 62021001, and the Youth Innovation Promotion Association CAS. It was also supported by GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of the USTC.

References

Mohamed Amin, Hesahm Hefny, and Mohammed Ammar. 2021. Sign language gloss translation using deep learning models. *International Journal of Advanced Computer Science and Applications*, 12.

Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022. Using neural machine translation methods for sign language translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–284, Dublin, Ireland. Association for Computational Linguistics.

Mikhail Belkin and Partha Niyogi. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Yong Cao, Wei Li, Xianzhi Li, Min Chen, Guangyong Chen, Long Hu, Zhengdao Li, and Kai Hwang. 2022. Explore more guidance: A task-aware instruction network for sign language translation enhanced with data augmentation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2679–2690, Seattle, United States. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, pages 17043–17056.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

Luis Chiruzzo, Euan McGill, Santiago Egea-Gómez, and Horacio Saggion. 2022. Translating Spanish into Spanish Sign Language: Combining rules and data-driven approaches. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 75–83, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Santiago Egea Gómez, Luis Chiruzzo, Euan McGill, and Horacio Saggion. 2022. Linguistically enhanced text to sign gloss machine translation. In *Proceedings of the International Conference on Applications of Natural Language to Information Systems*, pages 172–183.

Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 18–27, Online (Virtual Mode). INCOMA Ltd.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the International Conference on Machine Learning*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. 7:411–420.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Dongxu Li, Chenchen Xu, Liu Liu, Yiran Zhong, Rong Wang, Lars Petersson, and Hongdong Li. 2022. Transcribing natural languages for the deaf via neural editing programs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11991–11999.

Guanlin Li, Lemao Liu, Guoping Huang, Conghui Zhu, and Tiejun Zhao. 2019. Understanding data augmentation in neural machine translation: Two perspectives towards generalization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5689–5695, Hong Kong, China. Association for Computational Linguistics.

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.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

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.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. 2020. Boosting continuous sign language recognition via cross modality augmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1497–1505.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Proceedings of the European Conference on Computer Vision*, pages 687–705.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 5141–5151.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128:891–908.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Harry Walsh, Ben Saunders, and Richard Bowden. 2022. Changing the representation: Examining language representation for neural sign language production. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 117–124, Marseille, France. European Language Resources Association.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 31:1235–1270.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Xuan Zhang and Kevin Duh. 2021. Approaching sign language gloss translation as a low-resource machine translation task. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 60–70, Virtual. Association for Machine Translation in the Americas.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Rules Used in the Rule-based Heuristic for Creating Synthetic Data $\mathcal{D}_{U,r}$

A.1 Chinese Rules

For a spoken text $\boldsymbol{y}$,

1. Build the vocabulary for the gloss $\mathcal{V}$ and spoken word $\mathcal{S}$ from the original golden data $\mathcal{D}_L$.

2. Tokenize the spoken text at the word-level as $\boldsymbol{y} = \{y_1, y_2, \dots, y_T\}$ with T words.

3. Replace the words $y_t \in \boldsymbol{y}$ not in the word vocabulary $\mathcal{S}$ by a special token <UNK>.

4. Replace each spoken word $y_t \in \boldsymbol{y}$ by the most similar gloss in the gloss vocabulary $\mathcal{V} = \{v_1, v_2, \dots, v_{|\mathcal{V}|}\}$ based on the lexical similarity, which is formulated as:

$$Sim(y_t, v_i) = E(y_t) \cdot E(v_i) \tag{7}$$

where $E(\cdot)$ denotes L_2 normalized word embedding processes. The above information is obtained from the Chinese model (zh_core_web_lg) of spaCy.

A.2 German Rules

For a spoken text $\boldsymbol{y} = \{y_1, y_2, \dots, y_T\}$ with T words,

1. Build the vocabulary for the gloss $\mathcal{V}$ and spoken word $\mathcal{S}$ from the original golden data $\mathcal{D}_L$.

2. Replace the words $y_t \in \boldsymbol{y}$ not in the word vocabulary $\mathcal{S}$ by a special token <UNK>.

3. Lemmatize all the spoken words.

4. Replace the token $y_t \in \boldsymbol{y}$ that only matches parts of compounds glosses $v_i \in \mathcal{V}$ by it v_i.

The above information is obtained from the German model (de_core_news_lg) of spaCy.

B Statistics of Sign Language Datasets

As shown in Tab. 11 and Tab. 12, we present the key statistics of the PHOENIX14T and CSL-Daily dataset, respectively.

	Text			Gloss		
	TRAIN	DEV	TEST	TRAIN	DEV	TEST
Sentence	7,096	519	642	7,096	519	642
Vocabulary	2,887	951	1,001	1,085	393	411
Tot. Words	99.081	6,820	7,816	55,247	3,748	4,264
Tot. OOVs	-	57	60	-	14	19

Table 11: Statistic of the PHOENIX14T dataset.

	Text			Gloss		
	TRAIN	DEV	TEST	TRAIN	DEV	TEST
Sentence	18,401	1,077	1,176	18,401	1,077	1,176
Vocabulary	2,343	1,358	1,358	2,000	1,344	1,345
Tot. Words/Chars	291,048	17,304	19,288	133,714	8,173	9,002
Tot. OOVs	-	64	69	-	0	0

Table 12: Statistic of the CSL-Daily dataset.

Type	Amount	Source
Golden Data	7.096	PHOENIX14T
Monolingual Data	212,247	Wiki, German weather corpus
Golden Data	18,402	CSL-Daily
Monolingual Data	566,682	Wiki, CLUE, Web

Table 13: Statistic of the Training Data.

The PHOENIX14T dataset is about weather forecasting and does not contain any information that names or uniquely identifies individual people or offensive content. The CSL-Daily dataset is screened by its publishing team and is about daily life (shopping, school, travel, etc.). Does not contain any information that names or uniquely identifies individual people or offensive content.

C Statistics of the Training Data

The statistics of training data is shown in shown in Tab. 13. To collect more spoken language texts, we extract a subset of CLUE corpus (Xu et al., 2020) based on the topic of daily lives.

In the process of selecting this part of the data, the words in the PHOENIX14T and CSL-Daily datasets are used to select content on related topics, so most of the data is related to weather and daily life, which does not contain any information that names or uniquely identifies individual people or offensive content.