ConSLT: A Token-level Contrastive Framework for Sign Language Translation

Biao Fu$^{1,2,\dagger}$ Peigen Ye$^{1,2,\dagger}$ Liang Zhang$^{1,2}$ Pei Yu$^{1,2}$ Cong Hu$^{1,2}$
Yidong Chen$^{1,2,\ddagger}$ Xiaodong Shi$^{1,2}$

$^1$Department of Artificial Intelligence, School of Informatics, Xiamen University
$^2$Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism

{biaofu,pgye}@stu.xmu.edu.cn, {ydchen,mandel}@xmu.edu.cn

Abstract

Sign language translation (SLT) is an important technology that can bridge the communication gap between the deaf and the hearing people. SLT task is essentially a low-resource problem due to the scarcity of publicly available parallel data. To this end, inspired by the success of neural machine translation methods based on contrastive learning, we propose ConSLT, a novel token-level contrastive learning framework for Sign Language Translation. Unlike previous contrastive learning based works whose goal is to obtain better sentence representation, ConSLT aims to learn effective token representation by pushing apart tokens from different sentences. Concretely, our model follows the two-stage SLT method. First, in the recognition stage, we use a state-of-the-art continuous sign language recognition model to recognize glosses from sign frames. Then, in the translation stage, we adopt the Transformer framework while introducing contrastive learning. Specifically, we pass each sign glosses to the Transformer model twice to obtain two different hidden layer representations for each token as “positive examples” and randomly sample K tokens that are not in the current sentence from the vocabulary as “negative examples” for each token. Experimental results demonstrate that ConSLT achieves new state-of-the-art performance on PHOENIX14T dataset, with +1.48 BLEU improvements.

1 Introduction

Sign language is a visual language used for daily communication in the deaf community. Sign language translation (SLT) can play an important role in bridging the communication gap between the deaf and the hearing people. SLT has received increasing attention by the research community (Camgoz et al., 2018, 2020a; Zhou et al., 2021a,b; Yin et al., 2021). Recently, Camgoz et al. (2018) contributed the first publicly available SLT dataset, namely PHOENIX14T, which is also the most popular SLT dataset at present. The dataset contains less than 9K parallel sign language videos, sign glosses and spoken language sentences. In their work, the authors formalized the SLT into a Neural Machine Translation (NMT) problem. And they also found that using gloss as the representation of intermediate supervision can significantly improve the performance of the SLT model compared to an end-to-end method. More specifically, SLT can be divided into two stages: 1) using a Continuous Sign Language Recognition (CSLR) model to recognize sign glosses from sign videos, and 2) using a NMT model to translate the glosses into spoken language sentences. Figure 1 depicts two stages of Cascaded SLT. Based on the pipeline approach, Yin and Read (2020) proposed the STMC-Transformer model and achieved the state-of-the-art performance for SLT. STMC-Transformer first recognized glosses by applying the state-of-the-art CSLR model (Zhou et al., 2020) and then translated the glosses into spoken sentences by employing vanilla Transformer (Vaswani et al., 2017) as a NMT model. However, SLT systems cannot leverage large-scale annotation data to train deep model like NMT as the size of SLT dataset is several orders of magnitude smaller than NMT dataset. Table 1 describes the size of

| Corpus | P14T | EN-DE |
|--------|------|-------|
| Sentences | 8,257 | 1,920,209 |
| Words | 72,783 | 53,008,851 | 50,486,398 |

Table 1: Statistics of a SLT dataset and a NMT dataset. P14T denotes the largest publicly available SLT dataset PHOENIX14T. The EN-DE denotes the English-German parallel data from WMT18 (Bojar et al., 2018). The size of SLT dataset is several orders of magnitude smaller than NMT dataset. Therefore, SLT task is essentially a low-resource problem.
Figure 1: Cascaded Sign Language Translation. The sign language glosses are consistent meaning and order with sign gestures. However, the word order of sign glosses is different from spoken language sentences as the grammar of sign and spoken languages are vastly different. The spoken language sentence is more clear and coherent. A continuous sign language recognition model is used to recognize sign glosses from sign videos in the first stage. Then, a NMT model is applied to translate the glosses into spoken language sentences in the second stage.

The significantly important cause for this gap is that the collection and the annotation of sign language dataset is extremely difficult and expensive. Therefore, SLT task is essentially a low-resource problem.

Currently, contrastive learning has become a very popular technique in computer vision and natural language processing community with solid performance (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Wu et al., 2020; Fang et al., 2020; Carlsson et al., 2020; Giorgi et al., 2020; Yan et al., 2021). Recent works (Li et al., 2021; Pan et al., 2021; Qin et al., 2021) also demonstrate that contrastive learning has tremendous potential in low-resource scenarios, especially for NMT. Pan et al. (2021) proposes a multilingual contrastive learning framework mRASP2 for multilingual NMT which achieves significant improvements (10.26 BLEU) on the 30 zero-shot translation directions.

Inspired by the above works, we propose a novel token-level contrastive learning framework to alleviate the low-resource situation for sign language translation, called ConSLT. Unlike previous works whose goal is to obtain better sentence representation, our model aims to learn effective token representation by pushing apart tokens from different sentences. Following the two-stage SLT method, we first use the state-of-the-art CSLR model (Zhou et al., 2020) to recognize glosses from sign videos. Then, we pass each sign gloss to the Transformer model twice to obtain two different hidden layer representations for each token as “positive pairs”. For each token, we randomly sample K tokens that are not in the current sentence from the vocabulary as negative examples. An overview of ConSLT can be seen in Figure 2. We evaluate ConSLT on the challenging PHOENIX14T dataset. The experimental results demonstrate that our method improves the performance of SLT and outperforms previous state-of-the-art methods.

The main contributions of our work are summarized as follows:

- we propose a novel token-level contrastive learning method for sign language translation.
- To the best of our knowledge, we are the first to explore contrastive learning for sign language translation from the perspective of NLP.
- ConSLT achieves state-of-the-art performance in translation accuracy on the challenging public dataset PHOENIX14T.

2 Related Works

2.1 Sign Language Translation

Camgoz et al. (2018) first propose an end-to-end neural SLT model that combines Convolutional Neural Networks (CNNs) and the attention-based sequence-to-sequence model to jointly learn how to align and translate from sign videos to spoken language sentences. Camgoz et al. (2020a) proposes a multi-channel SLT architecture to model the inter and intra contextual relationships between different sign channels. Zhou et al. (2021b) design a spatial-temporal multi-cue network for SL recognition and translation to learn the spatial-temporal collaboration of different visual cues with
an end-to-end manner. Camgoz et al. (2020b) introduce a novel framework based on vanilla Transformer due to the superiority of Transformer in sequence modeling tasks. The framework jointly learns CSLR and SLT by introducing gloss as an intermediate supervision to alleviate the long-term dependency issues. Xie et al. (2021) adopt the Content-aware and Position-aware Temporal Convolution (CPTcn) and the Disentangled Relative Position Encoding (DRPE) to gather neighboring temporal features and inject position information for enhancing the frame-wise sign representation, respectively. Zhou et al. (2021a) propose a two-stage sign back-translation (SignBT) approach, i.e., text-to-gloss and gloss-to-sign, to generate synthetic parallel data as a supplement for SLT training. CDM (Jin and Zhao, 2021) employ contrastive constraints between the video frames and spoken language words for signer-independent SLT. Despite the fact that above works have achieved varied degrees of progress, they all pay attention to the improvement of visual feature module. However, we believe that NLP techniques are crucial to translation quality of SLT. Yin and Read (2020) propose the STMC-Transformer and explore SLT with Transformer on the translation system. Our work also focuses on the translation stage. Unlike the STMC-Transformer, we introduce contrastive loss in the translation stage to learn effective token representations.

2.2 Contrastive Learning

Contrastive learning aims to learn effective representations by contrasting positive pairs and negative pairs. It has been widely used to learn better visual representations for computer vision tasks (Oord et al., 2018; Grill et al., 2020; Zhuang et al., 2019). Chen et al. (2020a) apply a stochastic data augmentation module (cropping, color distortions and Gaussian blur) to generate two correlated views for given data examples. He et al. (2020) elaborately design a momentum update mechanism to build dynamic dictionaries for contrastive learning. Li et al. (2021) propose Prototype Contrastive Learning to implicitly encode the semantic structure of data into the embedding space.

Recently, contrastive learning has also gained significant attention in natural language processing community (Fang et al., 2020; Carlsson et al., 2020; Giorgi et al., 2020; Yan et al., 2021). Wu et al. (2020) carefully design and test different data augmentations and combinations to learn a noise-invariant sentence representation. mRASP2 model (Pan et al., 2021) leverages monolingual data and bilingual data under a unified training framework to close the representation gap of different languages by introducing contrastive learning and aligned augmentation. Gao et al. (2021) apply the dropout (Srivastava et al., 2014) twice acting as minimal data augmentation to obtain two different embeddings as “positive pairs” for each sentence. The above works mainly focus on obtaining better sentence representation. Wei et al. (2021) learn universal representations across languages by conducting instance-wise comparison at both sentence-level and word-level during pre-training. Different from word-level contrastive learning in (Wei et al., 2021) that aims to distinguish words semantically related to sentences for learning universal representations across languages. Our contrastive learning methods pay attention to generate more accurate and fluent tokens by learn token-level representation for sign language translation.

3 Approach

Given a sign video \( F_i = (f_1, \ldots, f_T) \) with \( T \) frames, we first employ a CSLR model to generate its corresponding sign gloss sequence \( G_i = (g_1, \ldots, g_U) \) with \( U \) glosses. Then, a NMT model translates the gloss sequence into spoken language sentence \( S_i = (w_1, \ldots, w_L) \) with \( L \) words. Moreover, \( V = (v_1, \ldots, v_M) \) is the vocabulary of natural spoken language with \( M \) tokens. An overview of our framework for SLT is illustrated in Figure 2.

3.1 Recognition

In the first stage, we apply the state-of-the-art CSLR model STMC (Zhou et al., 2020) to recognize sign glosses from sign videos. The STMC model consists of the spatial multi-cue module, the temporal multi-cue module and sequence learning module. We formulate the process as:

\[
G_i = \Phi(F_i)
\]

where \( \Phi(\cdot) \) denotes STMC network. \( F_i \) represents a sign video sequence and \( G_i \) represents a sign gloss sequence generated by STMC network. We refer to the original STMC paper (Zhou et al., 2020) for more architecture details.

3.2 Translation

In the second stage, we use Transformer to learn a mapping function to translate from a sign gloss
sequence to a spoken language sentence. A little different from vanilla Transformer (Vaswani et al., 2017), we choose a larger setting with a 2-layer encoder and a 2-layer decoder to prevent over-fitting owing to limited training data. The training loss is cross entropy defined as:

$$L_{ce} = \sum_{G_i, S_i \in D} - \log P_\theta (S_i | G_i)$$

(2)

where $D$ denotes all parallel datasets. $(G_i, S_i)$ denote $i$-th parallel sentence pair in $D$. $\theta$ is the parameter of Transformer model.

### 3.3 Token-level Contrastive Learning

Most previous works based on contrastive learning aim to obtain better sentence representations. However, NMT generates each word by conditioning on previously generated words. In other words, NMT is a token level NLP task. Thus, we propose a model based on token-level contrastive learning, which aims to learn effective token representation by pushing apart tokens from different sentences.

**Positive pairs.** Given a gloss-translation pairs $(G_i, S_i)$, we pass $G_i$ to the Transformer model twice. Since the dropout mechanism randomly drops part of units in fully-connected layers and attention probabilities of the Transformer, we can obtain two different hidden layer representations, denoted as $H_i = \{h_j\}_{j=1}^L$ and $H^+_i = \{h^+_j\}_{j=1}^L$, where $H_i, H^+_i \in \mathbb{R}^{L \times d}$. $L$ is the sequence length and $d$ is the hidden size. We construct the “positive pairs” $(h_j, h^+_j)$ from $H_i$ and $H^+_i$ for each token.

**Negative pairs.** We randomly sample $K$ token representations from candidate tokens set $C$ to construct a set of negative samples $Z_j = \{z_k\}_{k=1}^K$ for each token representation $h_j$. $C_i$ denotes all tokens from vocabulary $V$ but not in the current sentence $S_i$, i.e., $C_i \subset V \setminus S_i$. $W$ is the weight matrix of linear output layer, which is an $|V| \times d$ matrix, where $|V|$ denotes the vocabulary size. Each row of $W$ is the $d$-dimension vector representation of token from vocabulary. Thus, we can construct the “negative pairs” $(h_j, z_j)$ for each token.

**Distance Metric.** Previous work used inner product operation or cosine function to calculate similarity of positive and negative pairs. We adopt KL-divergence as the distance metric of contrastive learning as we believe that KL loss is more stringent compared to cosine loss. When the cosine loss of two vectors is 1, KL loss is not necessarily 0. For example, given two vectors $a = [1.0, 2.0, 3.0]$
|          | Dev     |        |        | Test    |        |        |        |
|----------|---------|--------|--------|---------|--------|--------|--------|
| **S2T**  |         | BLEU-1 | BLEU-2 | BLEU-3  | BLEU-4 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
| RNN-SLT  | 31.87   | 19.11  | 13.16  | 9.94    | 32.24  | 19.03  | 12.83  | 9.58    |
| SL-Transf| 46.56   | 34.03  | 26.83  | 22.12   | 47.20  | 34.46  | 26.75  | 21.80   |
| PiSLTRc  | 47.37   | 35.41  | 28.09  | 23.17   | 48.50  | 35.97  | 28.37  | 23.40   |
| SignBT   | **51.11** | **37.90** | **29.80** | **24.45** | 50.80  | 37.75  | 29.72  | 24.32   |
| MCT      | -       | -      | -      | 22.38   | -      | -      | -      | 21.32   |
| STMC-T   | 47.60   | 36.43  | 29.18  | 24.09   | 46.98  | 36.09  | 28.70  | 23.65   |

| **S2G2T** | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|-----------|---------|--------|--------|--------|---------|--------|--------|--------|
| RNN-SLT   | 42.88   | 30.30  | 23.02  | 18.40  | 43.29  | 30.39  | 22.82  | 18.13   |
| SL-Transf | 47.73   | 34.82  | 27.11  | 22.11  | 48.47  | 35.35  | 27.57  | 22.45   |
| SignBT    | 49.33   | 36.43  | 28.66  | 23.51  | 48.55  | 36.13  | 28.47  | 23.51   |
| STMC-Transf | 48.27 | 35.20  | 27.47  | 22.47  | 48.73  | 36.53  | 29.03  | 24.00   |
| ConSLT    | 50.47   | 37.54  | 29.62  | 24.31  | **51.29** | **38.62** | **30.79** | **25.48** |

Table 2: Comparison with methods for SLT on PHOENIX14T. **S2T** denotes the end-to-end SLT model. **S2G2T** denotes the cascaded SLT model. (the higher the better)

and $b = [20.0, 40.0, 60.0]$, then $\cos(a, b) = 1$, but $KL(a, b) = 4.04$. Experimental results in Table 3 show the effectiveness of our method. The objective of contrastive learning is to minimize the following loss:

$$L_{ctr} = -\sum_{g_i \in D} \sum_{h_j \in H_i} \log \frac{e^{\frac{\Omega(h_j, h_j^+)}{\tau}}}{\sum_{k \in K} e^{\frac{\Omega(h_j, z_k)}{\tau}}}$$  (3)

$$\Omega(h_j, h_j^+) = \frac{1}{2} (KL(h_j||h_j^+) + KL(h_j^+||h_j))$$  (4)

where $\tau$ is a temperature hyperparameter. We calculate the bidirectional KL-divergence as the KL-divergence is a non-symmetric measure.

Finally, during the training of ConSLT, the model can be optimized by jointly minimizing the translation loss and contrastive training loss:

$$L = L_{ce} + \alpha L_{ctr}$$  (5)

where $\alpha$ is the coefficient weight to balance the two training losses.

4 Experiments

4.1 Dataset and Metrics

PHOENIX14T (Camgoz et al., 2018) is the current benchmark for SLT, and we evaluate our model on this challenging dataset. PHOENIX14T is a dataset from the German Weather Forecast, a parallel corpus containing sign language videos, gloss annotations, and spoken language translations. The dataset is divided into Train, Dev and Test with sizes 7095, 519 and 642, respectively. The corpus contains sign languages from 9 different signers with 1066 different sign language glosses and 2887 different words translated into spoken German.

To measure the SLT performance, we apply BLEU (Papineni et al., 2002) scores, which are common metrics for NMT. BLEU measures the quality of translation based on precision. And we use BLEU-1,2,3,4 to measure the translation of different phrase levels based the multi-bleu.pl script¹.

4.2 Implementation Details

We follow the STMC-Transformer (Yin and Read, 2020) with PyTorch (Paszke et al., 2019) and OpenNMT (Klein et al., 2017) to implement ConSLT and set hyperparameters.

Network Details: Both encoder and decoder of Transformer have 2 layers. The size of the word embedding, the hidden and the feed-forward are 512, 512 and 2048, respectively. We use 8 attention heads for each layer. And the total parameters count of ConSLT is 17.7M.

Training: We init network parameters with Xavier (Glorot and Bengio, 2010) uniform. And we use a shared weight matrix for the input and output word embeddings in the decoder during training. We adopt the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.998$, and Noam learning rate schedule. Networks are trained with batch size 2,048 and initial learning rate 0.6. The dropout rate and label smoothing (Müller et al., 2019) is set to 0.3 and 0.1, respectively. we follow (Yan et al., 2021) to set the temperature $\tau$ in Eq.3 with

¹https://github.com/moses-smt/mosesdecoder
Table 3: Results of ablation study on the PHOENIX14T test set. (B-n: BLEU-n)

|                      | B-1 | B-2 | B-3 | B-4 |
|----------------------|-----|-----|-----|-----|
| w/o CL               | 48.73 | 36.53 | 29.03 | 24.00 |
| w/ S-CL + cos        | 48.48 | 37.17 | 29.49 | 24.36 |
| w/ T-CL + cos        | 49.91 | 37.60 | 29.88 | 24.59 |
| w/ S-CL + KL         | 50.82 | 38.04 | 30.23 | 25.07 |
| w/ T-CL + KL         | 51.57 | 38.81 | 30.91 | 25.48 |

Table 4: BLEU scores with different sampling strategies of negative samples on the PHOENIX14T test set.

|                      | B-1 | B-2 | B-3 | B-4 |
|----------------------|-----|-----|-----|-----|
| Batch                | 50.69 | 37.93 | 30.09 | 24.79 |
| Vocab                | 51.20 | 38.33 | 30.45 | 25.14 |
| Vocab-Sent           | 51.57 | 38.81 | 30.91 | 25.48 |

0.1. We carry out grid-search of the number of negative samples $K \in \{100, 200, 300, 400, 500\}$ and the $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1, 2.5, 5, 10, 20\}$ of Eq.5 on PHOENIX14T development set. The experimental results are shown in Figure 3 and Figure 5, respectively. And we set $K$ and $\alpha$ to 500 and 0.5 in our experiments, respectively. Note that we evaluate accuracy on the dev dataset every 100 steps and perform early stopping with patience 3. ConSLT requires to be trained on 1 NVIDIA TITAN RTX GPU with 24 GB memory for 4 hours.

**Decoding**: For decoding during the inference process, the generated $<$unk$>$ tokens are replaced by the source token with the highest attention weight. We also employ the beam search strategy with width 6.

### 4.3 Comparison Results

In Table 2, we compare ConSLT with several methods for the SLT on the PHOENIX14T dataset. Our results are averaged over 10 runs with different random seeds. The **RNN-SLT** (Camgoz et al., 2018) only adopt full frame features from Re-sign. **SL-Transf** (Camgoz et al., 2020b) models utilize pre-trained features from CNN-LSTM-HMM and jointly learn sign language recognition and translation. **PiSLTRc** (Xie et al., 2021) uses position-informed temporal convolution base on **SL-Transf**. **SignBT** (Zhou et al., 2021a) uses Sign Back-Translation for data augmentation. **MCT** (Camgoz et al., 2020a) and **STMC-T** (Zhou et al., 2021b) are evaluated under multi-cue setting. **STMC-Transf** (Yin and Read, 2020) uses vanilla Transformer for SL translation with STMC gloss. Our method can substantially improve translation quality for SLT, largely outperforming the previous state-of-the-art. Specifically, compared with the **SignBT**, ConSLT achieves competitive results on the PHOENIX-2014-T dev set as **SignBT** utilizes external monolingual data through sign back translation. To our surprise, ConSLT outperforms SignBT 1.16 BLEU-4 on the PHOENIX-2014-T test set. Moreover, ConSLT gains an improvement of 1.83 BLEU-4 on PHOENIX14 test set compared with the end-to-end spatial-temporal multi-cue network **STMC-T**, 1.48 BLEU-4 compared with **STMC-Transf**. Experiments demonstrates that introducing contrastive learning could bring notable performance gain for SLT.

### 5 Analysis

#### 5.1 Ablation Study

We perform the ablation study to investigate the influence of contrastive learning for SLT. First, following SimCSE (Gao et al., 2021), we train a model with sentence-level contrastive learning framework. More specifically, we pass the sign gloss to the vanilla Transformer (Vaswani et al., 2017) and obtain two sentence representations as “positive pairs”. We treat other sentence representations within a minibatch as negative examples. To make the comparisons fair, we apply cosine function and KL-divergence to calculate the distance of positive and negative pairs for the ablation experiments, respectively. We also set the number of negative samples to 500. For notation, w/o CL means no contrastive learning method, **S-CL** means sentence-level contrastive learning, **T-CL** means our token-level contrastive learning. As
shown in Table 3, we can see that the different contrastive learning methods used in SLT all bring significant improvements. Furthermore, ConSLT gains improvements of +0.41 BLEU scores, compared with sentence-level contrastive learning. This further demonstrates that ConSLT is more suitable for SLT as our approach can obtain fine-grained token representations. As illustrated in Table 3, ConSLT which adopt KL-divergence as the distance metric of contrastive learning is better than cosine function. We speculate about the possible reason is that learning similar representations via KL-divergence is more difficult than via cosine function. Therefore, ConSLT achieves better performance in SLT.

### 5.2 Sampling strategy of negative samples

We further conduct comparative experiments to analyze the effectiveness of sampling strategy of negative samples in contrastive learning. We explore three sampling strategies. For notation, Batch denotes randomly sampling negative samples for each token from a mini-batch. Vocab denotes randomly sampling negative samples for each token from the vocabulary of spoken language sentences. Vocab-Sent denotes randomly sampling negative samples for each token from the vocabulary but not in the current sentence. Table 4 presents the results. Vocab substantially outperforms Batch. This gap is because the tokens in the vocabulary are more diverse than those in a mini-batch. Moreover, Vocab-Sent performs better than other methods. This indicates that the tokens in the same sentence should be semantically similar due to the contextual relationship, which is similar to predicting the masked token based on its contextual information in the masked language model. It also enhances the relationship between the masked token and its context. Therefore, the different tokens from the same sentence cannot be pushed apart. The comparative experiments also demonstrate ConSLT can obtain better token representations for SLT.

### 5.3 The number of negative samples

Different number of negative samples may affect the translation performance, so we further test our model with respect to different number of negative samples. We vary $K$ in $\{100, 200, 300, 400, 500\}$ to see the difference. We report the results in Figure 3, from which we observe that the performance improves steadily when increasing the number of negative samples. The result supports the insight of Chen et al. (2020a) that more negative samples benefit the performance for contrastive learning. Note that when we increase $K$ to 600, the out-of-memory issues on GPU will be triggered.

### 5.4 Visualizing Token Embeddings

We also visualize the token embeddings learned by ConSLT on PHOENIX14T. The visualization is generated employing t-SNE (van der Maaten and Hinton, 2008) where each point represents a token from vocabulary. As shown in Figure 4 (a), we can obvious that the majority of token embeddings are squeezed into a more narrow space, which is consistent with the representation collapse phenomenon proposed by Gao et al. (2018). As a comparison shown in Figure 4 (b), this problem is alleviated by the token-level contrast of hidden representations. Our token-level contrastive learning can produce a more uniform representation space compared with STMC-Transformer to generate more accurate and fluent tokens for sign language translation.
5.5 Qualitative Analysis

We show several examples of different models in the SLT task. As the annotation of PHOENIX14T is in German, we provide its corresponding English translation. Compared with the STMC-Transf model, our method generates spoken language sentences with better accuracy and fluency. In the first three examples in the table 5, comparing the translation of our model with that of STMC-Transf, our model pays more attention to details and sentence fluency. And in the third example, it even translates conjunctions like und (and). In the fourth example, our model translates “am meer” (near the sea) into “an den küsten” (on the coasts), which has the similar meaning. Comparing the translations of the two models, it is obvious that our model translates the whole sentence more completely and smoothly. In the fifth example, “achten oktober” (eighth oktober) is mistranslated by STMC-Transf as “achtundzwanzigsten oktober” (twenty eighth oktober), due to the lack of sufficient training data and sufficient context, so specific numbers are often difficult to translate accurately. While our model can learn better token representations to obtain the correct translation. In the sixth example, the original word appears as “regen” (rain), and our model translates it as “dauerregen” (continuous rain). While STMC-Transf translates it as “sonnenschein” (sunshine), which is wrong. From these examples, we can see that our model generates more complete sentences and fewer under-translation sentences. Our model can achieve smoother and more accurate translations.

6 Conclusion

In this work, we propose ConSLT, a contrastive learning framework for SLT, which aims to learn effective token representations by pushing apart tokens from different sentences. We also explore the impact of different contrastive strategies and provide fine-grained analysis for interpreting how our approach works. The experimental results demonstrate that contrastive learning can significantly improve translation quality for SLT. We hope our work will provide a new perspective for future researches on SLT. As to the limitations of this work, ConSLT depends on generating an intermediate gloss sequence to perform the SLT task. However, gloss sequences are not always available due to the expensive annotation and difficult collection. In
future work, we will continue to study end-to-end SLT from the perspective of NLP.

References

Ondřej Bojar, Christian Federmann, Mark Fischel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Necati Cihan Camgoz, 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 Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Multi-channel transformers for multi-articulatory sign language translation. In European Conference on Computer Vision, pages 301–319. Springer.

Necati Cihan Camgoz, 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.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Helliqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In International Conference on Learning Representations.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 1597–1607. PMLR.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. In Advances in Neural Information Processing Systems, volume 33, pages 22243–22255. Curran Associates, Inc.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. arXiv preprint arXiv:2005.12766.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2018. Representation degeneration problem in training natural language generation models. In International Conference on Learning Representations.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821.

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. arXiv preprint arXiv:2006.03659.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, volume 9 of Proceedings of Machine Learning Research, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Geshlاغhi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738.

Tao Jin and Zhou Zhao. 2021. Contrastive Disentangled Meta-Learning for Signer-Independent Sign Language Translation, page 5065–5073. Association for Computing Machinery, New York, NY, USA.

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

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72. Vancouver, Canada. Association for Computational Linguistics.

Junnan Li, Pan Zhou, Caimei Xiong, and Steven Hoi. 2021. Prototypical contrastive learning of unsupervised representations. In International Conference on Learning Representations.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In Proceedings of the 59th Annual Meeting of the Association
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, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. 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 3350–3363, Online. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. Journal of Machine Learning Research, 9(86):2579–2605.

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 Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 6000–6010.

Xiangpeng Wei, Rongxiang Wang, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In International Conference on Learning Representations.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. arXiv preprint arXiv:2012.15466.

Pan Xie, Mengyi Zhao, and Xiaohui Hu. 2021. Pisltcr: Position-informed sign language transformer with content-aware convolution. IEEE Transactions on Multimedia, pages 1–1.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. 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 5065–5075, Online. Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malike 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.

Kayo Yin and Jesse Read. 2020. 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.

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

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2021b. Spatial-temporal multi-cue network for sign language recognition and translation. IEEE Transactions on Multimedia, pages 1–1.

Chengxu Zhuang, Alex Zhai, and Daniel Yamins. 2019. Local aggregation for unsupervised learning of visual embeddings. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6001–6011.

A Effect of Weight $\alpha$

We further investigate the effect of the contrastive learning loss weight $\alpha$. As mentioned in Section 4.2, we set $\alpha = 0.5$ for our experiments. Here we vary the $\alpha$ in $\{0.1, 0.3, 0.5, 0.7, 1, 2.5, 5, 10, 20\}$ and conduct experiments. As shown in Figure 5, either too small or too large $\alpha$ will make our model perform badly. When $\alpha = 0.5$ the model has the best performance. We select 0.5 as the contrastive learning loss weight in most of our experiments.
Figure 5: The effect of weight $\alpha$ on the PHOENIX14T test set.