Better Sign Language Translation with STMC-Transformer

Kayo Yin*
Language Technologies Institute
Carnegie Mellon University
kayo@cmu.edu

Jesse Read
LIX, Ecole Polytechnique
Institut Polytechnique de Paris
jesse.read@polytechnique.edu

Abstract

Sign Language Translation (SLT) first uses a Sign Language Recognition (SLR) system to extract sign language glosses from videos. Then, a translation system generates spoken language translations from the sign language glosses. This paper focuses on the translation system and introduces the STMC-Transformer which improves on the current state-of-the-art by over 5 and 7 BLEU respectively on gloss-to-text and video-to-text translation of the PHOENIX-Weather 2014T dataset. On the ASLG-PC12 corpus, we report an increase of over 16 BLEU.

We also demonstrate the problem in current methods that rely on gloss supervision. The video-to-text translation of our STMC-Transformer outperforms translation of GT glosses. This contradicts previous claims that GT gloss translation acts as an upper bound for SLT performance and reveals that glosses are an inefficient representation of sign language. For future SLT research, we therefore suggest an end-to-end training of the recognition and translation models, or using a different sign language annotation scheme.

1 Introduction

Communication holds a central position in our daily lives and social interactions. Yet, in a predominantly aural society, sign language users are often deprived of effective communication. Deaf people face daily issues of social isolation and miscommunication to this day (Souza et al., 2017). This paper is motivated to provide assistive technology that allow Deaf people to communicate in their own language.

In general, sign languages developed independently of spoken language and do not share the grammar of their spoken counterparts (Stokoe, 1960). For this, Sign Language Recognition (SLR) systems on their own cannot capture the underlying grammar and complexities of sign language, and Sign Language Translation (SLT) faces the additional challenge of taking into account the unique linguistic features during translation.

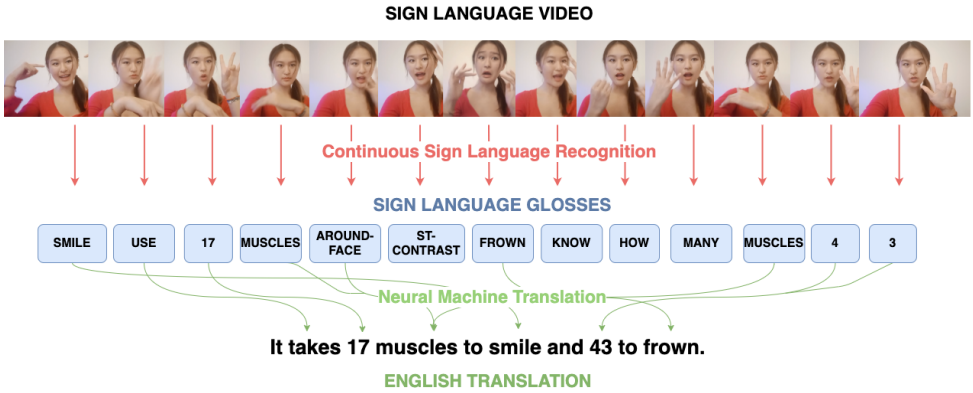

Figure 1: Sign language translation pipeline[1].

*Work carried out while at École Polytechnique.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5975–5989
Barcelona, Spain (Online), December 8-13, 2020

As shown in Figure 1, current SLT approaches involve two steps. First, a tokenization system generates glosses from sign language videos. Then, a translation system translates the recognized glosses into spoken language. Recent work (Orbay and Akarun, 2020; Zhou et al., 2020) has addressed the first step, but there has been none improving the translation system. This paper aims to fill this research gap by leveraging recent success in Neural Machine Translation (NMT), namely Transformers.

Another limit to current SLT models is that they use glosses as an intermediate representation of sign language. We show that having a perfect continuous SLR system will not necessarily improve SLT results. We introduce the STMC-Transformer model performing video-to-text translation that surpasses translation of ground truth glosses, which reveals that glosses are a flawed representation of sign language.

The contributions of this paper can be summarized as:

1. A novel STMC-Transformer model for video-to-text translation surpassing GT glosses translation contrary to previous assumptions

2. The first successful application of Transformers to SLT achieving state-of-the-art results in both gloss to text and video to text translation on PHOENIX-Weather 2014T and ASLG-PC12 datasets

3. The first usage of weight tying, transfer learning, and ensemble learning in SLT and a comprehensive series of baseline results with Transformers to underpin future research

2 Methods

Despite considerable advancements made in machine translation (MT) between spoken languages, sign language processing falls behind for many reasons. Unlike spoken language, sign language is a multidimensional form of communication that relies on both manual and non-manual cues which presents additional computer vision challenges (Asteriadis et al., 2012). These cues may occur simultaneously whereas spoken language follows a linear pattern where words are processed one at a time. Signs also vary in both space and time and the number of video frames associated to a single sign is not fixed either.

2.1 Sign Language Glossing

Glossing corresponds to transcribing sign language word-for-word by means of another written language. Glosses differ from translation as they merely indicate what each part in a sign language sentence mean, but do not form an appropriate sentence in the spoken language. While various sign language corpus projects have provided different guidelines for gloss annotation (Crasborn et al., 2007; Johnston, 2013), there is no universal standard which hinders the easy exchange of data between projects and consistency between different sign language corpora. Gloss annotations are also an imprecise representation of sign language and can lead to an information bottleneck when representing the multi-channel sign language by a single-dimensional stream of glosses.

2.2 Sign Language Recognition

SLR consists of identifying isolated single signs from videos. Continuous sign language recognition (CSLR) is a relatively more challenging task that identifies a sequence of running glosses from a running video. Works in SLR and CSLR, however, only perform visual recognition and ignore the underlying linguistic features of sign language.

2.3 Sign Language Translation

As illustrated in Figure 1, the SLT system takes CSLR as a first step to tokenize the input video into glosses. Then, an additional step translates the glosses into a valid sentence in the target language. SLT is novel and difficult compared to other translation problems because it involves two steps: extract meaningful features from a video of a multi-cue language accurately then generate translations from an intermediate gloss representation, instead of translation from the source language directly.

[1]Gloss annotation from `https://www.handspeak.com/translate/index.php?id=288`

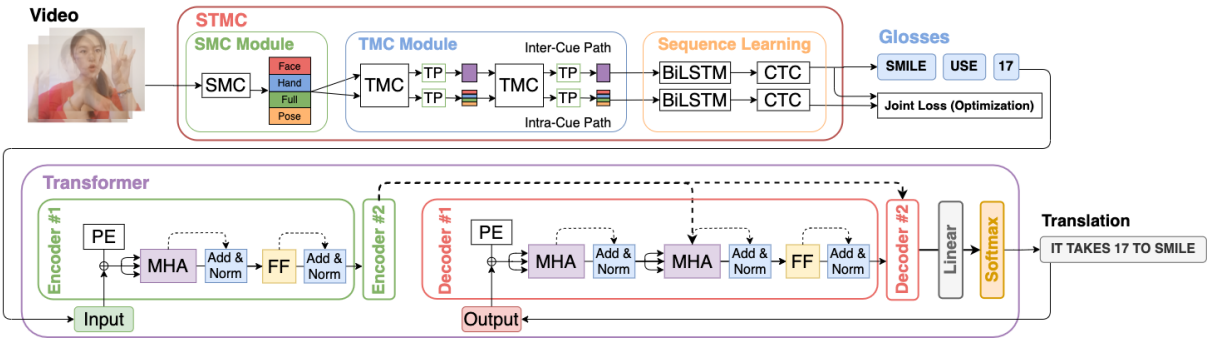

Figure 2: STMC-Transformer network for SLT. PE: Positional Encoding, MHA: Multihead Attention, FF: Feed Forward.

3 Related Work

3.1 Sign Language Recognition

Early approaches for SLR rely on hand-crafted features (Tharwat et al., 2014; Yang, 2010) and use Hidden Markov Models (Forster et al., 2013) or Dynamic Time Warping (Lichtenauer et al., 2008) to model sequential dependencies. More recently, 2D convolutional neural networks (2D-CNN) and 3D convolutional neural networks (3D-CNN) effectively model spatio-temporal representations from sign language videos (Cui et al., 2017; Molchanov et al., 2016).

Most existing work on CSLR divides the task into three sub-tasks: alignment learning, single-gloss SLR, and sequence construction (Koller et al., 2017; Zhang et al., 2014) while others perform the task in an end-to-end fashion using deep learning (Huang et al., 2015; Camgoz et al., 2017).

3.2 Sign Language Translation

SLT was formalized in Camgoz et al. (2018) where they introduce the PHOENIX-Weather 2014T dataset and jointly use a 2D-CNN model to extract gloss-level features from video frames, and a seq2seq model to perform German sign language translation. Subsequent works on this dataset (Orbay and Akarun, 2020; Zhou et al., 2020) all focus on improving the CSLR component in SLT. A contemporaneous paper (Camgoz et al., 2020) also obtains encouraging results with multi-task Transformers for both tokenization and translation, however their CSLR performance is sub-optimal, with a higher Word Error Rate than baseline models.

Similar work has been done on Korean sign language by Ko et al. (2019) where they estimate human keypoints to extract glosses, then use seq2seq models for translation. Arvanitis et al. (2019) use seq2seq models to translate ASL glosses of the ASLG-PC12 dataset (Othman and Jemni, 2012).

3.3 Neural Machine Translation

Neural Machine Translation (NMT) employs neural networks to carry out automated text translation. Recent methods typically use an encoder-decoder architecture, also known as seq2seq models.

Earlier approaches use recurrent (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014) and convolutional networks (Kalchbrenner et al., 2016; Gehring et al., 2017) for the encoder and the decoder. However, standard seq2seq networks are unable to model long-term dependencies in large input sentences without causing an information bottleneck. To address this issue, recent works use attention mechanisms (Bahdanau et al., 2015; Luong et al., 2015) that calculates context-dependent alignment scores between encoder and decoder hidden states. Vaswani et al. (2017) introduces the Transformer, a seq2seq model relying on self-attention that obtains state-of-the-art results in NMT.

4 Model architecture

For translation from videos to text, we propose the STMC-Transformer network illustrated in Figure 2.

4.1 Spatial-Temporal Multi-Cue (STMC) Network

Our work is the first to use STMC networks (Zhou et al., 2020) for SLT. A spatial multi-cue (SMC) module with a self-contained pose estimation branch decomposes the input video into spatial features of multiple visual cues (face, hand, full-frame and pose). Then, a temporal multi-cue (TMC) module with stacked TMC blocks and temporal pooling (TP) layers calculates temporal correlations within (inter-cue) and between cues (intra-cue) at different time steps, which preserves each unique cue while exploring their relation at the same time. The inter-cue and intra-cue features are each analyzed by Bi-directional Long Short-Term Memory (BiLSTM) (Sutskever et al., 2014) and Connectionist Temporal Classification (CTC) (Graves et al., 2006) units for sequence learning and inference.

This architecture efficiently processes multiple visual cues from sign language video in collaboration with each other, and achieves state-of-the-art performance on three SLR benchmarks. On the PHOENIX-Weather 2014T dataset, it achieves a Word Error Rate of 21.0 for the SLR task.

4.2 Transformer

For translation, we train a two-layered Transformer to maximize the log-likelihood

$$\sum_{(x_i,y_i)\in D} \log P(y_i|x_i,\theta)$$

where D contains gloss-text pairs (x_i, y_i).

Two layers, compared to six in most spoken language translation, is empirically shown to be optimal in Section 6.1, likely because our datasets are limited in size. We refer to the original Transformer paper (Vaswani et al., 2017) for more architecture details.

5 Datasets

	German Sign Gloss			German			American Sign Gloss			English		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Phrases	7,096	519	642	7,096	519	642	82,709	4,000	1,000	82,709	4,000	1,000
Vocab.	1,066	393	411	2,887	951	1,001	15782	4,323	2,150	21,600	5,634	2,609
tot. words	67,781	3,745	4,257	99,081	6,820	7,816	862,046	41,030	10,503	975,942	46,637	11,953
tot. OOVs	–	19	22	–	57	60	–	255	83	–	369	99
singletons	337	–	–	1,077	–	–	6,133	–	–	8,542	–	–

Table 1: Statistics of the RWTH-PHOENIX-Weather 2014T and ASLG-PC12 datasets. Out-of-vocabulary (OOV) words are those that appear in the development and testing sets, but not in the training set. Singletons are words that appear only once during training.

PHOENIX-Weather 2014T (Camgoz et al., 2018)

This dataset is extracted from weather forecast airings of the German tv station PHOENIX. This dataset consists of a parallel corpus of German sign language videos from 9 different signers, gloss-level annotations with a vocabulary of 1,066 different signs and translations into German spoken language with a vocabulary of 2,887 different words. It contains 7,096 training pairs, 519 development and 642 test pairs.

ASLG-PC12 (Othman and Jemni, 2012)

This dataset is constructed from English data of Project Gutenberg that has been transformed into ASL glosses following a rule-based approach. This corpus with 87,709 training pairs allows us to evaluate Transformers on a larger dataset, where deep learning models usually require lots of data. It also allows us to compare performance across different sign languages. However, the data is limited since it does not contain sign language videos, and is less complex due to being created semi-automatically. We make our data and code publicly available[2].

[2] https://github.com/kayoyin/transformer-slt

6 Experiments and Discussions

Our models are built using PyTorch (Paszke et al., 2019) and Open-NMT (Klein et al., 2017). We configure Transformers with word embedding size 512, gloss level tokenization, sinusoidal positional encoding, 2,048 hidden units and 8 heads. For optimization, we use Adam (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.998$, Noam learning rate schedule, 0.1 dropout, and 0.1 label smoothing.

We evaluate on the dev set each half-epoch and employ early stopping with patience 5. During decoding, generated $\langle unk \rangle$ tokens are replaced by the source token having the highest attention weight. This is useful when $\langle unk \rangle$ symbols correspond to proper nouns that can be directly transposed between languages (Klein et al., 2017). We perform a series of experiments to find the optimal setup for this novel application. We equally experiment with various techniques often used in classic NMT to SLT such as transfer learning, weight tying and ensembling to improve model performance.

For evaluation we use BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005). For BLEU, we report BLEU-1,2,3,4 scores and as ROUGE score we report the ROUGE-L F1 score. These metrics allow us to directly compare directly with previous works. METEOR is calculated in addition as it demonstrates higher correlation with human evaluation than BLEU on several MT tasks. All reported results unless otherwise specified are averaged over 10 runs with different random seeds.

We organize our experiments into two groups:

1. Gloss2Text (G2T) in which we translate GT gloss annotations to simulate perfect tokenization on both PHOENIX-Weather 2014T and ASLG-PC12

2. Sign2Gloss2Text (S2G2T) where we perform video-to-text translation on PHOENIX-Weather 2014T with the STMC-Transformer

6.1 Gloss2Text (G2T)

G2T is a text-to-text translation task that is novel and challenging compared to classic translation tasks between spoken languages because of the high linguistic variance between source and target sentences, scarcity of resources, and information loss or imprecision in the source sentence itself.

For ASLG-PC12, many ASL glosses are English words with an added prefix so during data preprocessing we remove all such prefixes. We also set all glosses that appear less than 5 times during training as $\langle unk \rangle$ to reduce vocabulary size.

	Raw data						Preprocessed data					
	Train		Dev		Test		Train		Dev		Test	
	ASL	en	ASL	en	ASL	en	ASL	en	ASL	en	ASL	en
Vocab.	15,782	21,600	4,323	5,634	2,150	2,609	5,906	7,712	1,163	1,254	394	379
Shared vocab.	10,048		2,652		1,296		4,941		899		287	
BLEU-4	20.97		21.16		20.63		38.87		38.74		38.37	

Table 2: Statistics of the ASLG-PC12 dataset before and after preprocessing.

Table 2 shows that the source and target corpora in ASLG-PC12 are more similar to each other with many shared vocabulary and a relatively high BLEU-4 score on raw data. This allows us to compare Transformer performance on a larger and less challenging dataset.

Model size

The original Transformer in (Vaswani et al., 2017) uses 6 layers for the encoder and decoder for NMT. However, our task differs from a standard MT task between two spoken languages so we first train Transformers with 1, 2, 4 and 6 encoder-decoder layers. Networks are trained with batch size 2,048 and initial learning rate 1.

	Dev Set						Test Set					
Layers	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
1	43.39	32.47	24.27	20.26	44.66	42.64	43.26	32.23	25.59	21.31	45.28	42.56
2	**45.31**	33.65	**26.73**	**22.23**	45.74	**43.92**	**44.57**	**33.08**	**26.14**	**21.65**	40.47	**42.97**
4	44.32	**32.87**	26.15	21.78	**45.86**	43.31	44.10	32.82	25.99	21.57	**45.44**	42.92
6	44.04	32.46	25.67	21.34	44.09	42.32	43.74	32.44	25.67	21.32	41.69	42.58

Table 3: G2T performance comparison of Transformers on PHOENIX-Weather 2014T with different number of enc-dec layers.

To choose the best model, we mainly take into account BLEU-4 as it is currently the most widely used metric in MT. We do find that our final model outperforms the other models across all metrics. Table 3 shows that on PHOENIX-Weather 2014T, using 2 layers obtains the highest BLEU-4. Because our dataset is much smaller than spoken language datasets, larger networks may be disadvantaged. Moreover, a smaller model has the advantage of taking up less memory and computation time. Repeating the same experiment on ASLG-PC12, we also find 2 layers to be the optimal model size. ASLG-PC12 is larger but less complex which may also explain why smaller networks are more suitable. We carry out the rest of our experiments using 2 enc-dec layers.

Embedding schemes

Press and Wolf (2017) shows that tying the input and output embeddings while training language models may provide better performance. Our decoder is in fact a language model conditioned on the encoding of the source sentence and previous outputs, we can tie the decoder embeddings by using a shared weight matrix for the input and output word embeddings.

In addition, models are often initialized with pre-trained embeddings for transfer learning. These embeddings are typically trained in an unsupervised manner on a large corpus of text in the desired language. We perform experiments on PHOENIX-Weather 2014T using two popular word embeddings: GloVe[3] (Pennington et al., 2014), and fastText (Bojanowski et al., 2017). To the best of our knowledge, weight-tying or pre-trained embeddings have never been employed in SLT.

	GloVe (de)	fastText (de)	GloVe (en)	fastText (en)
Dimension	300	300	300	300
Source match	0.08%	0.08%	96.23%	94.64%
Target match	90.53%	94.57%	97.71%	96.32%

Table 4: German and English pre-trained embeddings statistics

Table 4 shows there is only one matching token between German glosses and the pre-trained embeddings, while over 90% of the words in the German text appear in both pre-trained embeddings. We therefore initialize pre-trained embeddings on the decoder only, and keep random initialization for the encoder. The embedding layers are fine-tuned during training.

	Dev Set						Test Set					
WE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Vanilla embedding	45.81	34.06	**27.05**	**22.49**	**46.68**	**44.35**	**45.29**	**33.74**	**26.70**	**22.22**	**46.08**	43.75
Tied decoder	**45.90**	**34.10**	26.98	22.31	46.76	44.51	45.05	33.38	26.31	21.74	45.83	43.45
GloVe	44.37	32.65	26.00	21.41	45.03	42.38	44.69	32.93	25.73	21.04	42.70	**44.61**
fastText	44.91	33.23	26.60	22.04	46.17	43.70	44.21	32.90	25.94	21.64	45.55	42.95

Table 5: G2T performance comparison using different embedding schemes on PHOENIX-Weather 2014T.

Table 5 shows that the new embedding schemes do not improve performance on PHOENIX-Weather 2014T. It may be because pre-trained embeddings are shown to be more effective when used on the encoding layer (Qi et al., 2018). Another possible reason is the difference between the domain of our dataset and of the corpus the embeddings were trained on. We therefore keep random initialization of

[3] https://deepset.ai/german-word-embeddings

word embeddings for experiments on PHOENIX-Weather 2014T. Using this setting, we run a parameter search over the learning rate and warm-up steps, and we use initial learning rate 0.5 with 3,000 warm-up steps for the remaining experiments. Details of the parameter search are included in Appendix A.1.

Both GloVe and fastText English vectors have a reasonable overlap with the vocabulary of ASL glosses as well as the English targets (Table 4). Therefore on ASLG-PC12 we load pre-trained embeddings on only the decoder, as well as on both the encoder and decoder.

	Dev Set						Test Set					
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Vanilla embedding	90.15	84.92	80.27	75.94	94.72	94.58	90.49	85.64	81.31	77.33	94.75	95.16
Tied dec	**91.00**	**86.26**	**82.00**	**78.02**	**95.24**	**95.12**	**91.25**	**86.76**	**82.76**	**79.02**	**95.32**	**95.75**
GloVe dec	90.13	85.14	80.67	76.49	94.16	94.69	90.51	85.83	81.65	77.74	94.80	95.27
GloVe enc-dec	89.65	84.33	79.72	75.48	93.02	93.62	90.01	85.15	80.88	76.95	93.00	94.14
fastText dec	90.64	85.63	81.14	76.94	94.75	95.02	91.20	86.62	82.53	78.72	94.73	95.57
fastText enc-dec	90.02	85.01	80.56	76.41	93.68	94.10	90.94	86.58	82.01	76.23	93.61	94.42
fastText tied dec	90.16	85.26	80.85	76.72	95.03	94.60	90.44	85.25	81.69	77.28	95.11	95.04

Table 6: G2T performance comparison using different embedding schemes on ASLG-PC12.

Table 6 shows that fastText pre-trained embeddings for the decoder improves performance, and tied decoder embeddings with random initialization gives the best performance. Weight tying is more suited on this dataset likely because it acts as regularization and combats overfitting, while the previous dataset is more complex and therefore less prone to overfitting. For the remaining experiments, we use tied decoder embeddings, initial learning rate 0.2 and 8,000 warm-up steps.

Beam width

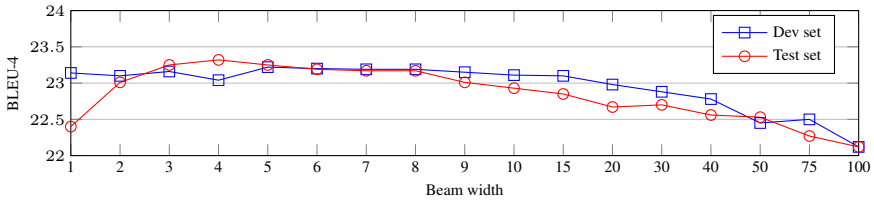

Figure 3: G2T decoding on RWTH-PHEONIX-WEATHER 2014T using different beam width.

A naive method for decoding is greedy search, where the model simply chooses the word with the highest probability at each time step. However, this approach may become sub-optimal in the context of the entire sequence. Beam search addresses this by expanding all possible candidates at each time step and keeping a number of most likely sequences, or the beam width. Large beam widths do not always result in better performance and take more space in memory and decoding time. We search and find the optimal beam width value to be 4 on PHOENIX-Weather 2014T and 5 on ASLG-PC12.

Ensemble decoding

Ensemble methods combine multiple models to improve performance. We propose ensemble decoding, where we combine the output of different models by averaging their prediction distributions. We chose 9 models from our experiments that gave the highest BLEU-4 during testing on PHOENIX-Weather 2014T. The number of models is chosen empirically, as using fewer models will lead to less ensembling but too many weaker models may lessen the quality of the ensemble model. These models are of the same architecture, but are initialized with different seeds and trained using different batch sizes and/or learning rates. These models give a BLEU-4 on testing between 22.92 and 23.41 individually.

	Dev Set						Test Set					
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Raw data	13.01	6.23	3.03	1.71	24.23	13.69	11.88	5.05	2.41	1.36	22.81	12.12
RNN Seq2seq (Camgoz et al., 2018)	44.40	31.93	24.61	20.16	46.02	–	44.13	31.47	23.89	19.26	45.45	–
Transformer (Camgoz et al., 2020)	**50.69**	**38.16**	**30.53**	**25.35**	–	–	**48.90**	36.88	29.45	24.54	–	–
Transformer	49.05	36.20	28.53	23.52	47.36	46.09	47.69	35.52	28.17	23.32	46.58	44.85
Transformer Ens.	48.85	36.62	29.23	24.38	**49.01**	**46.96**	48.40	**36.90**	**29.70**	**24.90**	**48.51**	**46.24**

Table 7: G2T on PHEONIX-WEATHER 2014T final results.

Table 7 gives a performance comparison on PHOENIX-Weather 2014T of the recurrent seq2seq model by Camgoz et al. (2018), Transformer trained concurrently by Camgoz et al. (2020), our single model, and ensemble model. We also provide the scores on the gloss annotations to illustrate the difficulty of this task.

Without any additional training, ensembling improves testing performance by over 1 BLEU-4. Also, we report an improvement of over 5 BLEU-4 on the state-of-the-art. A single Transformer also gives an improvement of over 4 BLEU-4 more than the state-of-the-art, which shows the advantage of Transformers for SLT, as shown also in Camgoz et al. (2020).

	Dev Set						Test Set					
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Raw data	54.60	39.67	28.92	21.16	76.11	61.25	54.19	39.26	28.44	20.63	75.59	61.65
Preprocessed data	69.25	56.83	46.94	38.74	83.80	78.75	68.82	56.36	46.53	38.37	83.28	79.06
Seq2seq (Arvanitis et al., 2019)	–	–	–	–	–	–	86.7	79.5	73.2	65.9	–	–
Transformer	**92.98**	**89.09**	83.55	**85.63**	82.41	95.93	92.98	**89.09**	85.63	82.41	95.87	**96.46**
Transformer Ens.	92.67	88.72	**85.22**	81.93	**96.18**	**95.95**	**92.88**	89.22	**85.95**	**82.87**	**96.22**	96.60

Table 8: G2T on ASLG-PC12 final results

We also use 5 of the best models from our experiments on ASLG-PC12 in an ensemble. Individually, these models obtain between 81.72 and 82.41 BLEU-4 on testing. Table 8 shows that the performance of our single Transformer surpasses the recurrent seq2seq model by Arvanitis et al. (2019) by over 16 BLEU-4. The ensemble model reports an improvement of 0.46 BLEU-4 over the single model. There is relatively less increase from ensembling possibly because there is less variance across different models.

6.2 German Sign2Gloss2Text (S2G2T)

In S2G2T, both gloss recognition from videos and its translation to text are performed automatically. Camgoz et al. (2018) claims the previous G2T setup to be an upper bound for translation performance, since it simulates having a perfect recognition system. However, this claim assumes that the ground truth gloss annotations give a full understanding of sign language, which ignores the information bottleneck in glosses. Camgoz et al. (2020) hypothesizes that it is therefore possible to surpass G2T performance without using GT glosses, which we confirm in this section.

We perform experiments on the PHOENIX-Weather 2014T dataset as it contains parallel video, gloss and text data. On the other hand, the ASLG-PC12 corpus does not have sign language video information.

	Dev Set						Test Set					
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
G2T (Camgoz et al., 2018)	44.40	31.93	24.61	20.16	46.02	–	44.13	31.47	23.89	19.26	45.45	–
S2G → G2T (Camgoz et al., 2018)	41.08	29.10	22.16	17.86	43.76	–	41.54	29.52	22.24	17.79	43.45	–
S2G2T (Camgoz et al., 2018)	42.88	30.30	23.03	18.40	44.14	–	43.29	30.39	22.82	18.13	43.80	–
Sign2(Gloss+Text) (Camgoz et al., 2020)	47.26	34.40	27.05	22.38	–	–	46.61	33.73	26.19	21.32	–	–
S2G → G2T	46.75	34.99	27.79	23.06	47.29	45.23	47.49	35.89	28.62	23.77	47.32	45.54
Bahdanau	45.89	32.24	24.93	20.52	44.46	43.48	47.53	33.82	26.07	21.54	45.50	44.87
Luong	45.61	32.54	26.33	21.00	46.19	44.93	47.08	33.93	26.31	21.75	45.66	44.84
Transformer	48.27	35.20	27.47	22.47	46.31	44.95	48.73	36.53	29.03	24.00	46.77	45.78
Transformer Ens.	**50.31**	**37.60**	**29.81**	**24.68**	**48.70**	**47.45**	**50.63**	**38.36**	**30.58**	**25.40**	**48.78**	**47.60**

Table 9: SLT performance using STMC for CSLR. The first set of rows correspond to the current state-of-the-arts included for comparison.

S2G →G2T

To begin, we use the best performing model for German G2T to translate glosses predicted by a trained STMC network. In Table 9 we can see that despite no additional training for translation, this model already obtains a relatively high score that beats the current state-of-the-art by over 5 BLEU-4.

Recurrent seq2seq networks

For comparison, we also train and evaluate STMC used with recurrent seq2seq networks for translation. The translation models are composed of four stacked layers of Gated Recurrent Units (GRU) (Chung et al., 2014), with either Luong (Luong et al., 2015) or Bahdanau (Bahdanau et al., 2015) attention.

In Table 9, recurrent seq2seq models obtain slightly better performance with Luong attention. Surprisingly, these models outperform previous models of similar architecture that translate GT glosses.

Transformer

For the STMC-Transformer, we train Transformer models with the same architecture as in G2T. Parameter search yields an initial learning rate 1 with 3,000 warm-up steps and beam size 4. We empirically find using the 8 best models in ensemble decoding to be optimal. These models individually obtain between 23.51 and 24.00 BLEU-4.

Again, we observe that STMC-Transformer outperforms the previous system with ground truth glosses and Transformer. While STMC performs imperfect CSLR, its gloss predictions may be more useful than ground-truth annotations during SLT and are more readily analyzed by the Transformer. Again, the ground truth glosses represent merely a simplified intermediate representation of the actual sign language, so it is not entirely unexpected that translating ground truth glosses does not give the best performance.

STMC-Transformer also outperforms Transformers that translate GT glosses. While STMC performs imperfect CSLR, its gloss predictions may be better processed by the Transformer. Glosses are merely a simplified intermediate representation of the actual sign language so they may not be optimal. This result also reveals, training the recognition model to output more accurate glosses will not necessarily improve translation.

Both our STMC-Transformer and STMC-RNN also outperform Camgoz et al. (2020)'s model. Their best model jointly train Transformers for recognition and translation, however it obtains 24.49 WER on recognition whereas STMC obtains a better WER of 21.0, which suggests their model may be weaker in processing the videos.

Moreover, Transformers outperform recurrent networks in this setup as well and STMC-Transformer improves the state-of-the-art for video-to-text translation by 7 BLEU-4.

7 Qualitative comparison

Example outputs of the G2T and S2G2T models (Table 10) show that the translations are of generally good quality, even with low BLEU scores. Most translations may have slight differences in word choice that do not change the overall meaning of the sentence or make grammatical errors, which suggests BLEU is not a good representative of human useful features for SLT. As for the comparison between the G2T and S2G2T networks, there does not seem to be a clear pattern between cases where S2G2T outperforms G2T and vice versa. One thing to note, though, is that the PHOENIX-Weather 2014T is restricted to the weather forecast domain, and a SLT dataset with a wider domain would be required to fully assess the performance of our model in more general real-life settings.

We also provide sample G2T outputs on the ASLG-PC12 corpus in Appendix A.2.

8 Conclusions and Future Work

In this paper, we proposed Transformers for SLT, notably the STMC-Transformer. Our experiments demonstrate how Transformers obtain better SLT performance than previous RNN-based networks. We also achieve new state-of-the-art results on different translation tasks on the PHOENIX-Weather 2014T and ASLG-PC12 datasets.

A key finding is we obtain better performance by using a STMC network for tokenization instead of translating GT glosses. This calls into question current methods that use glosses as an intermediate representation, since reference glosses themselves are suboptimal.

End-to-end training without gloss supervision is one promising step, though Camgoz et al. (2020)'s end-to-end model does not yet surpass their joint training model. As future work, we suggest continuing work on end-to-end training of the recognition and translation models, so the recognition model learns an intermediate representation that optimizes translation, or using a different sign language annotation scheme that has less information loss.

		BLEU-4
REF:	ähnliches wetter auch am donnerstag . (similar weather on thursday .)	
G2T:	GLEICH WETTER AUCH DONNERSTAG (SAME WEATHER ON THURSDAY)	
	ähnliches wetter auch am donnerstag . (similar weather on thursday .)	100.00
S2G2T:	GLEICH WETTER DONNERSTAG (SAME WEATHER THURSDAY)	
	ähnliches wetter dann auch am donnerstag . (similar weather then on thursday .)	48.89
REF:	der wind weht meist schwach aus unterschiedlichen richtungen . (the wind usually blows weakly from different directions .)	
G2T:	WIND SCHWACH UNTERSCHIED KOMMEN (WIND WEAK DIFFERENCE COME)	
	der wind weht meist nur schwach aus unterschiedlichen richtungen . (the wind usually blows only weakly from different directions .)	65.80
S2G2T:	WIND SCHWACH UNTERSCHIED (WIND WEAK DIFFERENCE)	
	der wind weht schwach aus unterschiedlichen richtungen . (the wind is blowing weakly from different directions .)	61.02
REF:	sonnig geht es auch ins wochenende samstag ein herrlicher tag mit temperaturen bis siebzehn grad hier im westen . (the weekend is also sunny and saturday is a wonderful day with temperatures up to seventeen degrees here in the west .)	
G2T:	WOCHENENDE SONNE SAMSTAG SCHOEN TEMPERATUR BIS SIEBZEHN GRAD REGION (WEEKEND SUN SATURDAY NICE TEMPERATURE UNTIL SEVENTEEN DEGREE REGION)	
	und am wochenende da scheint die sonne bei temperaturen bis siebzehn grad . (and on the weekend the sun shines at temperatures up to seventeen degrees .)	13.49
S2G2T:	WOCHENENDE SONNE SAMSTAG TEMPERATUR BIS SIEBZEHN GRAD REGION (WEEKEND SUN SATURDAY TEMPERATURE UNTIL SEVENTEEN DEGREE REGION)	
	am wochenende scheint die sonne bei temperaturen bis siebzehn grad . (on the weekend sun shines at temperatures up to seventeen degrees .)	12.55
REF:	es gelten entsprechende warnungen des deutschen wetterdienstes . (appropriate warnings from the german weather service apply .)	
G2T:	IX SCHON WARNUNG DEUTSCH WETTER DIENST STURM KOENNEN (IX ALREADY WARNING GERMAN WEATHER SERVICE STORM CAN)	
	es bestehen entsprechende unwetterwarnungen des deutschen wetterdienstes . (severe weather warnings from the german weather service exist .)	38.26
S2G2T:	DANN IX SCHON WARNUNG DEUTSCH WETTER STURM KOENNEN	
	es gelten entsprechende warnungen des deutschen wetterdienstes . (THEN IX ALREADY WARNING GERMAN WEATHER STORM CAN) (appropriate warnings from the german weather service apply .)	100.00
REF:	richtung osten ist es meist sonnig . (it is mostly sunny towards the east .)	
G2T:	OST MEISTENS SONNE (MOST EAST SUN)	
	im osten bleibt es meist sonnig . (in the east it mostly stays sunny .)	43.47
S2G2T:	OST REGION MEISTENS SONNE	
	im osten ist es meist sonnig . (MOST REGION EAST SUN) (in the east it is mostly sunny .)	80.91
REF:	am tag elf grad im vogtland und einundzwanzig grad am oberrhein . (during the day eleven degrees in vogtland and twenty one degrees in upper rhine .)	
G2T:	AM-TAG ELF VOGEL LAND (IN-THE-DAY ELEVEN BIRD LAND)	
	elf grad am oberrhein . (eleven degrees in upper rhine .)	18.74
S2G2T:	ELF VOGEL ZWANZIG (ELEVEN BIRD TWENTY)	
	am tag elf grad im vogtland und zwanzig grad im vogtland . (during the day eleven degrees in vogtland and twenty degrees in vogtland .)	54.91

Table 10: Qualitative comparison of G2T and S2G2T on RWTH-PHEONIX-WEATHER 2014T. Glosses are capitalized. REF refers to the reference German translation.

Acknowledgements

The Titan X Pascal used for this research was donated by the NVIDIA Corporation. The authors would also like to thank Jean-Baptiste Rémy for the helpful discussions and feedback with various aspects of this work, and Hao Zhou for sharing the details of her previous work.

References

Nikolaos Arvanitis, Constantinos Constantinopoulos, and Dimitris Kosmopoulos. 2019. Translation of sign language glosses to text using sequence-to-sequence attention models. pages 296–302, 11.

Stylianos Asteriadis, George Caridakis, and Kostas Karpouzis. 2012. Non-manual cues in automatic sign language recognition. *Personal and Ubiquitous Computing*, 18.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084.

N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. 2018. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. 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 (CVPR)*, June.

Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS 2014 Deep Learning and Representation Learning Workshop*, abs/1412.3555.

Onno Crasborn, Johanna Mesch, Dafydd Waters, Annika Nonhebel, Els van der kooij, Bencie Woll, and Brita Bergman. 2007. Sharing sign language data online: Experiences from the echo project. *International Journal of Corpus Linguistics*, 12:535–562, 01.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1610–1618.

Jens Forster, Oscar Koller, Christian Oberdörfer, Yannick L. Gweth, and Hermann Ney. 2013. Improving continuous sign language recognition: Speech recognition techniques and system design. In *SLPAT*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. pages 369–376, 01.

Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. 2015. Sign language recognition using 3d convolutional neural networks. *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Trevor Johnston. 2013. Auslan corpus annotation guidelines.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aäron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *ArXiv*, abs/1610.10099.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Sangki Ko, Chang Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9:2683.

Oscar Koller, Sepehr Zargaran, and Hermann Ney. 2017. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3416–3424.

Jeroen Lichtenauer, Emile Hendriks, and Marcel Reinders. 2008. Sign language recognition by combining statistical dtw and independent classification. *IEEE transactions on pattern analysis and machine intelligence*, 30:2040–6, 12.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. 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, September. Association for Computational Linguistics.

Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. pages 4207–4215, 06.

Alptekin Orbay and Lale Akarun. 2020. Neural sign language translation by learning tokenization. *ArXiv*, abs/2002.00479.

Achraf Othman and Mohamed Jemni. 2012. English-asl gloss parallel corpus 2012: Aslg-pc12.

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, July. 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 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110.

Janet L. Pray and I. King Jordan. 2010. The deaf community and culture at a crossroads: Issues and challenges. volume 9, pages 168–193. Routledge. PMID: 20730674.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain, April. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana, June. Association for Computational Linguistics.

Maria Fernanda Neves Silveira de Souza, Amanda Miranda Brito Araújo, Luiza Fernandes Fonseca Sandes, Daniel Antunes Freitas, Wellington Danilo Soares, Raquel Schwenck de Mello Vianna, and Arlen Almeida Duarte de Sousa. 2017. Main difficulties and obstacles faced by the deaf community in health access: an integrative literature review. *Revista CEFAC*, 19:395 – 405.

William C. Stokoe. 1960. Sign language structure: an outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10 1:3–37.

Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 4, 09.

Alaa Tharwat, Tarek Gaber, Shahin MK, Basma Refaat, and Aboul Ella Hassanien Ali. 2014. Sift-based arabic sign language recognition system. In *The 1st Afro-European Conference for Industrial Advancement,*, Addis Ababa, Ethiopia, November 17-19,.

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 *NIPS*.

Q X Yang. 2010. Chinese sign language recognition based on video sequence appearance modeling. *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 1537–1542.

Jihai Zhang, Wengang Zhou, and Houqiang Li. 2014. A threshold-based hmm-dtw approach for continuous sign language recognition. *ACM International Conference Proceeding Series*, 07.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2020. Spatial-temporal multi-cue network for continuous sign language recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 13009–13016. AAAI Press.

A Appendices

A.1 Experiments on German G2T learning rate

A learning rate that is too low results in a notably slower convergence, but setting the learning rate too high risks leading the model to diverge. To prevent the model from diverging, we apply the Noam learning rate schedule where the learning rate increases linearly during the first training steps, or the warmup stage, then decreases proportionally to the inverse square root of the step number. The number of warmup steps is a parameter that has shown to influence Transformer performance (Popel and Bojar, 2018) therefore we first run a parameter search over the number of warmup steps before finding the optimal initial learning rate.

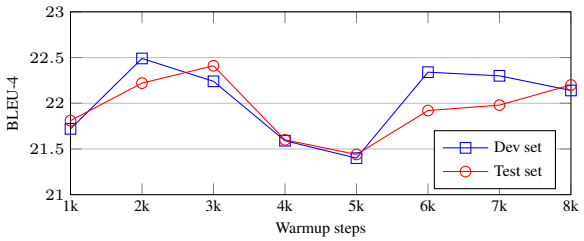

Figure 4: G2T performance on RWTH-PHEONIX-WEATHER 2014T with different warmup steps. Initial learning rate is fixed to 0.2.

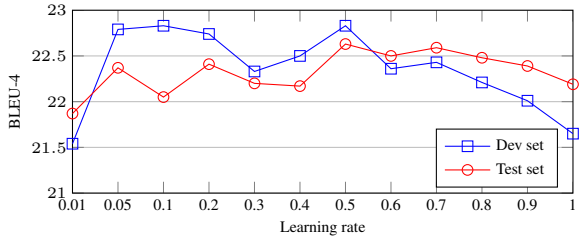

Figure 5: G2T performance on RWTH-PHEONIX-WEATHER 2014T with various initial learning rate.

A.2 Qualitative G2T Results on ASLG-PC12

Because quantitative metrics provide only a limited evaluation of translation performance, manual evaluation by viewing the translation outputs directly may give a better assessment of the quality of translations. Table 11 provides examples of SLT output on the ASLG-PC12 dataset. Here we can see how ASL glosses include prefixes that are not necessary to encapture the meaning of the phrase, which we have removed during data pre-processing before training. With a BLEU-4 testing score of 82.87, most predictions by our system are very close to the target English phrases and are able to convey the same meaning. We have also selected translation examples with lower BLEU-4 score and we can see that common errors include mistranslation of numbers and proper nouns. These are likely corner cases with infrequent examples during training.

		BLEU-4
ASL:	X-I BE DESC-PARTICULARLY DESC-GRATEFUL FOR EUROPEAN PARLIAMENT X-POSS DRIVE ROLE WHERE BALTIC SEA COOPERATION BE CONCERN	100.00
GT:	i am particularly grateful for the european parliament's driving role where the baltic sea cooperation is concerned .	
Pred:	i am particularly grateful for the european parliament's driving role where the baltic sea cooperation is concerned .	
ASL:	DESC-REFORE , DESC-MUCH WORK NEED TO BE DO IN ORDER TO DESC-FURR SIMPLIFY RULE	100.00
GT:	therefore , much work needs to be done in order to further simplify the rules .	
Pred:	therefore , much work needs to be done in order to further simplify the rules .	
ASL:	THIS PRESSURE BE DESC-PARTICULARLY DESC-GREAT ALONG UNION X-POSS DESC-SOURN AND DESC-EASTERN BORDER	100.00
GT:	this pressure is particularly great along the union's southern and eastern borders .	
Pred:	this pressure is particularly great along the union's southern and eastern borders .	
ASL:	MORE WOMAN DIE FROM AGGRESSION DESC-DIRECT AGAINST X-Y THAN DIE FROM CANCER .	73.15
GT:	more women die from the aggression directed against them than die from cancer .	
Pred:	more women die from aggression directed against them than die from cancer .	
ASL:	X-IT FUEL WAR IN CAMBODIUM IN 1990 AND X-IT BE ENEMY DEMOCRACY	25.89
GT:	it fuelled the war in cambodia in the 1990s and it is the enemy of democracy .	
Pred:	it fuel war in the cambodium in 1990 and it is an enemy of democracy .	
ASL:	DESC-N CHIEF INVESTIGATOR X-HIMSELF BE TARGET AND HOUSE CARD COLLAPSE .	21.29
GT:	then the chief investigator himself is targeted and the house of cards collapses .	
Pred:	then chief investigator himself is a target and a house card collapse .	
ASL:	U , X-WE TAKE DESC-DUE NOTE X-YOU OBSERVATION . AMENDMENT THANK X-YOU MR	15.93
GT:	otherwise we have to vote on the corresponding part of amendment thank you mrs ţicău , we take due note of your observation .	
Pred:	mr president , we took due note of your observation .	

Table 11: Examples of ASL translation with varying BLEU-4 scores