MLSLT: Towards Multilingual Sign Language Translation

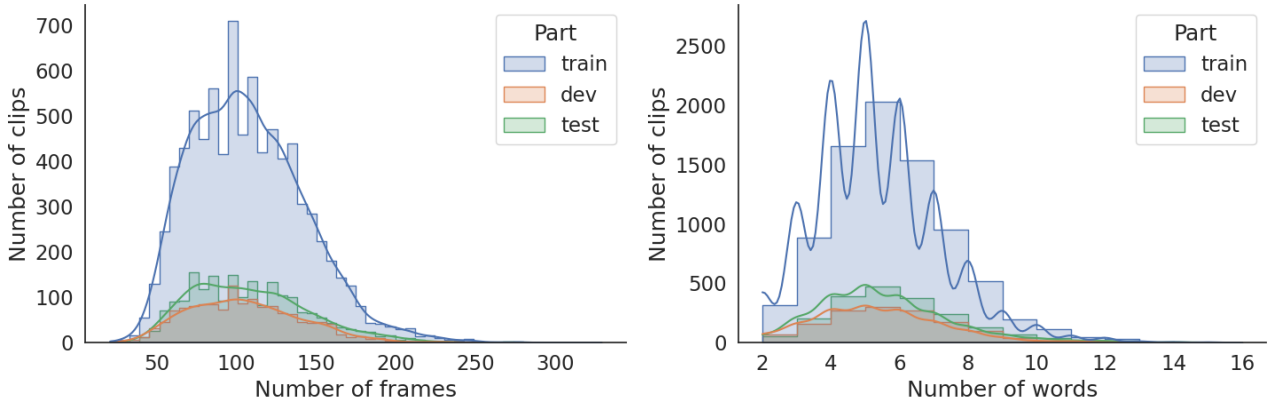
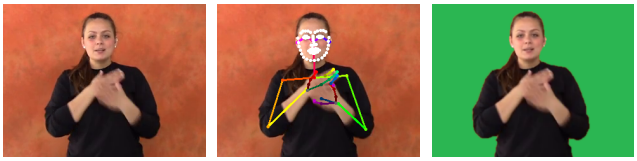


Figure 1. Distribution of the number of frames (left) and words (right) over sentence-level clips.

1. Dataset Details

The distribution of frames and words over all the clips for the 3 splits of the dataset can be seen in Figure 1. It can be seen that the three parts of the dataset have similar distributions in terms of the number of video frames and the length of the text, which proves that our division of the dataset is reasonable.



(a) RGB Image (b) Skeletal Data (c) Green Screen Image

Figure 2. An video example of SP-10 with RGB images, skeletal data and green screen images.

In Figure 2, we show examples of the three types of data contained in SP-10. For the transparent background video data, we replace its background with a standard green background, as shown in Figure 2c. In fact, in order to simulate natural scenes, we can arbitrarily set a variety of scenes as the background (such as subway stations with dense traffic, roadside with dense vehicles, etc.)

The distribution of video frames for different sign languages is shown in Figure 3. It can be seen that the length of sign language videos is mostly around 50-150 frames, and the distribution of different types of sign language is different. The video frame number distribution of some sign languages is quite different from that of other sign languages. For example, the number of video frames of BQN is relatively shorter, and the distribution is more concentrated.

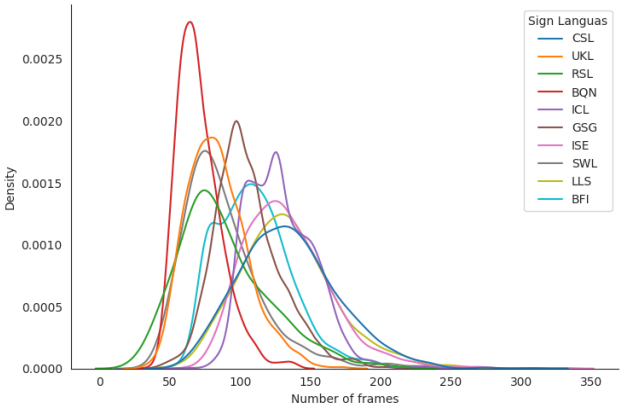


Figure 3. Distribution of video frames in different sign languages

The distribution of sentence lengths for different spoken languages is shown in Figure 4. It can be seen that the sentence length distributions of almost all spoken languages are

Table 1. Comparison of many-to-one translation results and BSLT baseline results. The metrics shown in the table are the results of averaging the ten source languages.

target language	Dev/BLEU4		Dev/ROUGE		Test/ROUGE		Test/ROUGE	
	single	MLSLT (ours)	single	MLSLT (ours)	single	MLSLT (ours)	single	MLSLT (ours)
zh	2.22	**2.69**	32.73	**33.91**	1.64	**4.22**	32.72	**33.84**
uk	3.40	**3.82**	28.68	**29.49**	1.25	**1.86**	27.79	**28.92**
ru	2.75	**3.19**	29.28	**31.07**	0.54	**1.50**	27.96	**29.42**
bg	2.84	**4.32**	29.24	**33.54**	1.06	**2.04**	27.13	**29.93**
is	3.52	**3.74**	30.91	**31.61**	1.65	**2.68**	29.62	**31.21**
de	4.96	**5.23**	30.99	**33.80**	2.41	**3.70**	29.64	**31.56**
it	3.64	**3.99**	27.44	**31.34**	1.51	**2.49**	25.39	**26.74**
sv	4.91	**5.47**	32.79	**33.57**	2.63	**3.40**	31.36	**32.75**
lt	2.13	**2.52**	29.01	**31.81**	0.43	**1.25**	29.27	**30.60**

Table 2. Comparison of one-to-many translation results and BSLT baseline results. The metrics shown in the table are the results of averaging the ten target languages.

source language	Dev/BLEU4		Dev/ROUGE		Test/ROUGE		Test/ROUGE	
	single	MLSLT (ours)	single	MLSLT (ours)	single	MLSLT (ours)	single	MLSLT (ours)
CSL	3.59	**4.06**	30.57	**32.47**	1.47	**3.03**	29.47	**32.51**
UKL	3.78	**3.94**	31.53	**32.72**	1.26	**1.47**	30.29	**30.77**
RSL	3.14	**3.58**	30.06	**32.40**	1.19	**2.12**	28.70	**30.91**
BQN	2.67	**2.81**	27.19	**29.65**	0.90	**1.74**	26.11	**28.58**
ICL	3.89	**4.94**	31.09	**35.14**	2.20	**3.85**	30.45	**33.58**
GSG	3.63	**4.00**	30.45	**32.30**	2.21	**3.41**	29.24	**30.78**
ISE	4.01	**4.21**	30.59	**31.75**	1.82	**2.74**	29.72	**31.06**
SWL	4.48	**4.68**	32.27	**33.57**	2.31	**3.28**	31.14	**32.42**
LLS	3.13	**3.61**	28.86	**31.36**	1.43	**2.72**	27.95	**30.01**

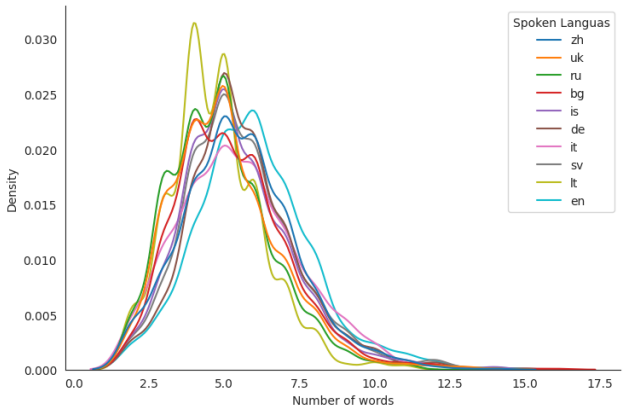


Figure 4. Distribution of sentence lengths in different spoken languages.

very similar, except for the slight differences in local positions.

2. Many to One

Table 1 shows the results of the remaining 9 spoken languages as the target language of many to one translation. It can be seen that in all cases, the average performance of our proposed model exceeds the BSLT baseline, while the number of parameters is only one-tenth of it. This shows that our proposed model is suitable for various many to one translation scenarios, not just those that are translated into English.

3. One to Many

Table 2 shows the results of the remaining 9 sign languages as the source language of one to many translation. It can be seen that for the other nine cases when performing one to many translation, the average performance of MLSLT also exceeds the baseline of BSLT. This proves the effectiveness of our proposed method in one to many sign language translation scenarios.

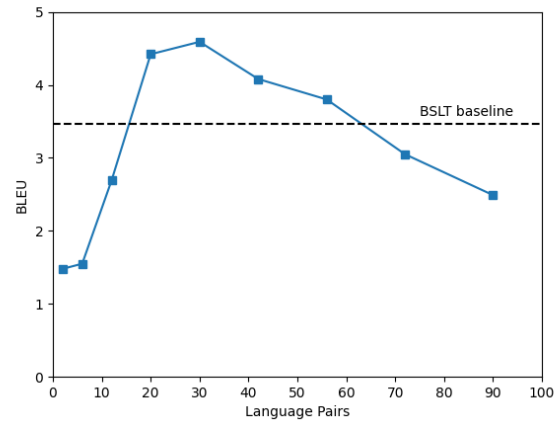


Figure 5. The influence of the number of language pairs participating in the training on the BLEU score of zero-shot translation.

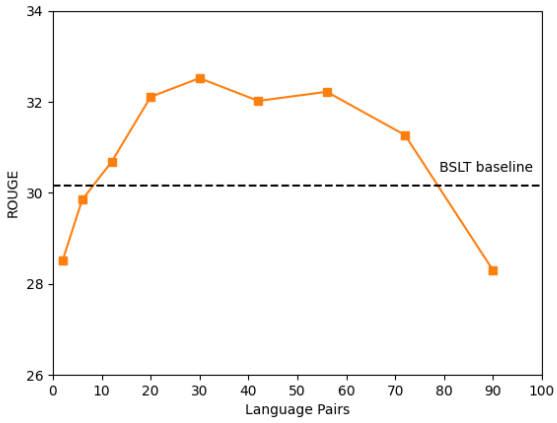


Figure 6. The influence of the number of language pairs participating in the training on the ROUGE score of zero-shot translation.

4. Zero-Shot Translation

In this section, we try to explore the impact of language pairs included in the training set on translation quality when performing zero-shot translation. We choose CSL→en as the zero-shot translation language pair to be observed, and gradually increase the language pairs included in the training set from 2×(2-1) to 10×(10-1). As shown in Figure 5 and Figure 6, as the language pairs used in training increase, the zero-shot translation performance of the model first rises rapidly and then slowly declines. In some cases, the performance of the zero-shot translation of the ML-SLT model even exceeds the performance of the supervised BSLT model. This shows that increasing the number of language pairs involved in training when the number of language pairs is small can help improve the performance of zero-shot translation, because it can help the model to build implicit bridges better. After the number of language pairs participating in the training exceeds a certain number, the zero-shot translation performance of the model begins to decrease. The increase of language pairs will create greater challenges for constructing shared semantic space and the generation of spoken text, which limits the performance of zero-shot translation.

5. Limitations and Potential Negative Effects

Our work does not have apparent negative effects, but since facial expressions are also an important part of sign language, it can convey a wealth of information, and the existing methods based on RGB video sequences may infringe on users' privacy. In order to protect the privacy of users, in the future, we can try to develop some facial blur methods that do not affect translation performance, and sign language translation methods based on skeletal sequences. The data used by our model are all sign language videos taken in an ideal environment, which may cause the translation performance of our model to decline in some actual scenes, such as a dimly lit environment or the camera's inaccurate focus.