SSL: Korean Disaster Safety Information Sign Language Translation Benchmark Dataset

Wooyoung Kim*, **Taeyong Kim***, **Byeongjin Kim***, **Myeongjin Lee***
Gitaek Lee*, **Kirok Kim***, **Jisoo Cha***, **Wooju Kim**†
Smart Systems Lab, Yonsei University
{timothy, kasamdi5, jin_kbj, myeongjin216, gitaekl, alfmalfm11, jisoo.cha, wkim}@yonsei.ac.kr

Abstract

Sign language is a crucial means of communication for deaf communities. However, those outside deaf communities often lack understanding of sign language, leading to inadequate communication accessibility for the deaf. Therefore, sign language translation is a significantly important research area. In this context, we present a new benchmark dataset for Korean sign language translation named **SSL:korean disaster Safety information Sign Language translation benchmark dataset**. Korean sign language translation datasets provided by the National Information Society Agency in South Korea have faced challenges related to computational resources, heterogeneity between train and test sets, and unrefined data. To alleviate the aforementioned issue, we refine the origin data and release them. Additionally, we report experimental results of baseline using a transformer architecture. We empirically demonstrate that the baseline performance varies depending on the tokenization method applied to gloss sequences. In particular, tokenization based on characteristics of sign language outperforms tokenization considering characteristics of spoken language and tokenization utilizing statistical techniques.
We release materials at our https://github.com/SSL-Sign-Language/Korean-Disaster-Safety-Information-Sign-Language-Translation-Benchmark-Dataset

Keywords: Sign Language Translation, Sign Language Recognition, Korean Sign Language

1. Introduction

Sign language is a primary method of communication for deaf communities. Sign language is a language that conveys meaning not only through hand movements (manual elements) but also through body language, facial expressions, and other non-manual elements. It has its own unique linguistic system that distinguishes it from spoken language (Stokoe, 1980). Most people outside deaf communities do not understand sign language, which makes social communication hard. Considering this situation, research on sign language translation holds significant value.
However, sign language translation is more challenging than general spoken language machine translation. To perform sign language translation, it requires the conversion of sign language videos or skeleton information extracted from videos into text (Cheok et al., 2019; Rastgoo et al., 2021; Núñez-Marcos et al., 2022). This task of bridging different modalities presents a difficulty in modeling. Furthermore, to train neural network models for this task, video-text paired data is necessary, and constructing such data is costly.
Fortunately, the National Information Society Agency (NIA) in South Korea has provided labeled data for Korean sign language research. However, there are many inconveniences in using this data for research. Therefore, we have reprocessed and distributed the data to make it suitable for benchmarking. Additionally, we have presented baseline performance using the Transformer architecture (Vaswani et al., 2017; Camgoz et al., 2020).

2. Related Works

2.1. Sign Language Recognition and Translation

Sign Language Recognition (SLR) is the task of converting sign language videos into gloss(or gloss sequence), which is the smallest semantic unit in sign language. SLR can be categorized into two types: Isolated Sign Language Recognition (ISLR) and Continuous Sign Language Recognition (CSLR). ISLR treats each video as referencing a single meaning, similar to translating individual words, while CSLR deals with videos where multiple meaningful gestures occur continuously, resembling the translation of sentences. Although ISLR has value in research fields, it comes with practical limitations. To ensure smooth practical applications, the focus should shift towards CSLR tasks.
Sign Language Translation (SLT) is a task that transforms sign language into spoken language sentences. Similar to spoken language machine translation, Sequence-to-Sequence structures are

*These authors contributed equally
†Coressponding Authoer

LREC-COLING 2024, pages 9948–9953
20-25 May, 2024.

	Count (Train/Test)	Frame (min/max/average)	FPS (min/max/average)	Resolution
Original Data	58699 / 7341	44 / 1804 / 418	2.9 / 60 / 28.4	1920x1080 RGB
SSL (Ours)	10170 / 2452	90 / 390 / 274	25 / 30 / 29.6	256x256 RGB

Table 1: Statistics of the Disaster Safety Information Sign Language Video Dataset (Original Data) and korean disaster Safery information Sign Language translation benchmark dataset (SSL).

widely used in Sign Language Translation (Camgoz et al., 2018; Guo et al., 2018; Ko et al., 2019). Recently, Camgoz et al. (2020) has utilized a CNN backbone network and a transformer architecture to jointly train SLR in the encoder and SLT in the decoder, resulting in significant improvements in SLT performance.

2.2. Sign Language Translation Datasets

For SLT, various countries worldwide have constructed and made sign language translation datasets publicly available. One of the most notable examples is the German Sign Language dataset **RWTH-PHOENIX-Weather-2014** (Forster et al., 2014). **How2Sign** (Duarte et al., 2021) for American Sign Language translation and **CSL-Daily** (Zhou et al., 2021) for Chinese Sign Language translation are widely used.

In South Korea, the National Information Society Agency (NIA) has also contributed to Korean Sign Language translation such as the **Sign Language Video Dataset** and the **Disaster Safety Information Sign Language Video Dataset**[1]. The **KETI Sign Language Dataset** (Ko et al., 2019) for korean sign language translation tasks is also widely used. These datasets serve as valuable resources for research but have certain limitations. For instance, the **Sign Language Video Dataset** misses spoken language text labels, only providing gloss level labels and oppositely **KETI Sign Language Dataset** has only provided spoken language texts, missing gloss level labels. The **Disaster Safety Information Sign Language Video Dataset** contains both spoken language text and gloss labels; however, it faces challenges due to variations in the number of frames per video and heterogeneity between train and test data.

3. Korean Sign Language Translation Dataset

To address the challenges posed by the Disaster Safety Information Sign Language Video Dataset, including 1) significant variations in Frames Per Second (FPS) and the number of frames, 2) disparities between the train set and test set, and 3) videos that contain excessive background noise, we have reprocessed the data and created a new Korean Sign Language Translation Dataset named **SSL:korean disaster Safety information Sign Language translation benchmark dataset**.

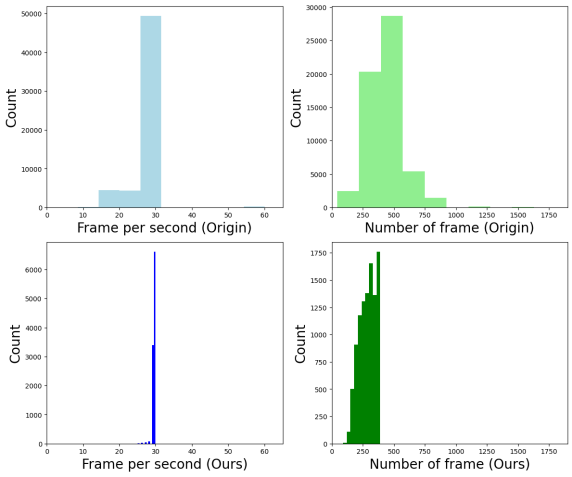

Figure 1: Statistics for FPS (blue) and the number of frames (green) for both the original data and the newly processed data. After processing, the FPS values fall within the range of 25 to 30, and the number of frames is limited to 400 or fewer.

3.1. FPS and The Number of Frames

If FPS is too low, videos appear unnaturally choppy and are not suitable for accurate sign language recognition and translation. Conversely, when FPS is excessively high or the number of frames is too large, it demands excessive computational resources.

In the source data, Disaster Safety Information Sign Language Video Dataset, FPS varies from 2 to 60. To address this, we refer to FPS of existing benchmark data (Von Agris et al., 2008; Forster et al., 2014; Huang et al., 2018; Adaloglou et al., 2021) and select data with FPS ranging from 25 to 30. Furthermore, while the average number of frames is 418, there are extremely long videos that exceed this average. When using the Transformer architecture with the same structure as Camgoz et al. (2020), training a 400-frame video requires about 21GB of GPU memory. Considering the spatial complexity of the Transformer architecture due to Self-Attention being $\mathcal{O}(n^2)$, handling videos with the large number of frames becomes challenging. Therefore, we choose to select data with fewer than 400 frames (Fig 1).

[1] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=636

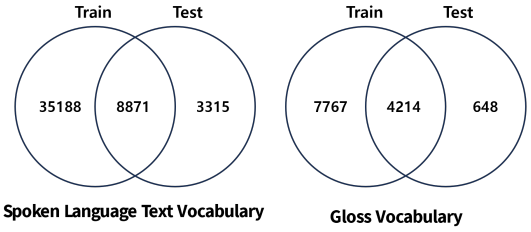

Figure 2: Heterogeneity between the train and test datasets in the Disaster Safety Information Sign Language Video Dataset (original data)

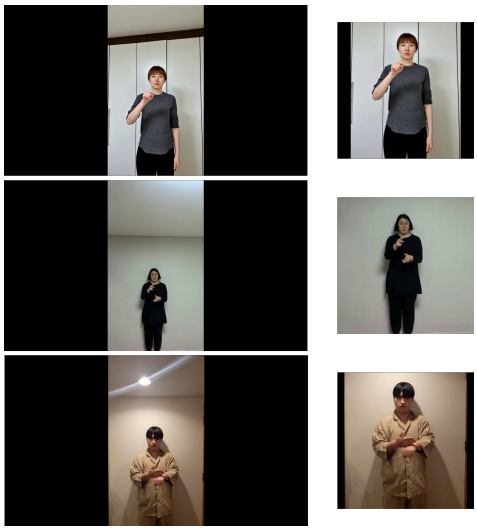

Figure 3: Examples before (left) and after (right) cropping and resizing. We refine the source data to fix the problems including slanted backgrounds, too-small subjects and distractive information.

3.2. Heterogeneity between the Train and Test Data

Our analysis of the vocabulary in the train and test datasets in the original data reveals significant disparities. In the case of spoken language text, the train and test datasets have 44,059 and 12,186 vocabularies. However, the overlapping vocabularies between them are only 8,871 words. Similarly, the gloss vocabularies in the train and test datasets numbered 11,981 and 4,862 with only 4,214 words overlapping (Fig 2). Such heterogeneity between the train and test datasets is not suitable for evaluation. Therefore, we extracted a new set of train and test data composed of the intersection of vocabulary between the original train dataset and the original test dataset.

3.3. Video Prepreocessing

The source video data contains a significant amount of undesirable information, including slanted backgrounds, too-small subjects, distractive information, and so on. Therefore, we utilize the human body skeleton information to crop and retain only the parts representing the individuals. After cropping, we resize the video to a resolution of 256x256 and make the extracted skeleton information available alongside the resized video (Fig 3). [2]

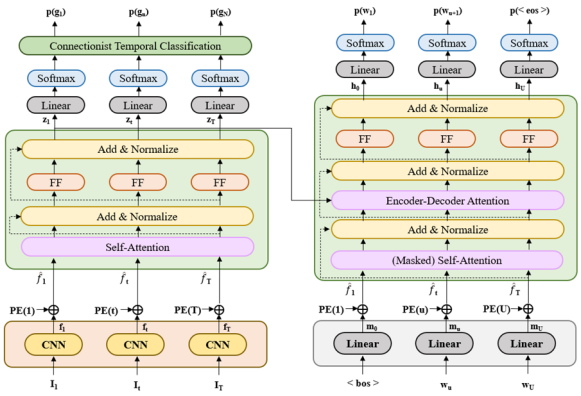

Figure 4: A detailed overview of a single layered Joint learning Sign Language Transformer.

4. Experiment

We conduct experiments using the Joint Learning Sign Language Transformer (JSLT) (Camgoz et al., 2020). In the JSLT, a convolutional neural network (CNN) extracts features from each frame in videos. These extracted features are forwarded through the transformer architecture. In the encoder part, it predicts glosses, while in the decoder part, it generates spoken language text through multi-task training.

The problem entails simultaneously optimizing the probabilities $P(G|V)$ and $P(S|V)$ for a given Frames (video) $V = (f_1, f_2 ... f_{|v|})$, aligned gloss sequence $G = (g_1, g_2 ... g_{|G|})$, and spoken language text sequence $S = (w_1, w_2 ... w_{|S|})$. The loss functions employed for $P(G|V)$ and $P(S|V)$ are Connectionist Temporal Classification Loss (L_g) and Cross-Entropy Loss (L_s). The goal of the training is to minimize $Lg + \lambda Ls$.

The λ is a hyperparameter that represents the weight between the two loss functions. In our experiments, we use 0.2 as default. We use the pretrained EfficientNet-7B (Tan and Le, 2019) model from ImageNet as the CNN backbone in our experiments. For training, we set the learning rate to 1e-3 and the weight decay to 1e-3, using the Adam optimizer (Kingma and Ba, 2014).

[2] For skeletons, we empoly the MediaPipe Holistic module. https://github.com/google/mediapipe.

Tokenizer	SLT				SLR
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	WER
GDT	**47.91**	**39.13**	**33.31**	**29.33**	**42.36**
Morpheme	45.67	36.77	30.88	26.96	55.14
BPE	47.07	38.09	32.23	28.22	56.98

Table 2: The experimental results for SLT and SLR based on different tokenization methods. It is evident that using a Gloss Dictionary based Tokenization (GDT) for gloss sequence tokenization performs better in both SLT and SLR. In the case of BLEU, higher values indicate better performance, while for WER, lower values indicate better performance.

4.1. Gloss Tokenization Method

In the development of language models, tokenization, which involves segmenting natural language sentences into tokens, is a crucial aspect as it directly impacts the input's smallest unit. In the case of spoken language, subword tokenization methods like Byte Pair Encoding (BPE) (Sennrich et al., 2015) are widely employed to address out-of-vocabulary (OOV) issues. However, there is currently no reported research on tokenization methods for gloss sequences used in SLR Tasks. Therefore, we conducted cross-experiments on the Gloss tokenization method.

- **Gloss Dictionary based Tokenization (GDT)**: The Origin Dataset (Disaster Safety Information Sign Language Video Dataset from NIA) contains segmented gloss information, considering the linguistic features of sign language, manually by manually annotated by experts and deaf. We conducted experiments using a tokenizer that uses the sign language linguistic features. We also provide this information together with our dataset.
- **Morpheme based Tokenization**: In typical natural language processing tasks, morphological tokenization methods are widely used. However, since gloss has linguistic characteristics different from spoken languages, it is necessary to verify whether (spoken language) morphological tokenization methods are suitable for sign language tasks.[3]
- **Statistical Subword Toeknization**: The tokenization utilizing a dictionary often encounters the out-of-vocabulary (OOV). To mitigate this, statistical-based subword tokenization methods have been researched and are currently the most prominently used in recent studies. We conduct experiments using one of the subword tokenization methods, Byte Pair Encoding (BPE) (Sennrich et al., 2015). The size of the vocabulary is set to 8000, referring to Gowda and May (2020).

[3]We employ the Okt (Twitter) module in the KoNLPy package. https://konlpy.org

4.2. Results

For the evaluation, we utilize BLEU for SLT and Word Error Rate (WER) Score for SLR. BLEU (Papineni et al., 2002a) is a commonly used metric in machine translation to evaluate the proportion of predicted sentence tokens that match the reference sentence tokens. Also, BLEU evaluates at the n-gram level (BLEU-n) (Papineni et al., 2002b). WER (Koller et al., 2015) is an evaluation metric that quantifies token-level errors between the predicted results and the actual reference.

Table 2 presents the experimental results. Depending on the tokenization method, BLEU-4 shows differences of up to 8%. The best performance is observed when using a GDT tokenizer for both SLT and SLR. An interesting observation is that the morpheme based tokenizer, designed considering the semantics of spoken language, induces the lowest performance. This implies the need for distinct semantic segmentation methods between spoken language and sign language. Furthermore, while the BPE Tokenizer is commonly used in spoken language translation, it results in a slight performance decrease in sign language translation. Through our experiments, we have demonstrated that sign language and spoken language exhibit different characteristics and, consequently, require different tokenization approaches.

5. Conclusion

We refine the Disaster Safety Information Sign Language Video Dataset from NIA in South Korea and release a new benchmark dataset, **SSL:korean disaster Safety information Sign Language translation benchmark dataset**. Through experiments, we disclose the baseline performance of our benchmark dataset. Furthermore, our experiments highlight significant performance differences based on the gloss tokenization method. These results indicate the need for additional research into gloss tokenization methods. We hope that our research will contribute to furthering sign language translation studies, ultimately ensuring communication rights for deaf communities.

6. Limitations

Through our experiments, we demonstrate performance differences based on tokenization methods. However, we are unable to conduct experiments with various models.
Furthermore, while we empirically demonstrated that GDT tokenization yields the best performance, we do not provide adequate linguistic evidence for this result (Yin et al., 2021). Collaborative efforts with sign language experts and linguists are necessary to delve into the linguistic structure of sign language, leading to more sophisticated experimental designs and further research.

7. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C100697011).

This work is financially supported by Korea Ministry of Land, Infrastructure and Transport(MOLIT) as Innovative Talent Education Program for Smart City.

Thank you to Eunyoung Lee and Donghyun Kim from the Institute of Korean Sign Language for their valuable insight to the research.

8. References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2021. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia*, 24:1750–1762.

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. 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*, pages 10023–10033.

Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. 2019. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10:131–153.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2018. Hierarchical lstm for sign language translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Adrián Núñez-Marcos, Olatz Perez-de Viñaspre, and Gorka Labaka. 2022. A survey on sign language machine translation. *Expert Systems with Applications*, page 118993.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. 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.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

William C Stokoe. 1980. Sign language structure. *Annual review of anthropology*, 9(1):365–390.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE international conference on automatic face & gesture recognition*, pages 1–6. IEEE.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*.

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