



# **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**

**LREC-COLING 2024**

1. Disclosure
2. Background
3. Motivation
4. Data Processing
5. Experiment
6. Result
7. Conclusion

# Contents

## Disclosure

### ● Assistance

- Our research was financially supported by the National Research Foundation of Korea (NRF)



- Our research received advice on sign language related matters from the Institute Korean Sign Language (IKSL)



## Background

### ● What is sign language?

- Primary method of communication for deaf communities
  - Variety of visual cues
  - hand movements (manual elements), body language, facial expressions, and other (non-manual elements)
- Sign language has its own unique linguistic system that distinguishes it from spoken language
  - So most people don't understand sign language
- Research on sign language translation holds significant value



## Background

### ● Sign language in AI research

- Task : Sign Language Recognition(SLR), Sign Language Translation(SLT), Sign Language Production(SLP) etc.
- Dataset : Pair Sign language Video, Gloss, Text

- Gloss : Minimal sign language lexicon

Labeling require specialized expertise in sign language and entail high costs



**Korean Text** : "광주광역시 홍수주의보 발령, 안전에 유의하시기 바랍니다."  
("Flood warning issued in Gwangju, please be safe.")

**Korean Gloss Sequence** : "광주1 지역1 비내리다1# 차오르다1# 주의보1 지시2# 조심1"  
("Gwangju1 region1 rain1# fill up1# advisory1 instruction2# careful1")

## Motivation

### ● Sign Language Translation Datasets

- Various countries worldwide have constructed and made sign language translation datasets publicly available

- In South Korea

The National Information Society Agency (NIA)

- Sign Language Video Dataset

- Disaster Safety Information Sign Language Video Dataset

KETI(Korea Electronics Technology Institute) Sign Language Dataset

(Ko, Sang – Ki, et al. "Neural sign language translation based on human keypoint estimation." *Applied sciences* 9.13 (2019): 2683.)

*This paper used datasets from 'Disaster Safety Information Sign Language Video Dataset (AI – Hub, S. Korea)'.*

*All data information can be accessed through 'AI – Hub (www.aihub.or.kr)'*

## Motivation

### ● South Korea Sign Language Translation Datasets

#### ▪ Limitations

- Sign Language Video Dataset : Gloss level labels only (no spoken language text labels)
- KETI Sign Language Dataset : Spoken language text labels only (no gloss level labels)
- Disaster Safety Information Sign Language Video Dataset : Gloss level labels, Spoken language text labels but, variations in the number of frames per video and heterogeneity between train and test data

- **So, we have reprocessed the data and created a new Korean Sign Language Translation Dataset and we report experimental results of baseline using a transformer architecture.**
- **Baseline performance varies depending on the tokenization method applied to gloss sequences**

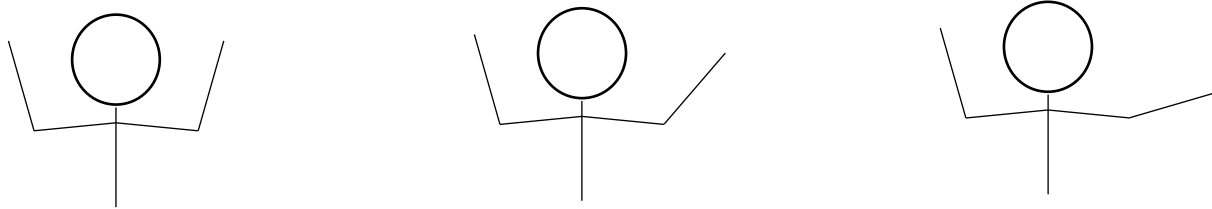
*This paper used datasets from 'Disaster Safety Information Sign Language Video Dataset (AI – Hub, S. Korea)'.*

*All data information can be accessed through 'AI – Hub (www.aihub.or.kr)'*

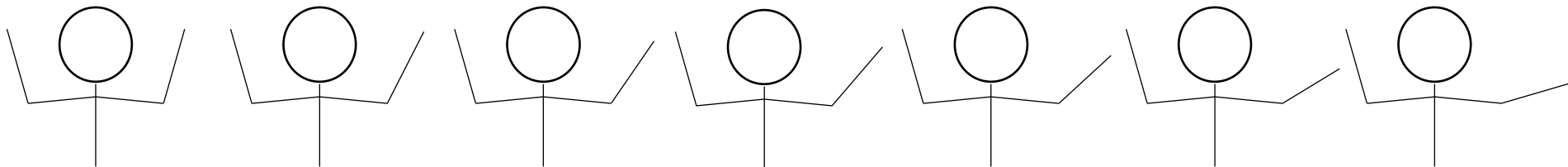
# Data reprocessing

- FPS

- Low FPS



- High FPS

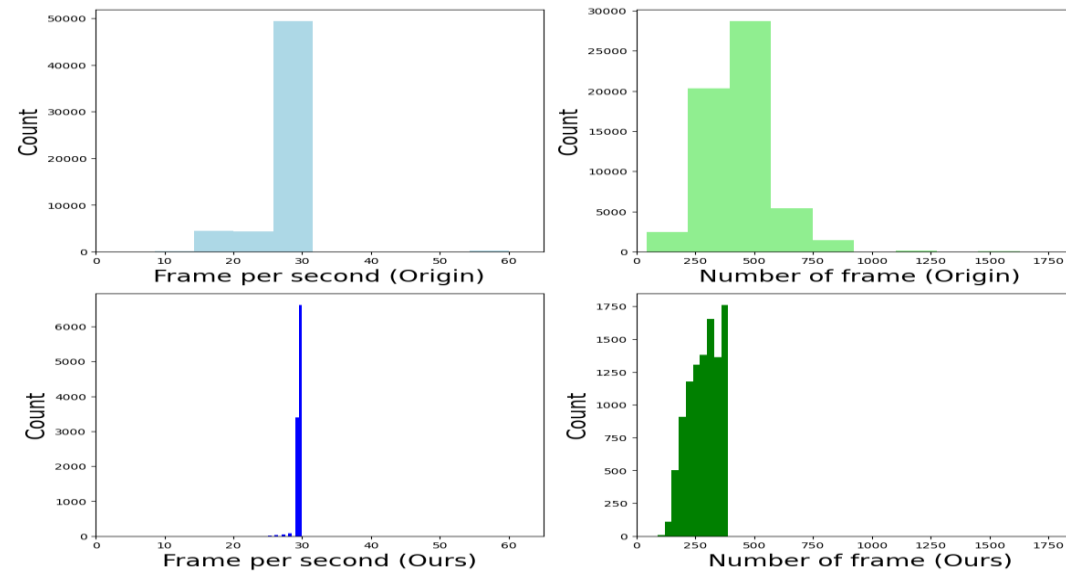


- In the source data, Disaster Safety Information Sign Language Video Dataset, FPS varies from 2 to 60
    - Selected data within the 25~30 FPS range, referencing existing benchmark datasets

# Data reprocessing

## ● The Number of Frames

- In the source data the average number of frames is 418, there are extremely long videos that exceed this average.
  - Considering the spatial complexity of the Transformer architecture due to Self-Attention being  $O(n^2)$
- We choose to select data with fewer than 400 frames

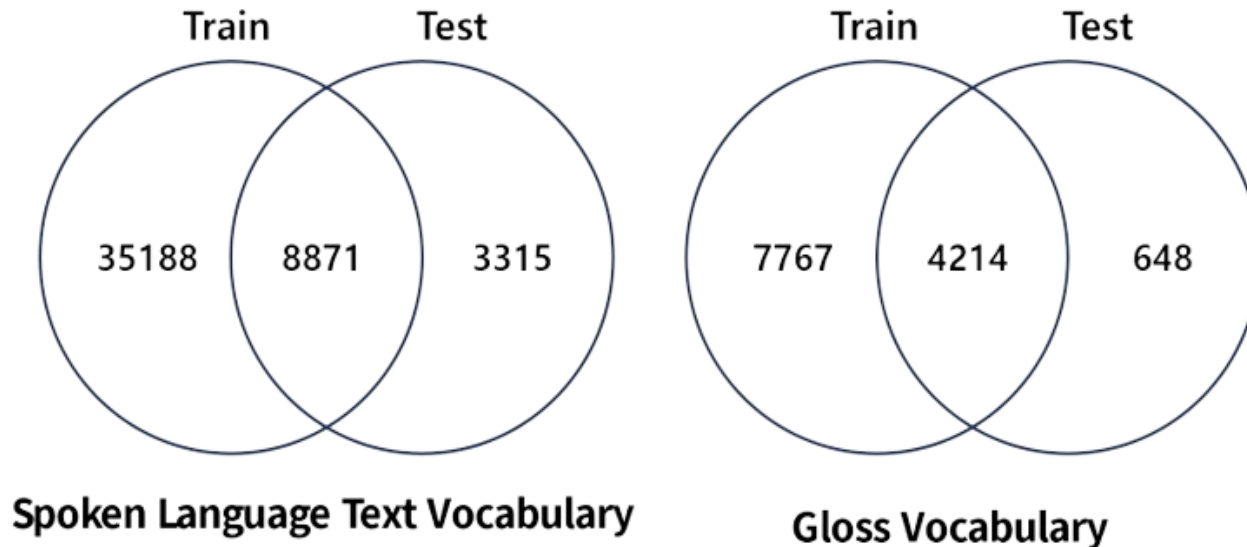


	<i>Count(Train/Test)</i>	<i>Frame (min/max/average)</i>	<i>FPS (min/max/average)</i>	<i>Resolution</i>
<i>Original Data</i>	58699 / 7341	44 / 1804 / 418	2.9 / 60 / 28.4	1920x1080 RGB
<i>SSL (Ours)</i>	10170 / 2452	90 / 390 / 274	25 / 30 / 29.6	256x256 RGB

## Data reprocessing

### ● Heterogeneity between the Train and Test Data

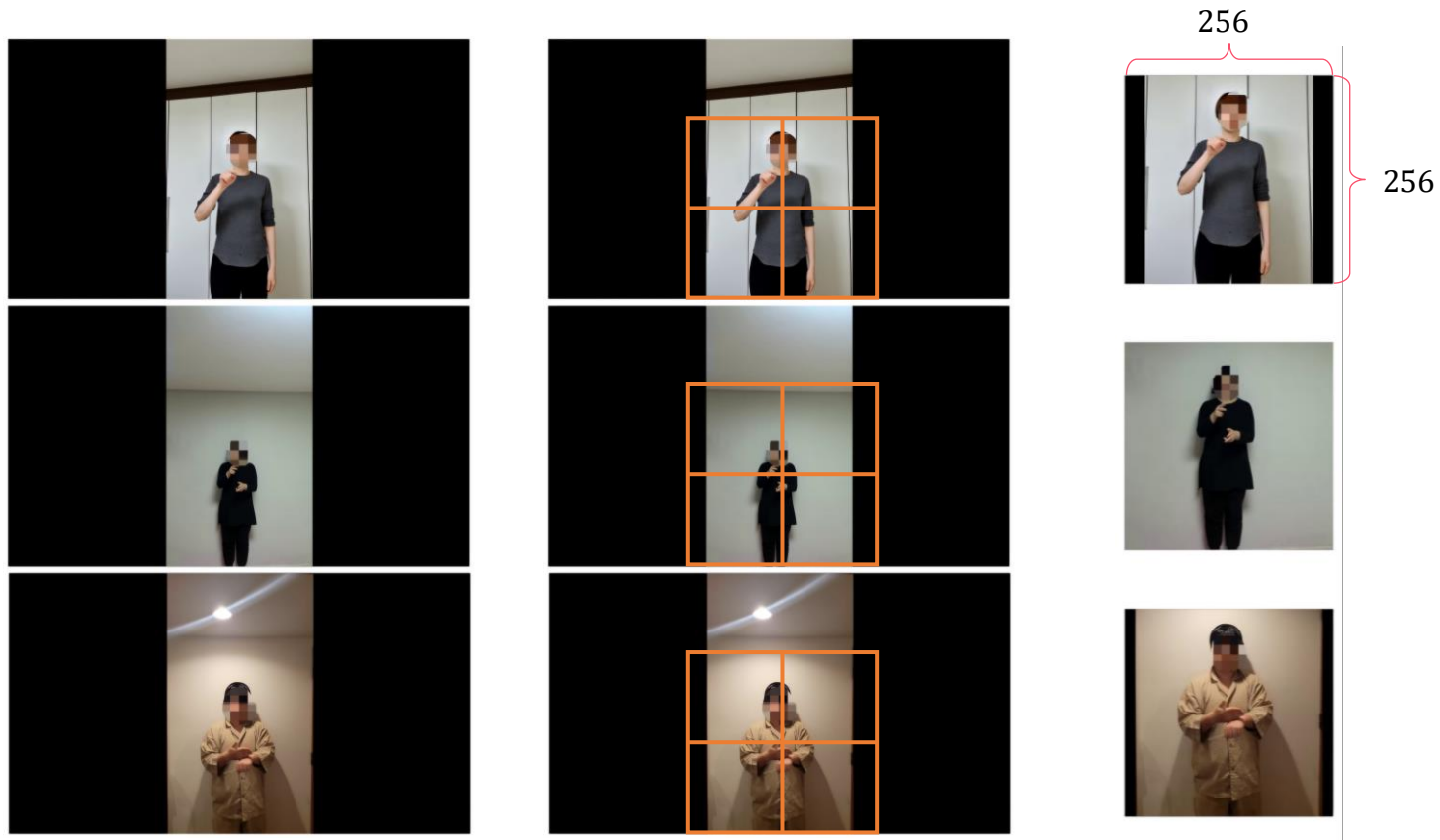
- Our analysis of the vocabulary in the train and test datasets in the original data reveals significant disparities
- Composed of the intersection of vocabulary between the original train dataset and the original test dataset



# Data reprocessing

## ● Video Preprocessing

- Significant amount of undesirable information (slanted background, too-small subject etc.)



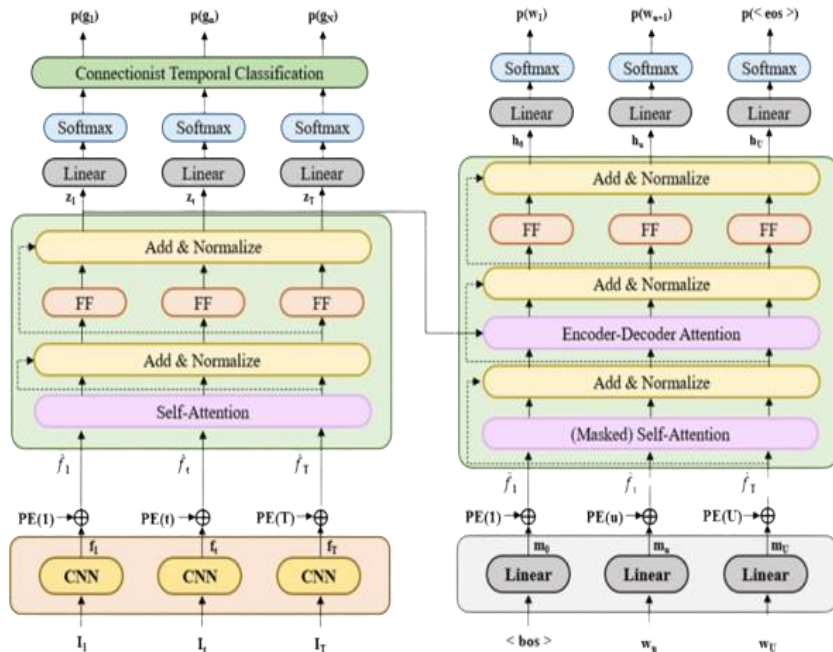
# Experiment

## ● Baseline

### ■ Use Joint Learning Sign Language Transformer (JSLT)

(Camgoz, Necati Cihan, et al. "Sign language transformers: Joint end – to – end sign language recognition and translation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.)

### ■ Architecture



### ■ Training step

- Extracts features from each frame in videos EfficientNet-7B (Frozen)
- Multi-task training  
Transformer encoder part : predict gloss  
Transformer decoder part : generates spoken language text
- Simultaneously optimizing the probabilities  $P(G|V), P(S|V)$   
( $V$  : Frames (video),  $G$  : Gloss sequence,  $S$  : spoken language text sequence)
- Connectionist Temporal Classification Loss + Cross-Entropy Loss

## Experiment

### ● Gloss Tokenization Method

- Spoken language often uses subword tokenization methods such as Byte Pair Encoding (BPE) to handle out-of-vocabulary (OOV) problems
- No reported research on tokenization methods for gloss sequences used in SLR/SLT Tasks
- Cross-experiments on the Gloss tokenization method (Spoken language text tokenization : BPE)
- Gloss Dictionary based Tokenization (GDT)
  - Segmented gloss information (linguistic features of sign language, manually annotated by experts and deaf)
- Morpheme based Tokenization
  - Morphological tokenization methods
- Statistical Subword Tokenization (BPE, The size of the vocabulary : 8000)
  - Statistical-based subword tokenization method

## Experiment

### ● Gloss Tokenization Example

- Spoken language Korean text
  - "광주광역시 홍수주의보 발령, 안전에 유의하시기 바랍니다."
  - "아침 최저기온은 17도에서 21도, 낮 최고기온은 19도에서 25도가 되겠습니다."
- Gloss Dictionary based Tokenization (GDT)
  - "광주1, 지역1, 비내리다1#, 차오르다1#, 주의보1, 지시2#, 조심1"
  - "아침1, 꼴찌1, 온도1, 온도1, 17, 21, 낮1, 최고1, 온도올라가다1, 19, 25"
- Morpheme based Tokenization
  - "광주, 1, 지역, 1, 비내리다, 1, #, 차오르다, 1, #, 주의보, 1, 지시, 2, #, 조심, 1"
  - "아침, 1, 꼴찌, 1, 온도, 1, 온도, 1, 17, 21, 낮, 1, 최고, 1, 온도, 올라가다, 1, 19, 25"
- Statistical Subword Tokenization (BPE, The size of the vocabulary : 8000)
  - "\_광주, 1, \_지역, 1, \_비내리다, 1#, \_차오르다, 1#, \_주의보, 1, \_지시, 2#, \_조심, 1"
  - "아침, 1, \_꼴찌, 1, \_온도, 1, \_온도, 1, \_17, \_21, \_낮, 1, \_최고, 1, \_온도올라가다, 1, \_19, \_25"

## Experiment

### ● Gloss Tokenization Example (Version translated from Korean to English)

- Spoken language Korean text
  - "Flood warning issued in Gwangju, please be safe."
  - "Morning lows will be 17 to 21 degrees and daytime highs will be 19 to 25 degrees."
- Gloss Dictionary based Tokenization (GDT)
  - "Gwangju1, region1, rain1#, fill up1#, advisory1, instruction2#, careful1"
  - "Morning1, lowest1, temperature1, temperature1, 17, 21, afternoon1, highest1, temperature\_rise1, 19, 25"
- Morpheme based Tokenization
  - "Gwangju, 1, region, 1, rain, 1, #, fill up, 1, #, advisory, 1, instruction, 2, #, careful, 1"
  - "Morning, 1, lowest, 1, temperature, 1, temperature, 1, 17, 21, afternoon, 1, highest, 1, temperature, rise, 1, 19, 25"
- Statistical Subword Tokenization (BPE, The size of the vocabulary : 8000)
  - "\_Gwangju, 1, \_ region, 1, \_rain, 1#, \_fill up, 1#, \_advisory, 1, \_instruction, 2#, \_careful, 1"
  - "Morning, 1, \_lowest, 1, \_temperature, 1, \_temperature, 1, \_17, \_21, \_afternoon, 1, \_highest, 1, \_temperature\_rise, 1, \_19, \_25"

## Results

- Experimental results SLT & SLR using various gloss tokenization methods

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

- Sign language and spoken language exhibit different characteristics and, consequently, require different tokenization approaches

## Conclusion

### ● Conclusion

- Release a new benchmark dataset

SSL : korean disaster Safety information Sign Language translation benchmark dataset

(Pair Sign language Video, Gloss, Text, Keypoint)

- Disclose the baseline performance of our benchmark dataset
- Significant performance differences based on the gloss tokenization method
- Need for additional research into gloss tokenization methods



**Best regards**