Action-Concentrated Embedding Framework: This is your captain sign-tokening

Hyunwook Yu[1*], Suhyeon Shin[2*], Jungu Heo[3], Hyuntaek Shin[4], Hyosu Kim[5†], Mucheol Kim[6†]
Computer Science and Engineering, Chung-Ang University[1,2,5,6]
UNIVIA Inc.[3,4]
{[1]yu990410, [2]girinssh, [5]hskimhello, [6]kimm}@cau.ac.kr,
[3]heojunku@univia.co.kr,[4]dorim123@gmail.com

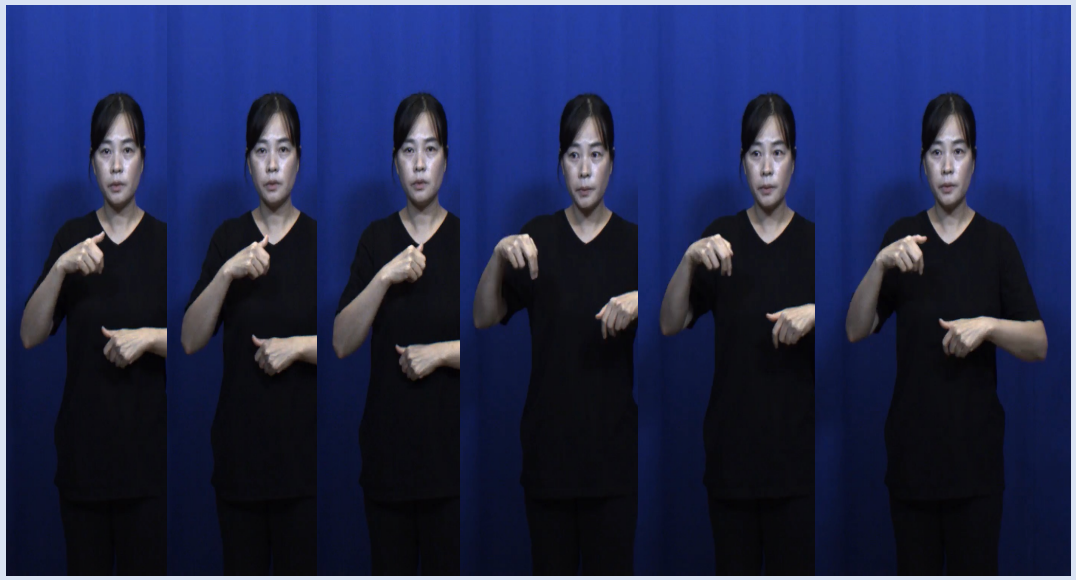
https://sites.google.com/dilabcau/
https://hcslab.cau.ac.kr/

Contents

1. Introduction

Sign language translation

What is Sign language translation?



23시 강원도 춘천시 3명 실종 및 유괴 상황 발생,자녀의 하루 일과와 친한 친구들을 알아두시기 바랍니다.

(At 11 PM in a Asheville, North Carolina, there has been an incident involving the disappearance and abduction of three individuals. Parents are advised to keep track of their children's daily activities and know their close friends.)

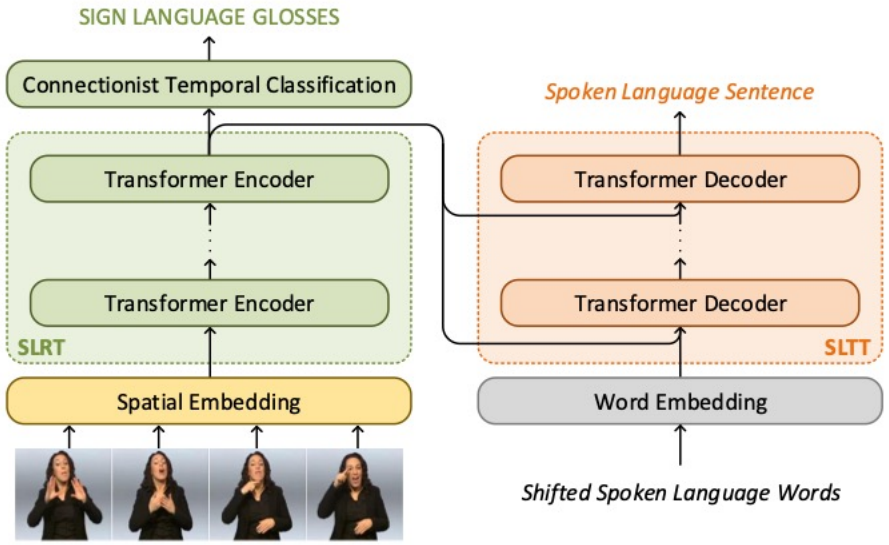
〈Sign language, a language used by deaf people, convey meaning through gestures and facial expressions〉.

〈Natural language conveys and records meaning in the form of letters〉.

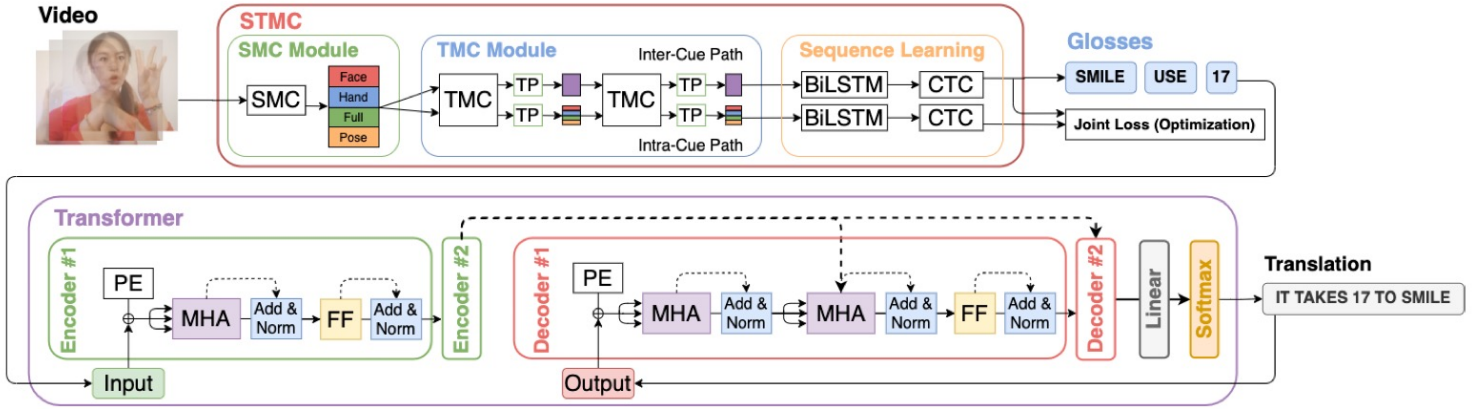
1. Introduction

Sign language translation with Transformer model

Transformer architecture leveraged in a state of the art of sign language translation research



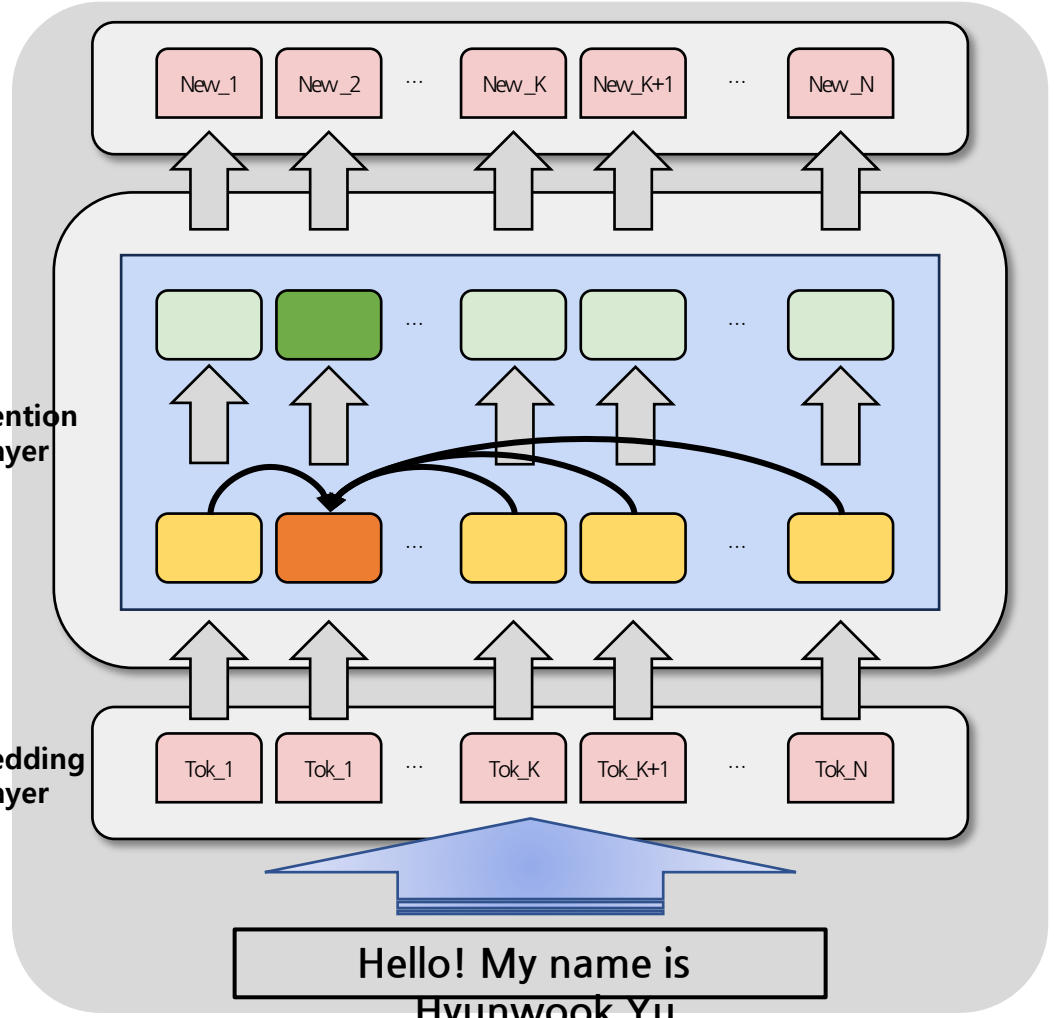
Sign language transformers: Joint end-to-end sign language recognition and translation



Better Sign Language Translation with STMC-Transformer

1. Introduction

Features of the Attention layers that compose Transformer



〈Attention architecture with a focus on context between tokens〉

- Transformer, a Sequence to Sequence model that focuses on contexts between tokens
- Transformer, a model where the tokenize algorithm and embedding layer have a significant impact on performance.

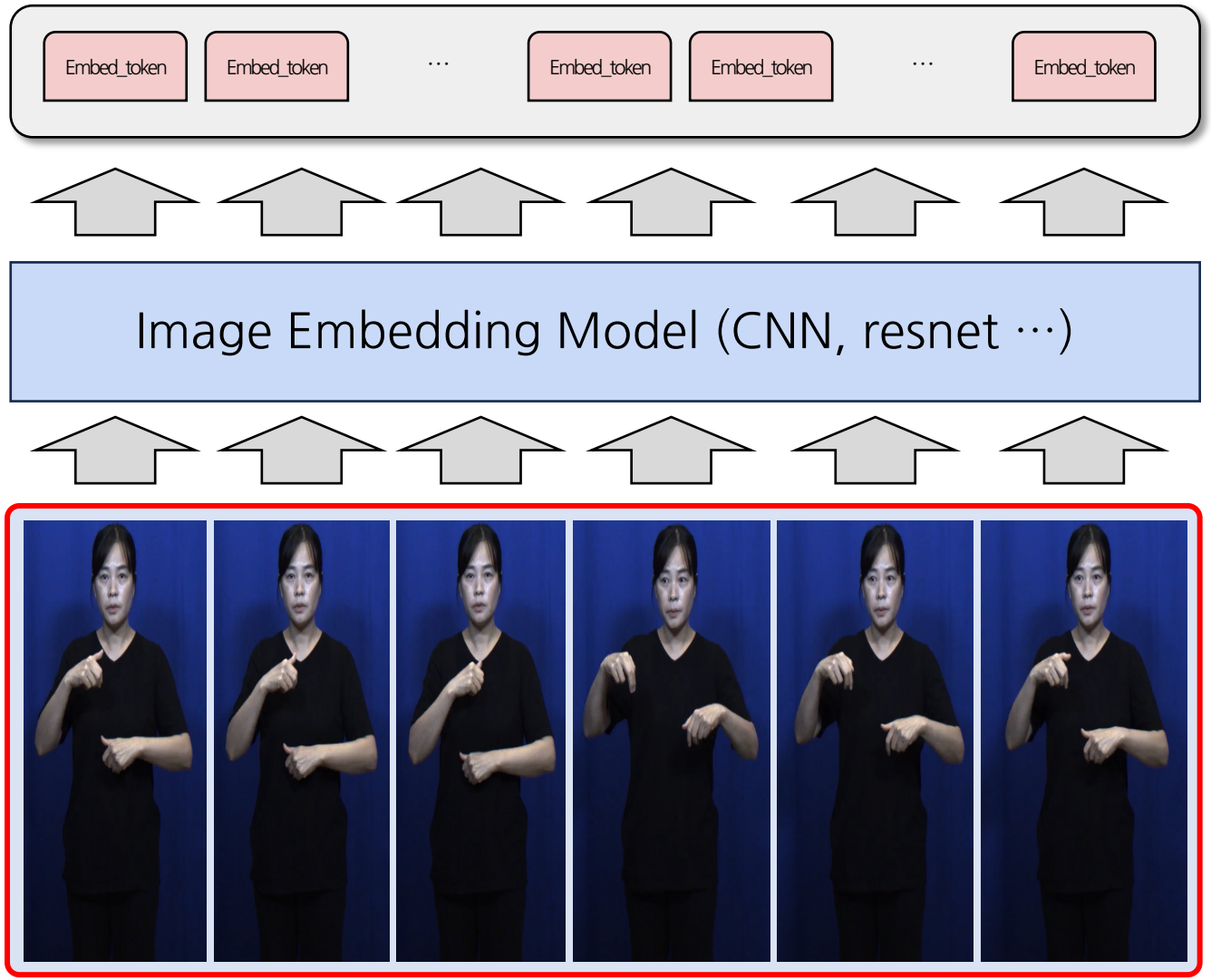
1. Introduction

- [x] How have we been tokenizing sign language and embedding them?

How have we been tokenizing sign language and embedding them?

1. Introduction

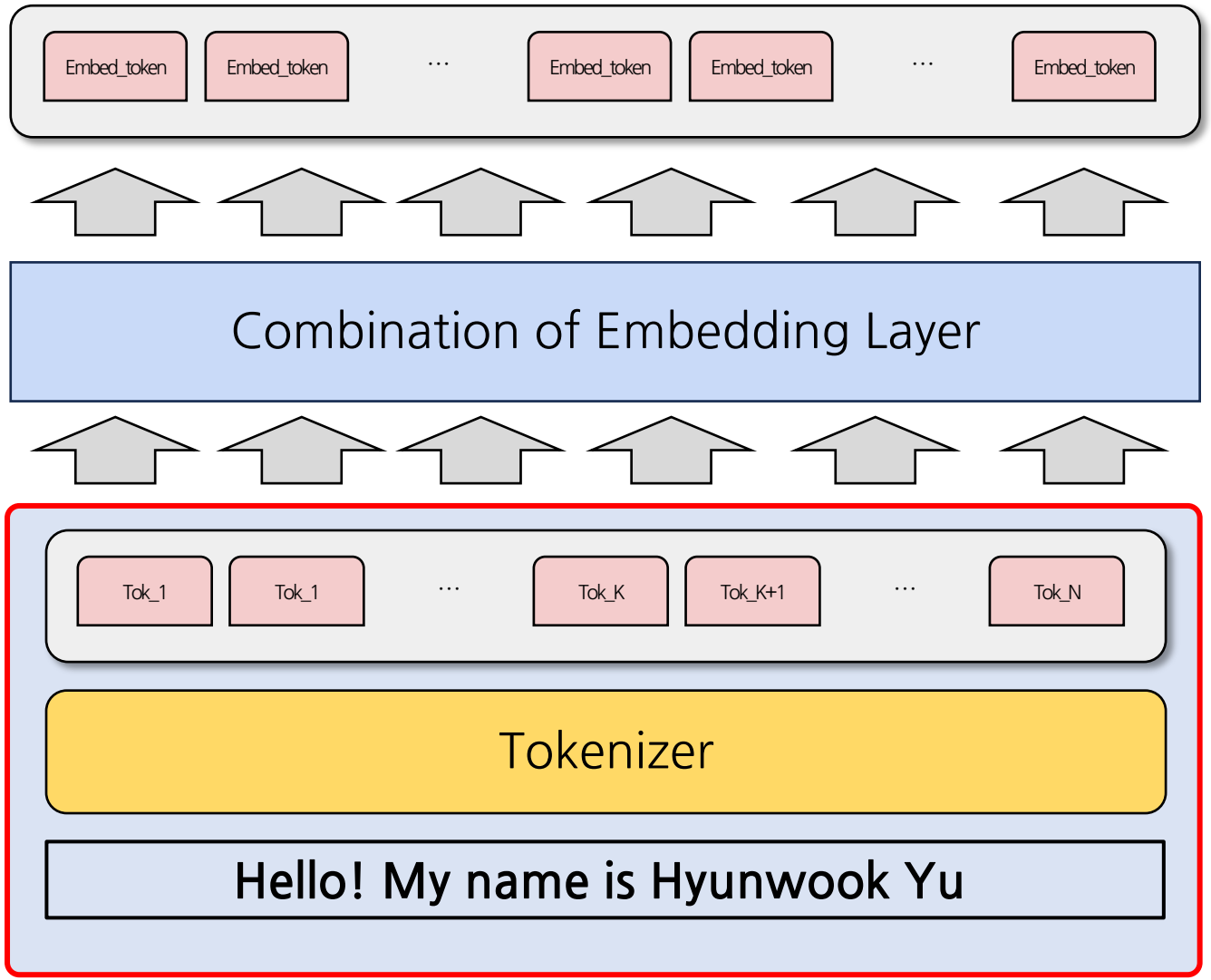
Frame-based token embedding framework



Consider frame(image) as token

1. Introduction

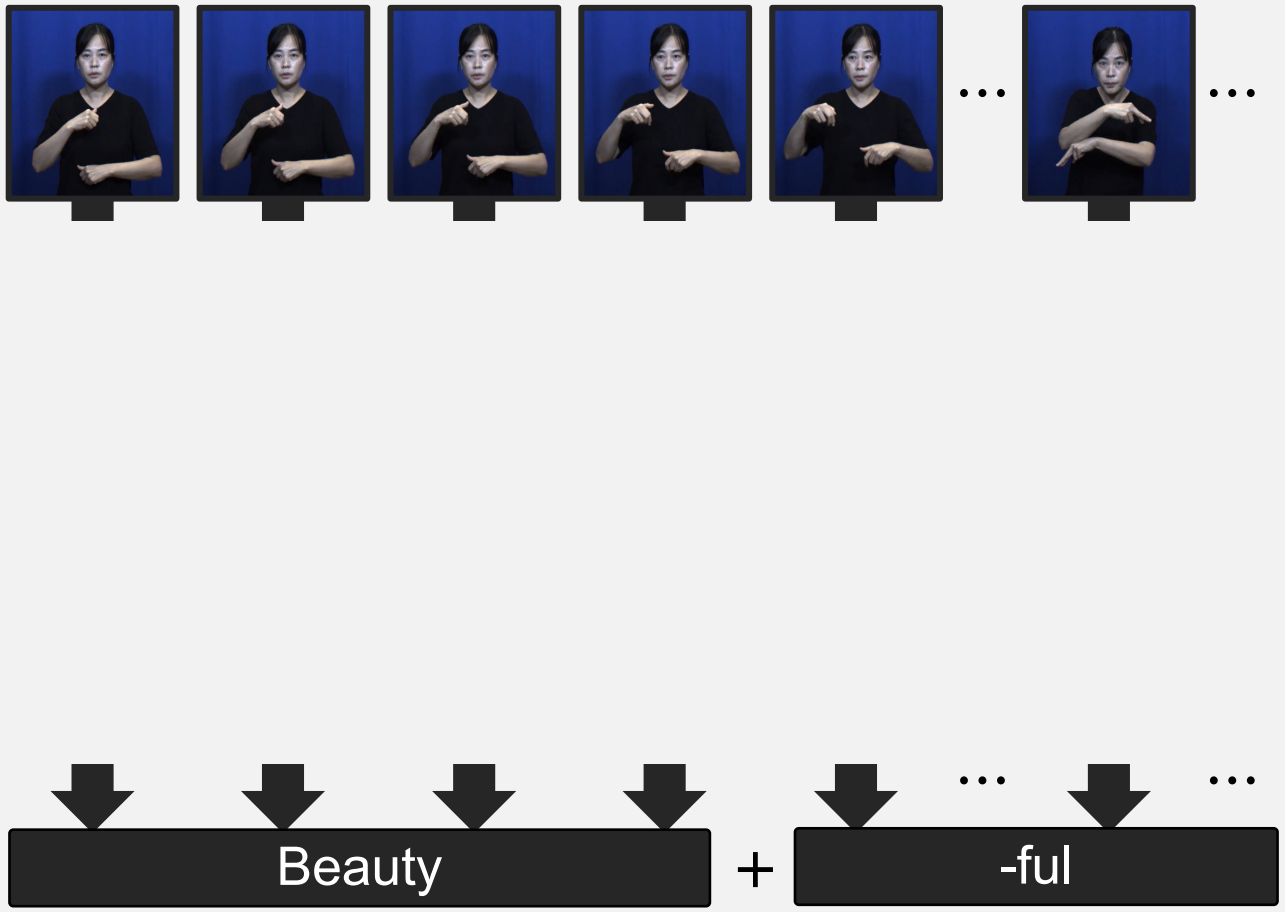
Tokenization framework for language models using the Attention



Consider the smallest unit with meaning (word, morpheme, etc.) as a token

1. Introduction

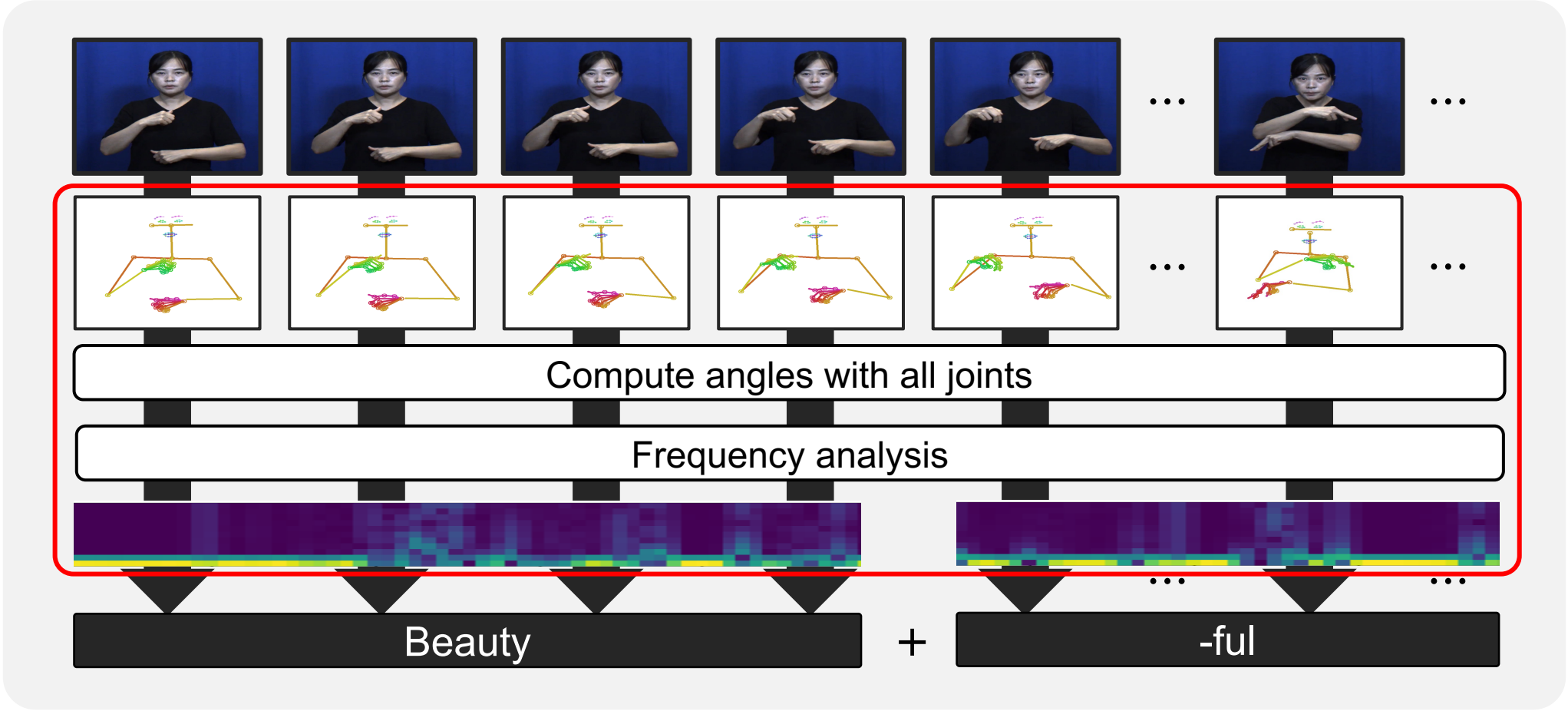
A linguistic system of sign languages that uses combinations of actions (frames) to form meanings.



- A tokenization unit of a sign language that should be a set of frames, not an frame.
  - ➢ Image is just alphabet(or stroke)
- A frame-based sign language embedding framework that overlooks the linguistic nature of sign language
  - ➢ Difficulty reflecting the complex linguistic system of sign languages, where each action represents a stem and affix.
- Frame-based tokenization increases the number of tokens, which increases the time complexity of the transformer model, which is proportional to the square of the number of tokens

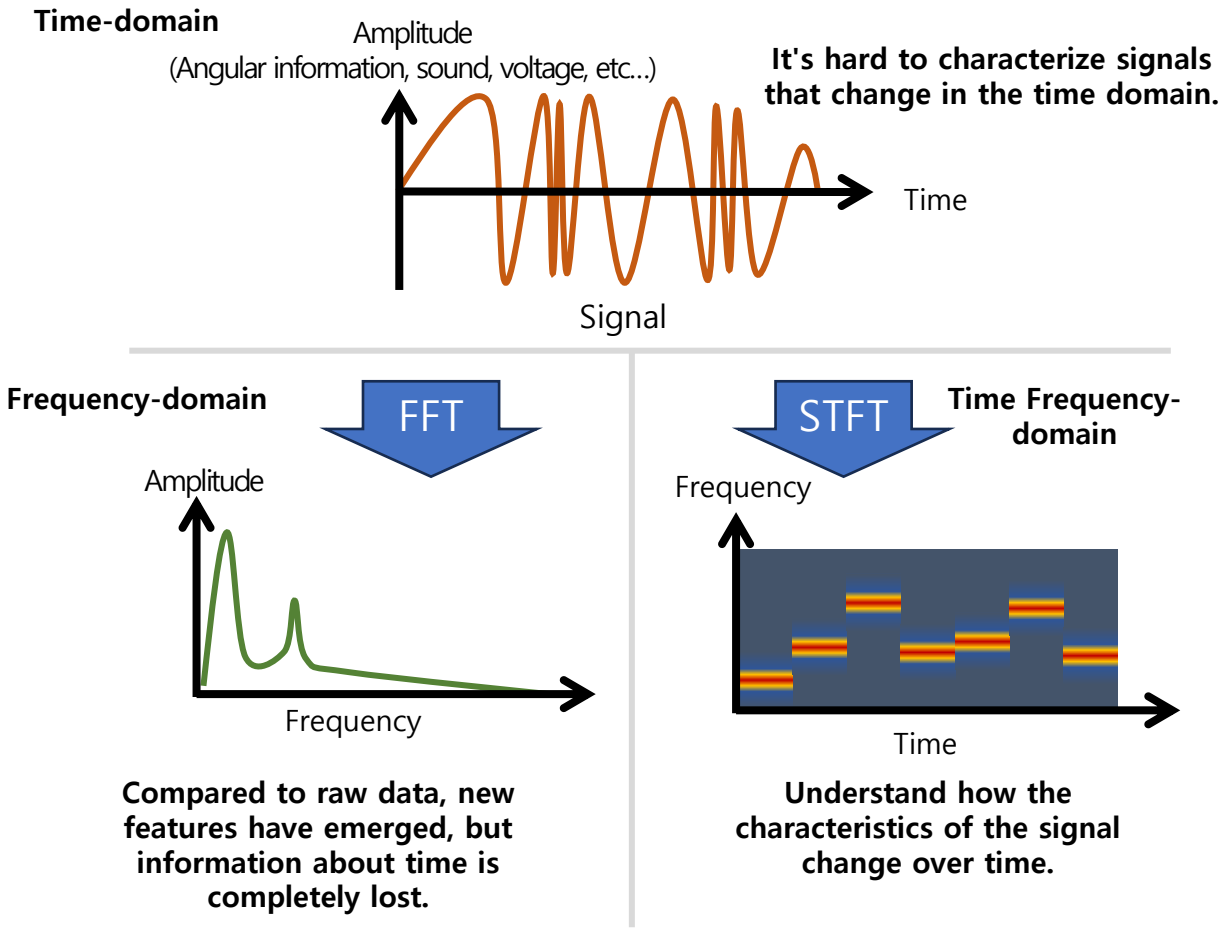
1. Introduction

A linguistic system of sign languages that uses combinations of actions (frames) to form meanings.



1. Introduction

Apply short-time Fourier transform into sign language embeddi ng



Advantages of applying STFT to sign languages

- View sign language as an action over time rather than a pose(frame) over time.
- Sliding window method allows tokenization of sign language into semantic units
  - ➢ Action over time with linguistic meaning can be used as a sign language token.
  - ➢ Contextual meaning between windows can be interpreted through the Attention layer

Contribution

■ Action-concentrated embedding (ACE) framework

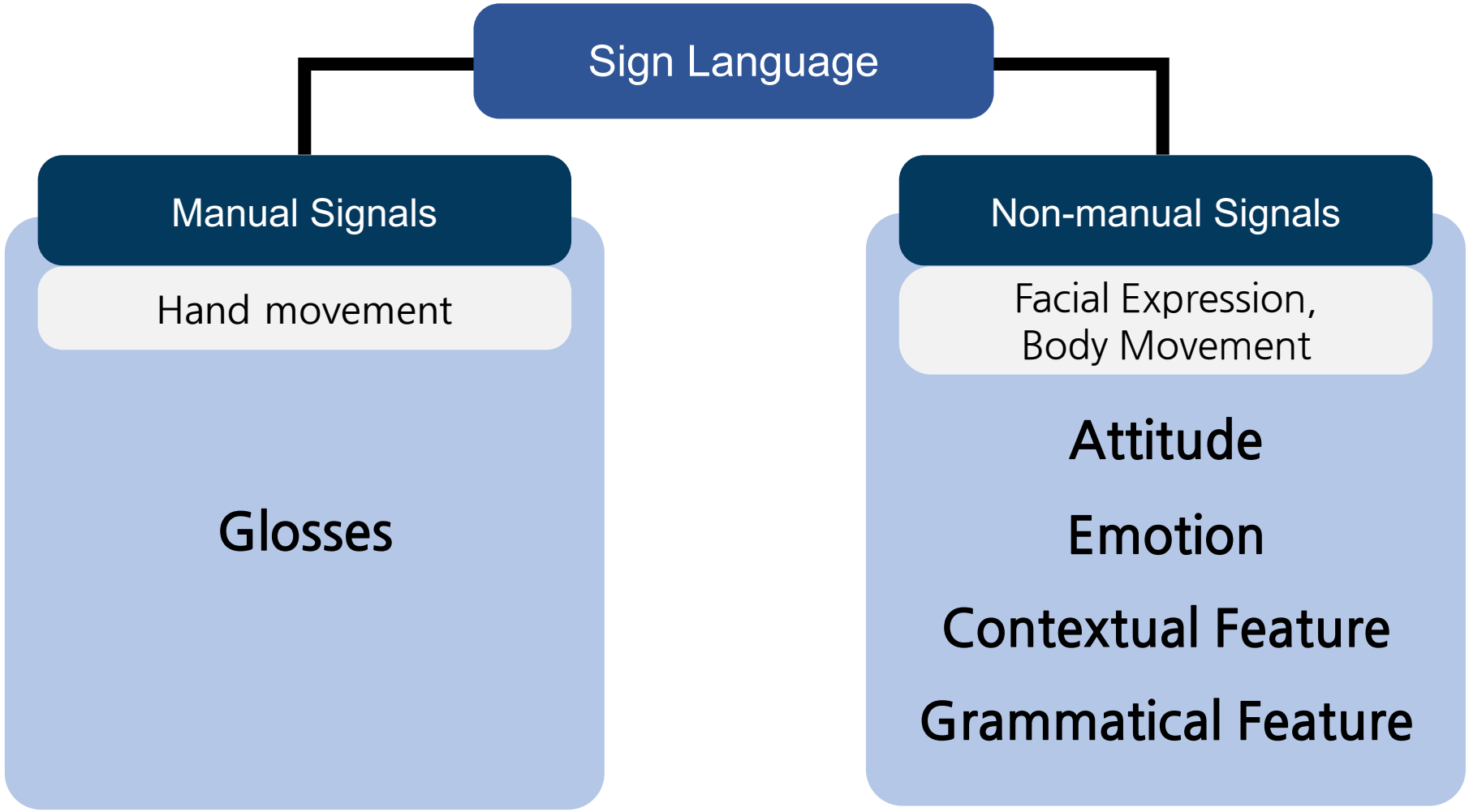
➢ Focuses on changes in the speaker's hands, face, and body motions

- ✓ The ACE framework captures the posture changes of body parts over time and provides token embeddings using the STFT.
- ✓ The proposed framework understands each action in the sign as a linguistic system of stems and affixes.
- ✓ The proposed approach effectively represents various actions of sign language and facilitates the understanding of sign language based on patterns of part-specific movement changes.
- ✓ The proposed framework improves sign language translation using a transformer-based model.

Contents

2. Methodology

Components of Sign language



2. Methodology

ACE Architecture

Sign Language Video

$Frame_1$ ... $Frame_i$ ... $Frame_T$

Keypoint Extraction

Edge-pair Construction

Angular Information Computation

$V: v_1 \quad v_2 \quad \dots \quad v_i \quad \dots \quad v_T$

Short-time Fourier Transform

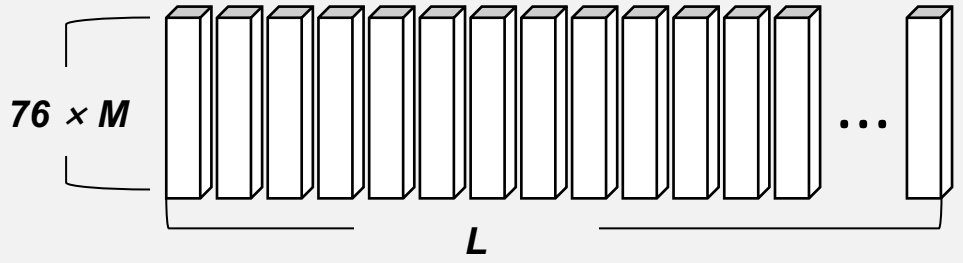
Sliding Window Tokenization

Overlapped window

Token Embedding with Frequency Component

DFT for each edge-pair

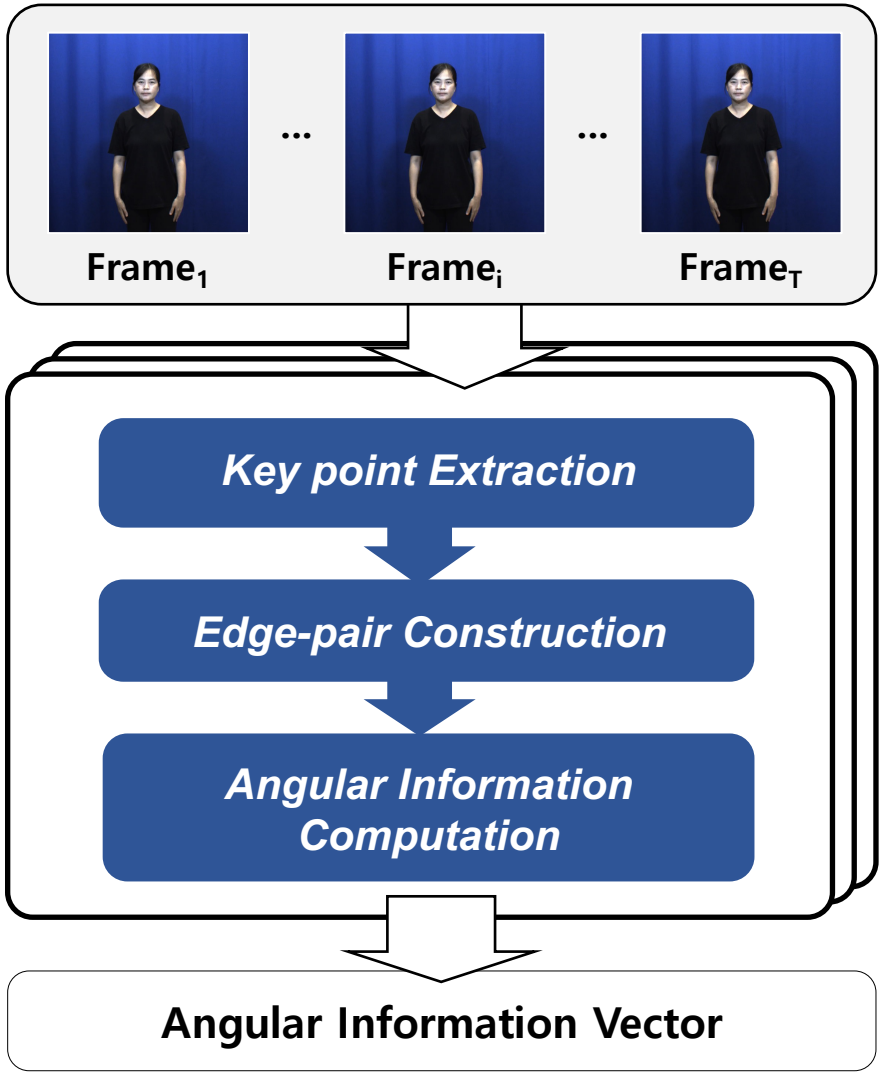
Get Magnitudes

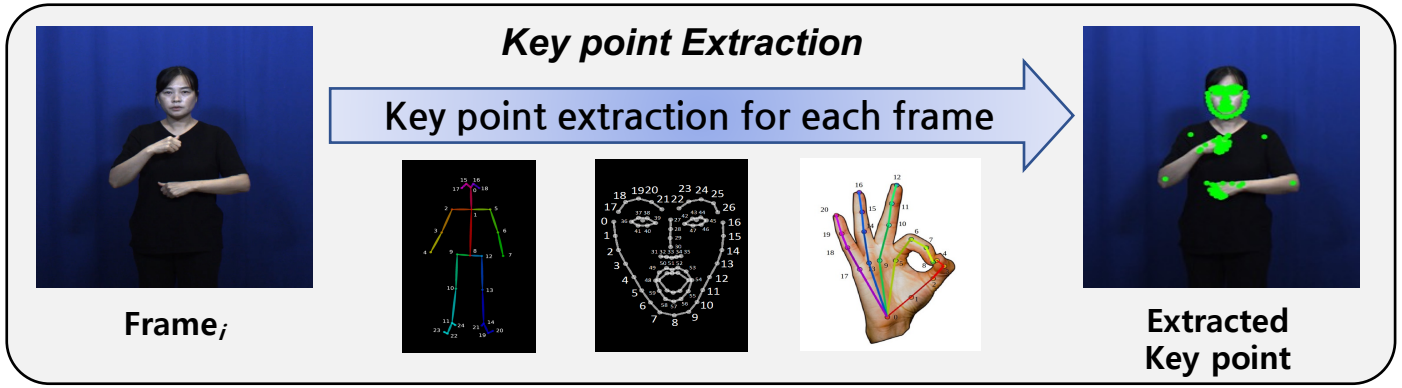


L Tokens

2. Methodology

Transforming the Frame to Angular Information



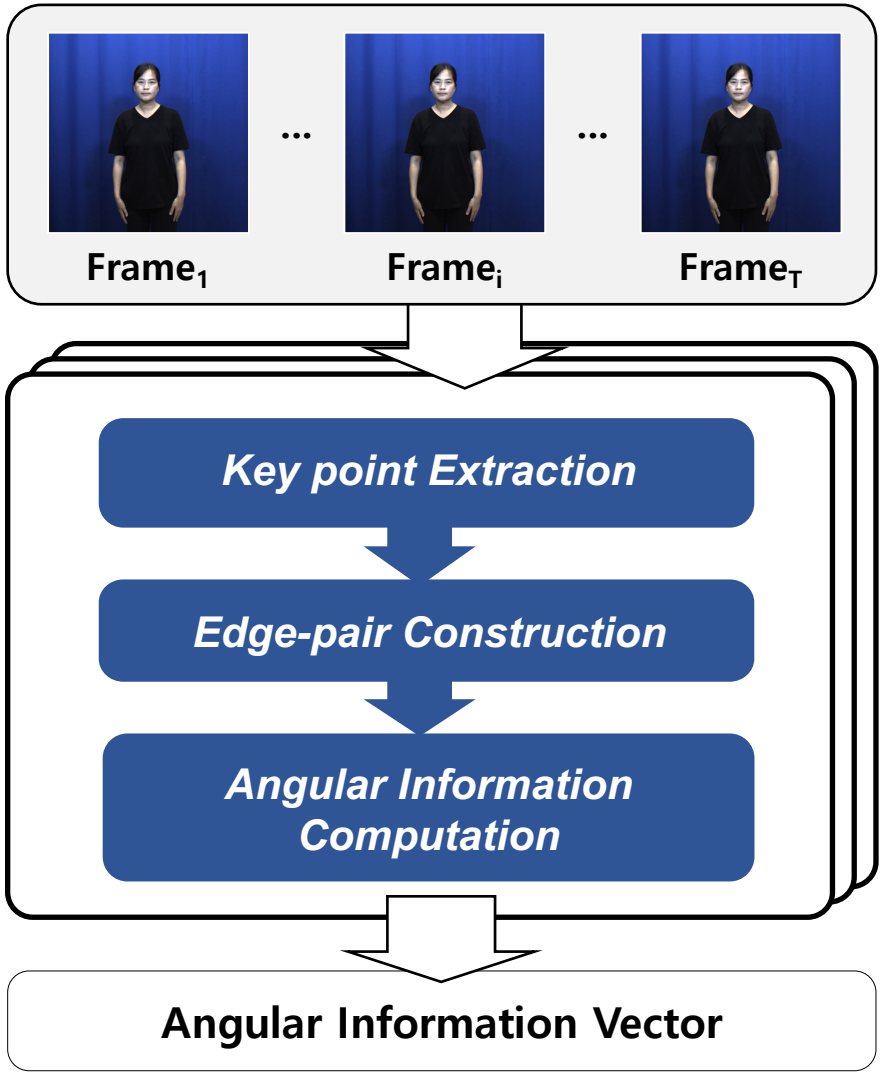


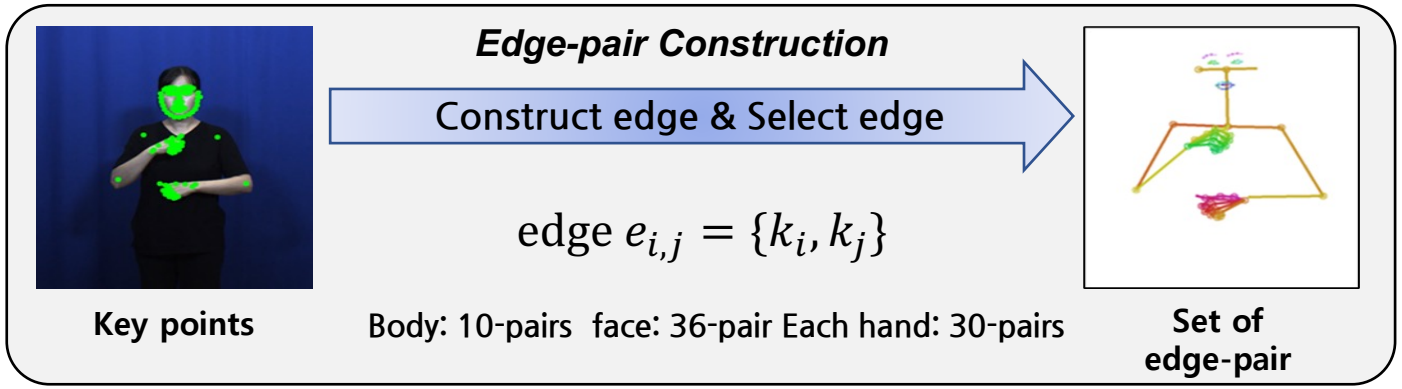
- Key points from Openpose
  - Body(BODY_ 25): 25ea
  - Each hand: 21ea
  - Face: 70ea
- Key points using this work
  - Body: 12ea
  - Each hand: 21ea
  - Face: 70ea
- Excludes the lower half of the body, which is not used to express sign language.
- The j-th keypoint k_j is represented by the x, y coordinates of the frame.

$$k_j = (x_j, y_j)$$

2. Methodology

Transforming the Frame to Angular Information

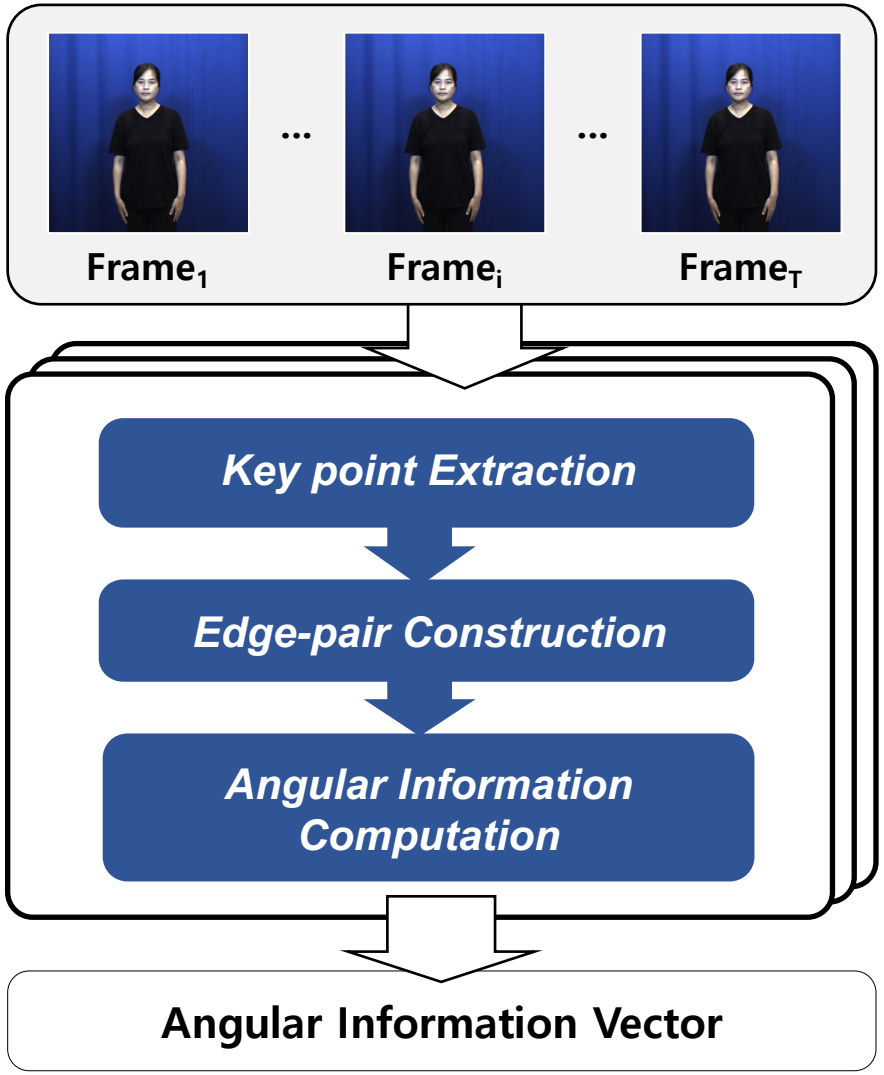


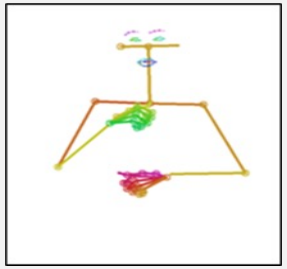


- An edge is composed of two key points (ki) that are core to the body's structure (often neighboring key points).
- In total, 76 edge-pairs were constructed to represent major joints such as elbows, wrists, and shoulders, as well as gestures, facial expressions, and hands such as fingers, eyebrows, and lips.

2. Methodology

Transforming the Frame to Angular Information





Angular Information Computation

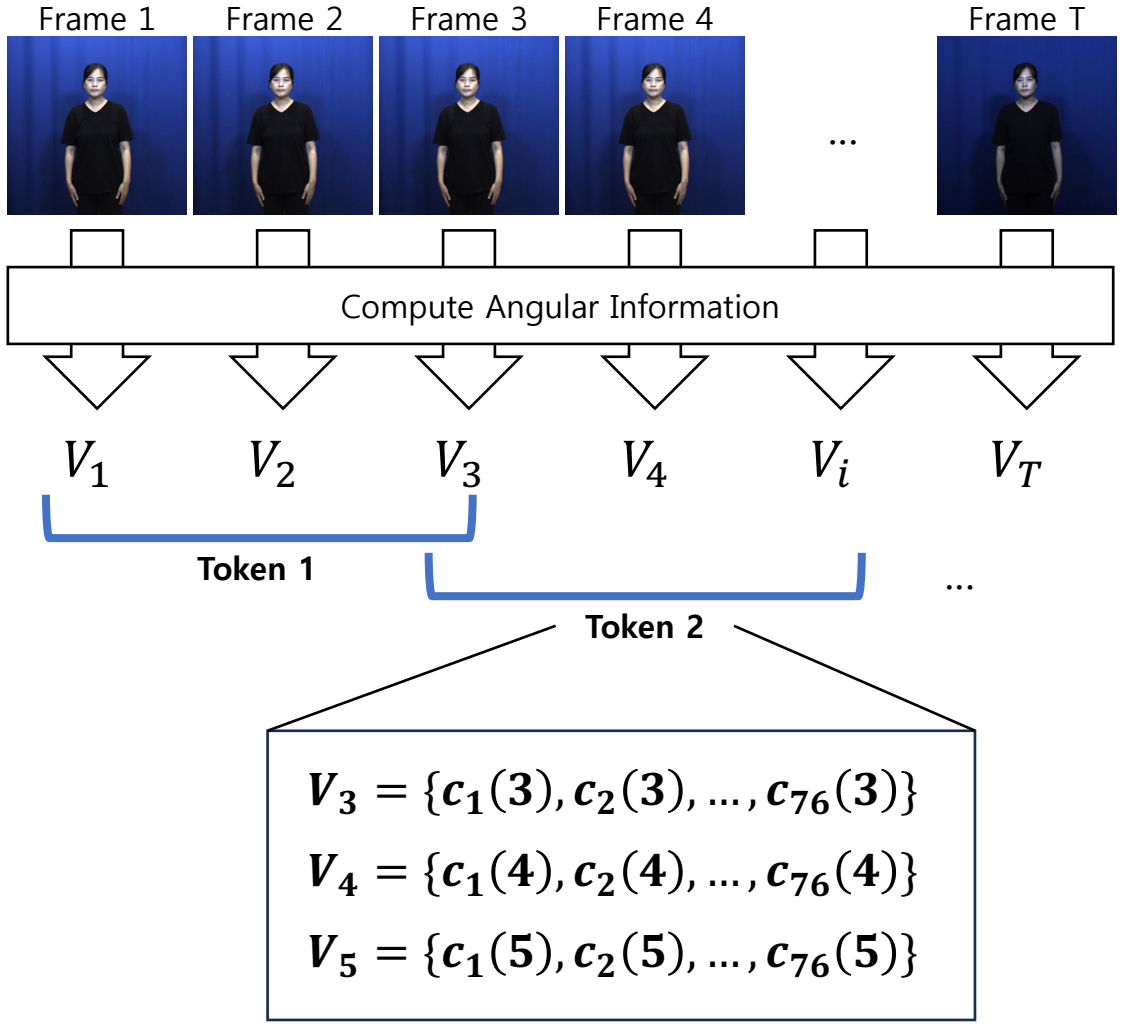
compute cosine value

$$c = \frac{e_{i,j} \cdot e_{k,l}}{\|e_{i,j}\| \|e_{k,l}\|}$$

- Angular information serves as a pivotal bridge between physical gestures and linguistic meaning.
- For i-th input video frame, a set of cosine values denoted $\boldsymbol{V_i}$, where $\boldsymbol{V_i} = \{\boldsymbol{c_1(i)}, \boldsymbol{c_2(i)}, \dots, \boldsymbol{c_{76}(i)}\}$.
- ACE outputs an angular information set V, which contains the cosine values obtained from each video frame (i.e $\boldsymbol{V} = \{\boldsymbol{V_1}, \boldsymbol{V_2}, \dots, \boldsymbol{V_T}\}$).
- The cosine values of edge pairs offer a snapshot of the signer's morphology at that particular moment, encoding both manual and non-manual signals.

2. Methodology

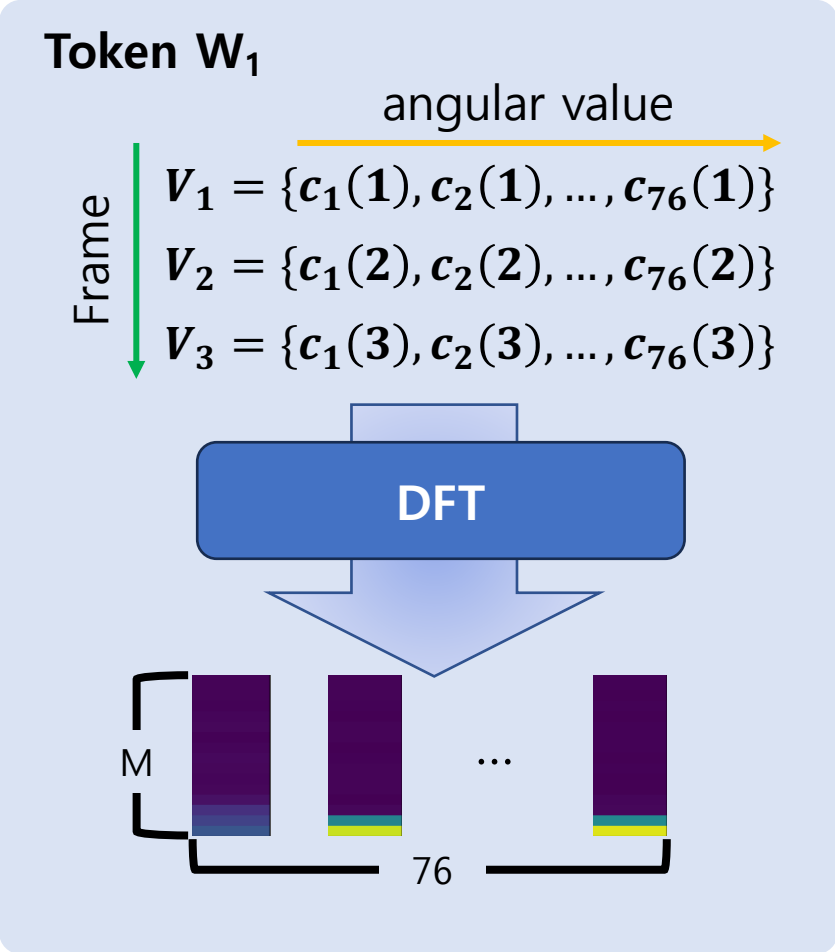
Sliding window-based tokenization



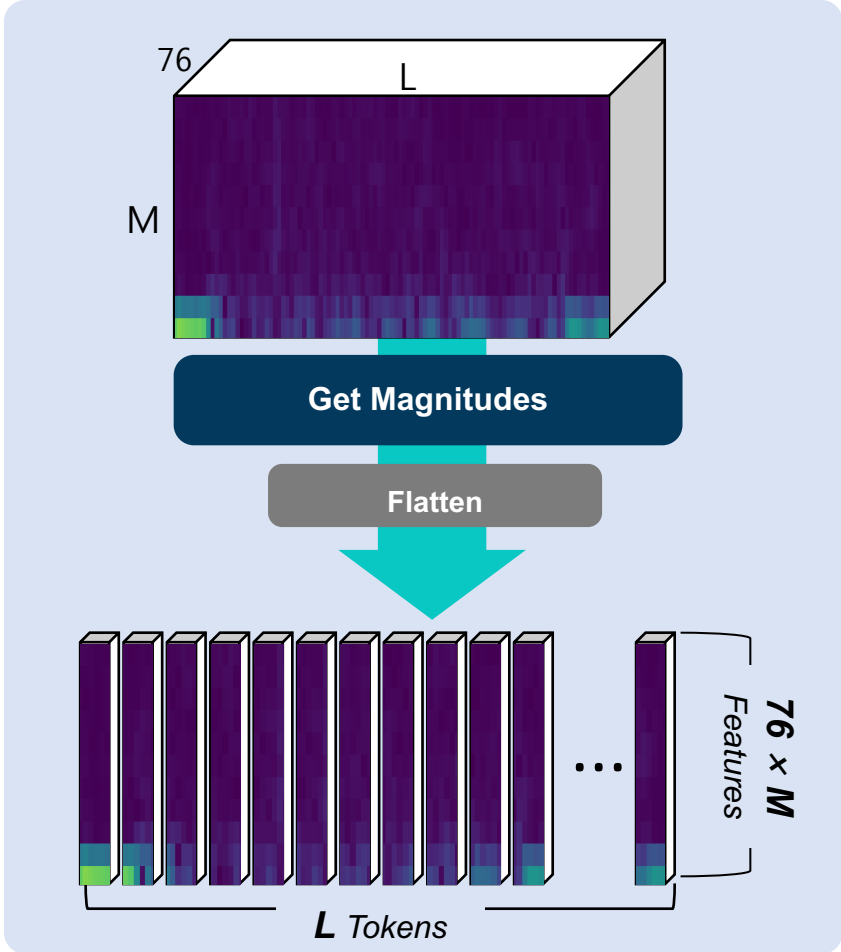
- The sliding window method requires a window size N and a window overlap parameter O.
- The i-th token W_i consists of N consecutive angular information, which can be represented as $W_i = \{V_j | H \times (i-1) + 1 \leq j \leq H \times (i-1) + N\}$.
- Here, H is the window interval, which is N minus O.
- The window size N determines the frequency resolution of the DFT.

2. Methodology

Token embedding with frequency components



Perform for all tokens



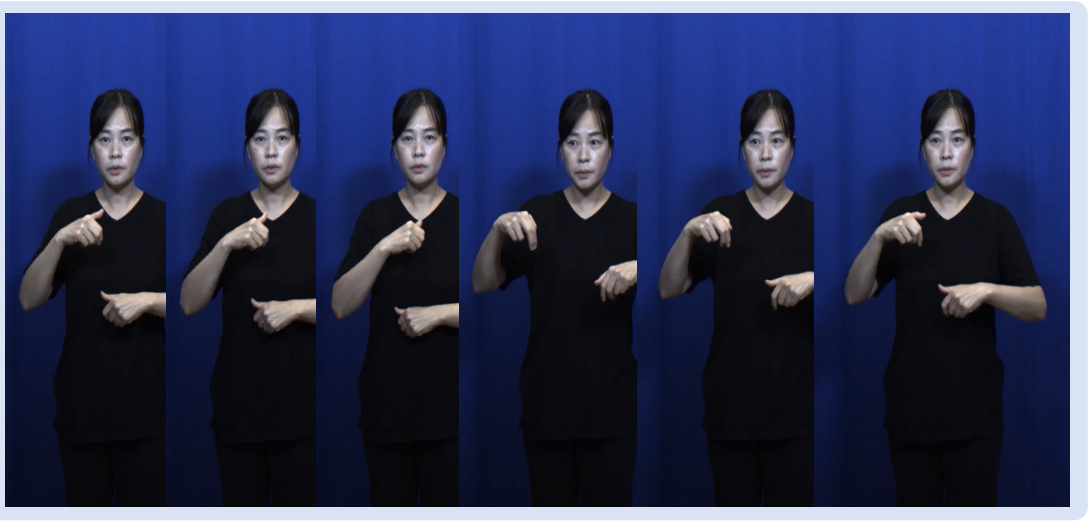
Contents

Setup

■ Dataset

➢ Sign language for disaster safety information[1)]

- ✓ AI datasets extracted from source data (sentences, sign language videos) containing disaster safety information, including morphological and non-morphological information and key point data.
- ✓ Contains sentences with matching Korean grammar, morphemes, temporal information about gestures, and non-numerical information such as facial expressions.



23시 강원도 춘천시 3명 실종 및 유괴 상황 발생,자녀의 하루 일과와 친한 친구들을 알아두시기 바랍니다.

(At 11 PM in a Asheville, North Carolina, there has been an incident involving the disappearance and abduction of three individuals. Parents are advised to keep track of their children's daily activities and know their close friends.)

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

3. Experiments

Setup

■ Metric

➢ BLEU-4

✓ A high BLEU-4 score indicated that the machine output closely resembled that of a human, focusing on the accuracy of the text produced.

➢ ROUGE-L

✓ ROUGE-L emphasized the coverage or thoroughness between the translated and reference text.

➢ METEOR

✓ METEOR metric provided a comprehensive and nuanced assessment, which included the accuracy and comprehensiveness of the translated material.

■ Setup

➢ Hyperparameter of experiments

✓ Three transformer layers, each with a dimension of 256 and 4 heads, vocabulary size of 22,000 for decoder,
✓ Train 50 epochs, a dropout rate of 0.1, 32 batch size with Nvidia RTX 3090 GPU, Adam optimizer with a learning rate of 0.001

➢ Data proportion

✓ Of the 160,677 training samples (excluding incomplete data), 95% was allocated for training and 5% for testing

3. Experiments

Exploring the Impact on Sign Language Translation

➢ RQ1: To what extent does increasing overlap size O influence the ability to represent sign language from diverse analytical perspectives?

N	O	Avg. Token length	BLEU-4	ROUGE-L	METEOR
30	20	44.18	36.72	40.10	41.44
30	10	22.36	32.36	32.83	38.24
30	5	17.92	29.67	30.23	36.40
30	0	15.11	26.86	26.96	33.99
20	15	89.83	38.21	40.69	42.65
20	10	45.19	36.44	39.02	41.56
20	5	30.23	33.18	34.68	38.91
20	0	22.82	30.91	31.09	37.32
15	10	90.83	38.80	41.29	42.74
15	5	45.64	36.73	39.14	41.32
15	0	30.65	32.53	34.12	38.29

Table 1: Performance metrics for various window sizes and overlap ratios.

3. Experiments

Exploring the Impact on Sign Language Translation

➢ **RQ2: What size of window N maximizes the quality of sign language representation, balancing the need for detail and sufficient context?**

N	Overlap Ratio	Sentence Information Size	BLEU-4	ROUGE-L	METEOR
80	75%	61,765	32.71	34.06	38.40
60	75%	65,162	35.60	36.72	40.15
50	74%	64,332	35.63	37.64	40.71
40	75%	68,924	37.00	38.90	41.60
30	76.66%	76,470	37.87	40.28	42.20
20	75%	75,090	38.21	41.00	42.79
15	73.33%	68,965	38.57	41.79	42.64

Table 2: Evaluation metrics for different window sizes while maintaining similar overlap ratios. Results are averaged over three different random seeds to account for variability.

✓ Increasing the overlap size improved the interpretation of sign language patterns, although an equilibrium with the window size was crucial to avoid losing the meaning of the message.

*sentence information size is derived by multiplying the average token length with the token embedding dimension.

3. Experiments

Impact of Body Part Features on ACE

➢ The influence of various body part features on the performance of ACE

Model	Metric	Features			
		Hands	Hands, Body	Hands, Face	Hands, Body, Face
Baseline	BLEU-4	29.65	30.31	30.43	30.82
	ROUGE-L	32.87	33.08	32.9	33.58
	METEOR	36.6	37.37	37.42	37.21
ACE	BLEU-4	31.97	35.23	33.16	36.78
	ROUGE-L	36.14	38.86	36.56	38.39
	METEOR	37.54	40.44	38.65	41.51

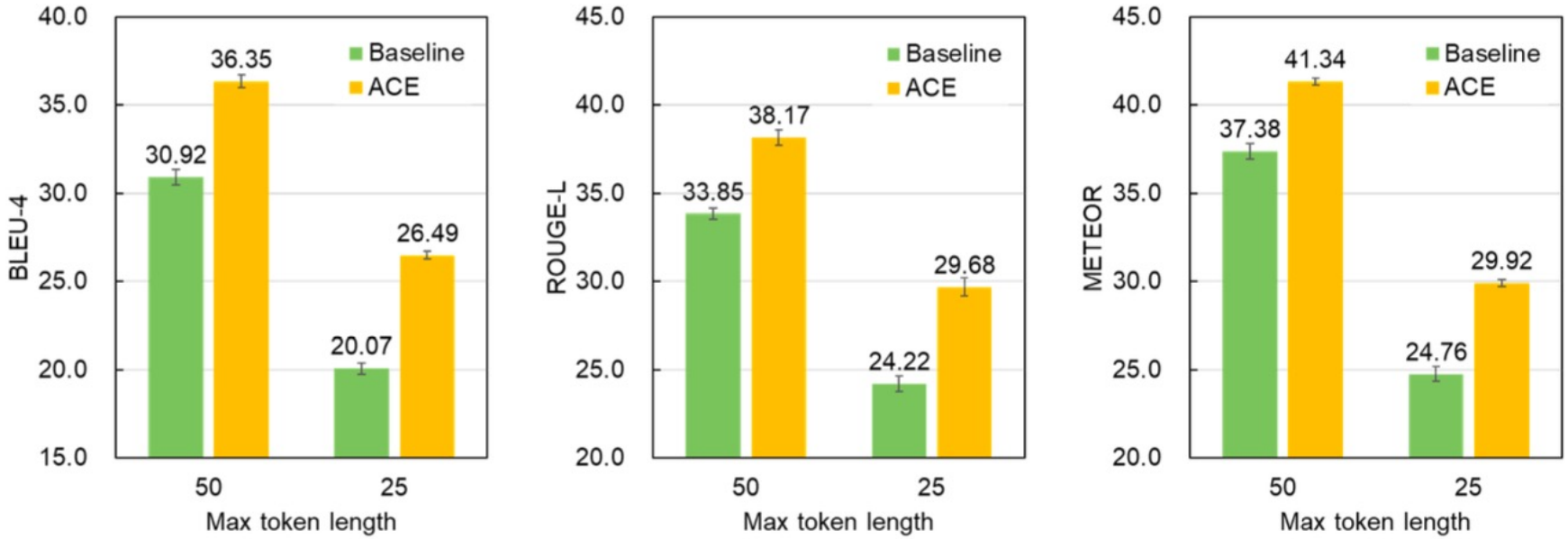
Table 3: Influence of body part features on the performance of ACE and the baseline model (Ko et al.,2019).

- ✓ When confined to the "Hands" feature, ACE's potential diminished because "Hands" alone only offers a constrained range of angular data, overlooking intricate relationships with other body parts (such as the torso and shoulders)
- ✓ This result also highlighted the essential role of dynamic body movements in advancing sign language translation algorithms.

3. Experiments

Comparison with Prior Work

- Comparison the machine translation performance of the proposed and existing models in a constrained environment.



Comparison of automatic translation performance metrics for various maximum token lengths between the proposed ACE and the baseline by (Ko et al., 2019). Results are averaged over three different random seeds to account for variability.

3. Experiments

Examples

Ko et al. method

ENG

PRED

Road explosion near Cheonan, Chungnam, please be aware of the safety of nearby residents.

TRUE

An explosion occurred in the Cheonan area of Chungnam. Please be careful not to get hit by fallen objects on the road.

KOR

PRED

충남 천안시 인근 도로 폭발사고 발생, 인근 주민은 안전에 유의하시기 바랍니다.

TRUE

충남 천안 지역 폭발사고 발생. 도로에 떨어진 물건에 사고가 발생하지 않도록 주의바랍니다.

Our method

ENG

PRED

Heavy rain warning in effect in Jeju at 11:40 today, please evacuate landslide risk areas, refrain from going out, etc.

TRUE

Heavy rain warning in effect in Jeju at 11:40 today, you need to evacuate landslide risk areas, refrain from going out, etc.

KOR

PRED

오늘 11시 40분 제주 호우경보 발효, 산사태 위험지역 대피, 외출 자제 등 안전에 주의 바랍니다

TRUE

오늘 11시 40분 제주 호우경보 발효. 산사태 위험지역 대피, 외출 자제 등 안전에 주의하세요.

Contents

4. Conclusion

Conclusion

- By focusing on the dynamics of body movements and leveraging the STFT for tokenization, ACE displayed promising results in sign language-to-sentence translation.
- The experimental results confirmed ACE's adaptability and efficiency regarding the number of tokens. Especially under the restriction of a limited number of tokens, the increase in BLEU-4 and ROUGE-L metrics was remarkable.
- This study provided a strong and comprehensive depiction of sign language, while also establishing a solid framework for future research.

Thank you
감사합니다.