Journal of Korea Multimedia Society Vol. 25, No. 10, October 2022(pp. 1535-1545)
https://doi.org/10.9717/kmms.2022.25.10.1535

수어 동작 키포인트 중심의 시공간적 정보를 강화한 Sign2Gloss2Text 기반의 수어 번역

김민채[†*], 김정은[†††], 김하영[†††]

Sign2Gloss2Text-based Sign Language Translation with Enhanced Spatial-temporal Information Centered on Sign Language Movement Keypoints

Minchae Kim[†*], Jungeun Kim[†††], Ha Young Kim[†††]

ABSTRACT

Sign language has completely different meaning depending on the direction of the hand or the change of facial expression even with the same gesture. In this respect, it is crucial to capture the spatial-temporal structure information of each movement. However, sign language translation studies based on Sign2Gloss2Text only convey comprehensive spatial-temporal information about the entire sign language movement. Consequently, detailed information (facial expression, gestures, and etc.) of each movement that is important for sign language translation is not emphasized. Accordingly, in this paper, we propose Spatial-temporal Keypoints Centered Sign2Gloss2Text Translation, named STKC-Sign2Gloss2Text, to supplement the sequential and semantic information of keypoints which are the core of recognizing and translating sign language. STKC-Sign2Gloss2Text consists of two steps, Spatial Keypoints Embedding, which extracts 121 major keypoints from each image, and Temporal Keypoints Embedding, which emphasizes sequential information using Bi-GRU for extracted keypoints of sign language. The proposed model outperformed all Bilingual Evaluation Understudy(BLEU) scores in Development(DEV) and Testing(TEST) than Sign2Gloss2Text as the baseline, and in particular, it proved the effectiveness of the proposed methodology by achieving 23.19, an improvement of 1.87 based on TEST BLEU-4.

Key words: Deep Learning, Sign Language Translation, Human Pose Estimation, Keypoints, Sign2Gloss2Text, Transformer, Spatial-temporal Structure Information

1. 서 론

수어는 청각장애인의 모국어로써 의사소통과 정보 공유를 위한 중요한 수단이다[1]. 수어를 이해하고 이를 구어로 번역하는 것은 비장애인과 청각장애인을 소통할 수 있게 하며, 이를 통해 청각장애인들의 사회 참여도를 높일 수 있다. 수어는 손의 모양과 움직임을 통해 언어적 의미가 표현되며, 얼굴의 표정

*These authors contributed equally.
※ Corresponding Author : Ha Young Kim, Address: (03722) Yonsei-ro 50, Seodaemun-gu, Seoul, Republic of Korea, TEL : +82-2-2123-4194, FAX : +82-2-2123-8654, E-mail : hayoung.kim@yonsei.ac.kr
Receipt date : Sep. 29, 2022, Revision date : Oct. 17, 2022
Approval date : Oct. 17, 2022
† Graduate School of Information, Yonsei University (E-mail : inthemingcha@yonsei.ac.kr)
††† Dept. of Artificial Intelligence, Graduate School, Yonsei University (E-mail : jekim5418@yonsei.ac.kr)
††† Graduate School of Information, Yonsei University
※ This research was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT; Ministry of Science and ICT) (2020R1F1A1071527).

을 통해 감정과 의도가 표현된다[2]. 따라서, 손의 모양, 위치, 얼굴의 표정, 팔의 위치 등이 종합적으로 고려돼야 한다. 예를 들어, 오른손의 엄지손가락과 집게손가락을 붙인 상태에서 코 왼쪽에서 오른쪽으로 붙인 손가락을 약하게 튕기며 웃는 동작은 한국어 수화 중 '신난다.'를 나타낸다[3]. 또한, 수어는 구어의 문법 체계와 다른 고유한 문법 체계를 가지고 있다[4]. 따라서, 비장애인들이 수어를 이해하고 표현하는 것은 새로운 언어 체계를 배우는 것만큼 어려운 일이다. 이러한 수어를 구어로 번역하기 위해 딥러닝을 활용한 많은 연구가 진행되었다[5, 6].

수어 번역을 위해 [7]에서 처음으로 RWTH-PHOENIX-Weather 2014 T(PHEONIX-2014 T) 데이터셋을 제시하였으며, 인코더(encoder)와 디코더(decoder)가 Recurrent Neural Network(RNN)으로 이루어진 seq-to-seq 번역 모델인 Sign2Text를 제시하였다. Sign2Text는 주어진 연속적인 수어 비디오를 구어(독일어)로 직접 번역하도록 설계된 모델이다. [7] 논문이 제시된 후 PHOENIX-2014 T를 이용한 연구들과 Sign2Text 모델의 성능을 비교하는 많은 연구가 진행되었다[8, 9]. 자연어처리 분야의 성능 발전에 혁신을 일으킨 Transformer 모델[11]이 등장한 이래로, 최근에는 이를 활용한 Sign2Gloss2Text[10]와 같은 모델이 제시되었다. Sign2Gloss2Text는 Sign2Text와 마찬가지로 인코더-디코더 구조로 이루어진 vanilla Transformer 기반 모델이다. 그러나, 수어 이미지에서 구어로 직접 번역하는 Sign2Text와 달리, Sign2Gloss2Text의 인코더는 연속적인 수어 이미지에 대한 글로스(Gloss)로의 예측인 수어 인식(Continuous Sign Language Recognition, CSLR)을, 디코더는 예측된 글로스들로부터 구어로의 번역(Translation)을 수행한다.

그러나, Sign2Gloss2Text 모델은 수어자의 전체 이미지를 입력(input)으로 하며, 수어를 구성하는 중요한 요소 중 하나인 손 모양 및 위치, 몸의 위치나 수어자의 표정 등이 잘 반영되기 힘들다. 수어자의 포괄적인 특징과 세부적인 정보를 동시에 파악하는 것이 중요한 이유는 표정이나 손의 모양을 통해 비슷한 수어 동작 간의 구분에 더 도움을 줄 수 있기 때문이다. 이를 바탕으로, 본 연구에서는 수어자의 전체적인 이미지와 더불어 수어자의 상반신과 양손의 모양 및 얼굴 표정을 파악하기 위해 human pose estimation으로 121개의 키포인트들(keypoints)을 추출해 순차적 특징을 임베딩하는 모듈(Spatial-temporal Keyponts Embedding)과 Sign2Gloss2Text를 결합한 번역 모델인 STKC-Sign2Gloss2Text를 제안한다. 이전 연구 중에도 키포인트 정보를 사용한 연구들이 존재한다 [12,13]. [12]의 경우, Transformer 등장 이전에 자주 활용되던 seq-to-seq 구조를 사용하였으며, 자체적으로 수집한 데이터인 KETI(Korea Electronics Technology Institute) 수어 데이터셋에 대해서 키포인트를 추출해 실험을 진행하였다. [13]은 Sign2Gloss2Text 모델을 기반으로 하며 수어자의 키포인트들을 사용하였으나, 이를 얼굴 및 손의 이미지를 자르는 데에만 사용하였다. 이들 선행연구와는 다르게 본 논문에서 제안하는 STKC-Sign2Gloss2Text는 수어자의 세부적인 동작들의 시간적 및 공간적인 정보를 활용하기 위해 추출된 키포인트들을 RNN 계열 중 하나인 Bidirectional Gated Recurrent Units(Bi-GRU)로 임베딩 하였다. 이는 Long Short-term Memory(LSTM)[14]이나 Gated Recurrent Unit(GRU)[15]를 사용했을 때보다 실험적으로 우수함을 보였기에 선택되었다. 이를 통해, 연속적인 수어 이미지들의 순차적인 특징을 고려할 수 있었으며, PHOENIX-2014 T 데이터셋에서 베이스라인인 Sign2Gloss2Text의 성능을 능가하였다. 뿐만 아니라, DEV에서는 22.92, TEST에서는 21.87의 BLEU-4 score를 달성하였다.

2. 관련 연구

초기 수어에 관한 연구들은 연속적인 수어 이미지들을 수어 동작의 의미를 나타내는 글로스로 전사하는 연속 수어 인식이 중점이 되었다. 수어 모델의 입력 데이터인 연속적인 수어 비디오의 순차적 모델링을 위해 딥러닝 모델 중 하나인 Convolutional Neural Network(CNN)[16], RNN[17] 및 GRU[31] 모델을 사용하였다. 이후, [7]에서 처음으로 수어 번역(Sign Language Translation, SLT)을 위해 독일의 공영 방송사인 PHOENIX의 일별 뉴스와 일기 예보를 수어로 번역한 것을 녹화한 PHOENIX-2014 T 데이터셋을 구축하였다. 해당 데이터셋은 수어 비디오, 수어 비디오를 전사한 연속적인 글로스들, 독일어 번역문의 세 가지 데이터로 이루어졌으며, 글로스는 수어를

음성 언어로 표현한 것이다[18]. [7]에서 데이터셋을 제시함과 동시에 seq-to-seq 기반의 수어 번역 모델인 Sign2Text도 함께 제시되었다. Sign2Text 모델이 제시된 후 수어 번역에 대한 관심이 증가하였으며, 관련 연구 중 하나인 [12]에서는 수어자의 키포인트들을 Openpose [19,20,21]와 같은 라이브러리를 이용해 이미지마다 상체에서 12개, 양손에서 각각 21개, 그리고 얼굴에서 70개의 키포인트들을 추출하였다. 각 키포인트들은 한 이미지에 존재하는 모든 키포인트들에 대해서 정규화되었다. 정규화된 키포인트들은 GRU로 구성된 seq-to-seq 내부 모듈 중 하나인 인코더의 입력이 되며, 이때 수어 이미지 자체는 사용되지 않는다.

Sign2Text를 기반으로 하는 모델들은 수어 이미지에서 구어로의 직접적인 번역을 수행하였으나, 신경망 기계 번역(Neural Machine Translation, NMT)에서 획기적인 발전을 이루었던 Transformer 구조를 사용한 [8] 논문에서 처음으로 수어 인식과 수어 번역을 동시에 수행하는 모델인 Sign2Gloss2Text가 제시되었다. Sign2Gloss2Text의 인코더의 결과(output)는 Connectionist Temporal Classification (CTC)[22] 층을 통해 수어 이미지에 해당하는 글로스를 예측하며, 이 과정을 Sign Language Recognition Transformer(SLRT)로 명명하였다. 또한, 디코더는 예측된 글로스를 자연스러운 구어 문장으로 번역하는 역할을 하고, 이를 Sign Language Translation Transformer(SLTT)라고 명명하였다. 수어 인식과 수어 번역을 동시에 학습하는 모델을 설계하기 위해 인코더의 결과가 디코더의 모듈 중 하나인 encoder-decoder attention 층의 입력이 되도록 하였다. Sign2Gloss2Text는 기존 모델 대비 큰 성능 향상을 보였으며, PHOENIX-2014 T 데이터셋에서 21.32의 BLEU-4 score를 달성하였다. 이후, Sign2Gloss2Text 구조를 기반으로 제시된 연구 중 [13]은 [23]에서 제안한 Spatial-Temporal Multi-Cue(STMC) 모듈을 Sign2Gloss2Text의 인코더에 적용하였다. STMC는 Spatial Multi-Cue(SMC)와 Temporal Multi-Cue (TMC) 모듈로 구성되어 있다. SMC 모듈은 컨볼루션(Convolution) 층을 통해 생성된 특징(feature)을 deconvolution layer [24] 및 soft-argmax trick [25]을 사용해 수어자의 얼굴과 양손의 키포인트들을 추출한 후 키포인트에 맞게 이미지를 자른다. 잘린 이미지들은 CNN으로 구성된 모듈을 거쳐 얼굴 및 양손과 같은 시각적 단서(visual cue)의 특징을 생성한다. 이후, 생성된 특징은 TMC 모듈의 입력이 되며, 이는 시각적 단서(visual cue)의 각 특징을 학습하는 intra-cue와 합쳐진 특징을 학습하는 inter-cue로 구성되어 있다. Sign2Gloss2Text 모델에 STMC 모듈을 활용함으로써 각각의 수어 이미지에 대해 적절한 글로스 표현(representation) 생성을 가능하게 하였으며, 이는 곧 성능 향상에 기여하였다[13].

3. 연구 방법

수어는 손의 위치, 방향, 손 및 손가락 모양을 비롯한 상반신 동작과 얼굴 표정이라는 시각적 특성만으로 의사를 전달하는 고유 언어로서 표현 방식은 물론 문법 역시 음성 언어와는 다른 독자적인 체계를 가지고 있다[1,10]. 한편 수어는 같은 동작일지라도 손의 방향 혹은 표정 변화에 따라 전혀 다른 의미가 표현된다는 점에서 시각적 특성으로 이루어진 각 동작 키포인트들의 시공간적 구조 정보(spatial-temporal structure information)를 포착하는 것이 매우 중요하다고 볼 수 있다. 그러나 기존의 Sign2Gloss2Text에 기반한 수어 번역 연구에서는 수어 이미지 전체에 대한 spatial embedding과 positional encoding을 통해 포괄적인 수어 정보만을 전달할 뿐 수어 번역에 주요한 각 동작 부위의 세부적인 정보는 강조하지 못하고 있다. 이에 따라 우리는 Sign2Gloss2Text에서 수어 동작 포인트의 시공간적 정보를 보충해주는 spatial-temporal keypoints embedding 모듈을 추가함으로써 수어 동작을 인식 및 번역하는 데에 핵심적인 순서 및 의미론적 정보(sequential and semantic information)를 확보하고자 했다. 제안 모델인 STKC-Sign2Gloss2Text Translation의 전체 도식은 Fig. 1과 같다.

3.1 Spatial-temporal Keypoints Embedding

수어 동작이 담고 있는 시공간적 구조 정보를 모델에 반영하기 위해 우리는 Sign2Gloss2Text 의 spatial embedding 층과 병렬적으로 진행되는 spatial-temporal keypoints embedding 모듈을 추가하였다. Spatial-temporal keypoints embedding 모듈은 키포인트들을 추출해 공간적 구조 정보를 생성하

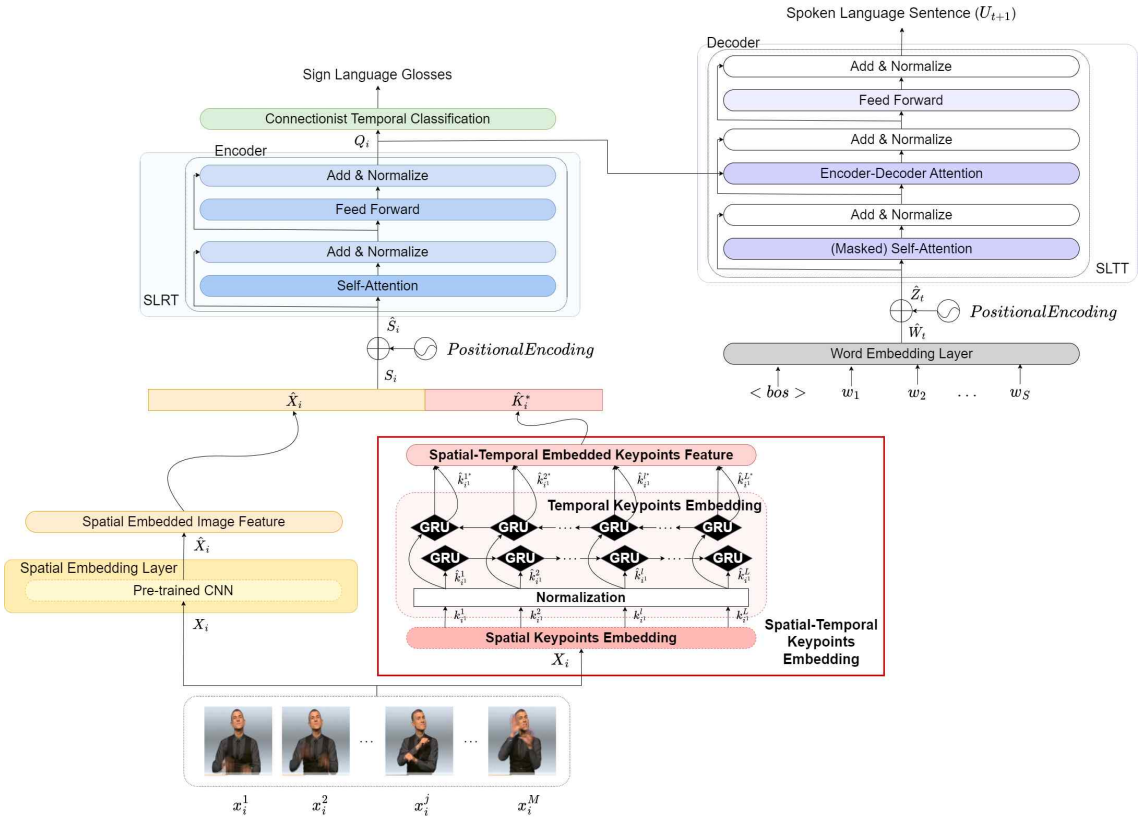

Fig. 1. An Overview of Spatial–temporal Keypoints Centered Sign2Gloss2Text Translation. Spatial–temporal Keypoints Embedding, which is boxed in red, is the module presented in this paper.

는 spatial keypoints embedding 단계와 이를 바탕으로 키포인트들의 순서 정보를 확보하는 temporal keypoints embedding 단계로 이루어진다. 첫 번째 단계인 spatial keypoints embedding에서는 Open Pose의 human pose estimation을 사용해 수어 비디오 X_i 내의 j번째 이미지 x_i^j에 대해 양손, 상반신 및 얼굴의 키포인트들을 추출한다. 이때, 상반신을 비롯한 양손 및 얼굴의 키포인트들은 OpenPose[19]를 바탕으로 추출되었으며, 그중 양손과 얼굴은 OpenPose의 [20]이 함께 적용되어 진행되었다. 이를 수식으로 표현하면 다음과 같다.

$$I_{video} = \{X_i \mid i = 1,2,\cdots \mathrm{N}\}, \quad (1)$$

N is the number of total videos.

$$X_i = \{x_i^j \mid i = 1,2,\cdots, N, j = 1,2,\cdots \mathrm{M}\}, \quad (2)$$

M is the number of frames in a video.

$$K_i = SpatialKeypointsEmbedding(X_i), \quad (3)$$

$$K_i = \{k_i^j = (\alpha_i^j, \beta_i^j) \mid i = 1,2,\cdots, N, j = 1,2\cdots, M,\ l = 1,2,\cdots, \mathrm{L}\}, \quad (4)$$

L is the number of keypoints in a frame.

이를 통해 우리는 수어 비디오 X_i로부터 수어 동작을 인식하는 데에 주요한 상반신 9개, 양손 42개 및 얼굴 70개 총 121개의 키포인트들 K_i를 추출하였다. 여기서 키포인트들은 i번째 비디오의 j번째 프레임마다 l개씩 생성되며 (4)와 같이 x축과 y축의 좌표로 이루어진 $k_i^j = (\alpha_i^j, \beta_i^j)$이다. 한편, 추출된 키포인트 좌표값 k_i^j은 신체 부위의 크기 혹은 위치에 따라 좌표값의 차이가 존재할 수 있으므로, [12]에 따라 한 이미지 내에 존재하는 모든 키포인트들에 대해 각 축을 기준으로 정규화(normalization)하는 과정을 거쳤다. 과정은 식 (5)와 같다. 이때, 키포인트들의 x축에 대한 평균값과 표준편차는 $\bar{\alpha}_i^j$, $\sigma(\alpha_i^j)$로, y축에 대한 평균값과 표준편차는 $\bar{\beta}_i^j$, $\sigma(\beta_i^j)$로 표현되었다.

$$\hat{\alpha}_i^j = \frac{\alpha_i^j - \bar{\alpha}_i^j}{\sigma(\alpha_i^j)}, \quad \hat{\beta}_i^j = \frac{\beta_i^j - \bar{\beta}_i^j}{\sigma(\beta_i^j)} \tag{5}$$

$$\hat{k}_i^j = (\hat{\alpha}_i^j, \hat{\beta}_i^j) \tag{6}$$

$$\hat{K}_i = \left\{ \hat{k}_i^j \mid i = 1, 2, \cdots, N, j = 1, 2 \cdots, M, \ l = 1, 2, \cdots, \mathrm{L} \right\}$$

다음으로 우리는 정규화된 키포인트들인 $\hat{K}_i$를 바탕으로 두 번째 단계인 temporal keypoints embedding을 진행하였다. 수어는 시각적 정보를 통해 의사를 전달하는 언어이며 시간의 흐름에 따라 동작이 진행된다는 점에서 키포인트들의 순서 정보를 보존하는 것이 중요하기 때문이다. 따라서 우리는 121개의 키포인트들인 $\hat{K}_i$를 1개의 층으로 이루어진 순차 모델링 층인 Bi-GRU($\theta(\bullet)$)에 통과시킴으로써 학습 파라미터를 최소한으로 사용하면서도 추출한 키포인트들의 순서 및 의미론적 구조 정보를 강화하고자 하였다. 이때 2단계로 임베딩된 키포인트들은 식 (7)의 $\hat{K}_i^*$로 표현하였으며 위의 2단계를 통칭하여 (8)을 (9)로 정의하고, 아래와 같이 표기하였다.

$$\hat{K}_i^* = \mathit{TemporalKeypointsEmbedding}(\hat{K}_i) \tag{7}$$

$$\hat{K}_i^* = \mathit{TemporalKeypoinsEmbedding}(\mathit{SpatialKeypointsEmbedding}(X_i)) \tag{8}$$

$$\hat{K}_i^* = \mathit{Spatial\ TemporalKeypointsEmbedding}(X_i) \tag{9}$$

3.2 Spatial-temporal Keypoints Centered Sign2Gloss2Text Translation

본 연구는 수어 번역 성능의 주축이 되는 글로스 표현을 향상하는 데에 있다. 이를 위해 Fig. 1과 같이 SLRT 및 SLTT 구조로 이루어진 Sign2Gloss2Text에 수어 동작에 관한 주요 키포인트들의 시공간적 구조 정보를 추가 반영할 수 있는 spatial-temporal keypoints embeddding 모듈을 추가하였다. 이에 따라 우리는 [10]과 동일하게 수어 비디오 X_i를 pre-trained CNNs($\phi(\bullet)$)에 통과하여 수어 이미지 전체에 관한 특징 $\hat{X}_i$를 얻은 후, 이를 제안한 식 (9)에 따라 추출한 $\hat{K}_i^*$와 결합하여 SLRT의 positional encoding에 넣어줌으로써 Sign2Gloss2Text 보다 수어 동작의 키포인트들을 강조한 모델 학습이 이루어지도록 하였다. 이를 식으로 표현한 바는 아래와 같으며, 식 (11)에 따라 키포인트들의 시공간적 구조 정보가 반영된 S_i는 Transformer의 positional encoding[11]과 동일한 방법으로 위치 정보를 인코딩하게 된다.

$$\hat{X}_i = \mathrm{Pr}\mathit{etrained\,CNNs}(\phi(X_i)) \tag{10}$$

$$S_i = \hat{X}_i ; \hat{K}_i^* \ \text{(concatenation)} \tag{11}$$

$$\hat{S}_i = S_i + \mathit{PositionalEncoding}(i) \tag{12}$$

SLRT 단계에서는 프레임별 순서 정보(positional information)가 인코딩된 $\hat{S}_i$을 입력받아 multi-head로 이루어진 self-attention 층을 지나 non-linear point-wise feed forward 층을 통과하게 된다. 이를 통해 모델은 키포인트들이 강조된 시공간적 구조 정보에 self-attention[11]을 수행함으로써 다양한 시점에서 얻은 각 동작의 연관 정보를 수집할 수 있게 된다. 각 층은 모델 학습을 돕기 위해 연산 과정이 완료될 때마다 해당 층의 입력값을 다시 더해주는 잔차 연결(residual connection)과정과 층 정규화(layer normalization)를 거친다. SLRT의 결과는 각 이미지에 해당하는 글로스를 예측하고, cross-entropy[26] 손실 함수(loss function)와 CTC 층을 통해 최적화된다. SLRT에서 글로스를 예측하는 과정은 SLRT 네트워크를 통과한 각 프레임의 특징들이 글로스 표현에 가까워지도록 설계되었다. 이는 곧, 글로스가 수어 번역을 위한 중간 감독자(intermediate supervisor)의 역할을 하게 된다[10]. 우리는 이 점을 이용해 SLRT의 입력값으로 키포인트들의 시공간적 구조 정보를 보충해 넣어줌으로써 유사한 수어에 대해 구분이 가능한 최적의 글로스 표현 Q_i를 생성하고자 했다. SLRT를 수식으로 표현한 바는 다음과 같다.

$$Q_i = \mathit{SLRT}(\hat{S}_i | \hat{S}_{i:I_{Video}}), i \in \{1, 2, \cdots | I_{Video}|\} \tag{13}$$

SLTT 단계에서는 생성된 글로스 표현을 바탕으로 본격적인 수어 번역이 이루어진다. SLTT는 자기회귀(autoregressive) 과정의 Transformer 디코더로 구성되어 있어 이전 단어를 바탕으로 다음 단어를 예측한다. Fig. 1에서 볼 수 있듯, 각 프레임에 해당하는 t번째 구어 문장 W_t는 S개의 단어 w_s로 이루어져 있으며 word embedding 층을 통해 t번째 문장의 임베딩 벡터 $\hat{W}_t$가 된다. 이를 수식으로 표현하면 다음과 같다.

$$W_t = \{w_s | s = 1,2, \cdots, S\}, \quad (14)$$

S is the number of words in a sentence.

$$\hat{W}_t = WordEmbedding(W_t) \quad (15)$$

이후 $\hat{W}_t$는 SLRT와 동일하게 positional encoding을 통과함으로써 구어 문장의 순서 정보를 유지하게 된다. SLTT는 masked self-attention 층과 encoder-decoder attention 층 및 non-linear point-wise feed forward 층으로 구성되어 있다. 각 층을 세부적으로 설명하면 다음과 같다. Masked self-attention 층에서는 positional encoding을 통과한 단어 벡터 $\hat{Z}_t$에 mask를 사용하여 오직 이전 단어만으로 다음 단어를 예측하도록 학습하면서 문맥 정보(contextual information)를 추출한다. 추출된 문맥 정보는 SLRT의 결괏값인 글로스 표현 $Q_{i:I_{Video}}$과 결합되어 encoder-decoder attention 층을 통과하게 된다. 이를 통해 SLTT의 구어 표현(spoken language representation)은 SLRT의 글로스 표현과 서로 대응되어 multi-heads attention 과정을 거치면서, 구어 번역에 도움이 되는 다양한 정보를 수집하게 된다. SLTT의 각 층은 연산이 완료된 후 SLRT와 마찬가지로 잔차 연결과 층 정규화를 거치며 cross- entropy 손실함수를 통해 수어 번역에 최적화된다. 본 과정에서 이루어진 positional encoding은 Transformer에서 사용된 식과 같으며[11], SLTT를 통해 디코딩된 U_{t+1}을 수식으로 표현한 바는 아래와 같다.

$$\hat{Z}_t = \hat{W}_t + PositionalEncoding(t) \quad (16)$$

$$U_{t+1} = SLTT(\hat{Z}_t | \hat{Z}_{t:Tword}, Q_{i:I_{Video}}) \quad (17)$$

4. 실험 결과 및 고찰

4.1 실험 데이터

본 연구에서 제시한 STKC-Sign2Gloss2Text Translation의 실험을 위해 우리는 수어 연구의 벤치마크 데이터셋인 PHOENIX-2014 T [7]를 사용했다. PHOENIX-2014 T는 독일의 공영 방송국인 PHOENIX의 일간 뉴스와 일기 예보에 관한 수어 통역 영상과 함께 이에 관한 글로스 표기법 및 독일어 번역 코퍼스(corpus)를 제공하고 있다. 수어 영상에는 총 9명의 서로 다른 수어자가 있으며, 각각의 수어 비디오는 그중 1명의 수어 과정을 정면에서 촬영하였다. 데이터셋은 총 8,257개의 영상으로 구성되어 있으며, Train 7,096개, Dev 519, Test 642개로 분할되어있다. 수어 통역 영상 이미지(Frames)에 관한 글로스와 독일어 번역 코퍼스의 세부 정보는 Table 1과 같다.

4.2 실험 세부 정보

본 연구의 베이스라인으로는 vanilla Transformer를 기반으로 한 Sign2Gloss2Text로 설정하였다. 제안하는 STKC-Sign2Gloss2Text Translation은 베이스라인과 동일하게 인코더와 디코더 모두 512개의 은닉 유닛(hidden units)과 3개의 층으로 이루어졌다. 다만, 8개의 multi-head self-attention 층으로 이루어진 베이스라인의 인코더 및 디코더와 달리 STKC-Sign2Gloss2Text Translation은 인코더와 디코더 모두 베이스라인의 1/2에 해당하는 4개의 multi-head로 이루어져 있다. 학습은 베이스라인과 마찬가지로 미니배치 (mini-batch)의 수가 32이며, learning rate 및 weight decay는 10^{-3}로 설정하였고, 최적화 알고리즘은 Adam [27]을 사용하였다. 이때, Adam의 인자인 β_1과 β_2는 각각 0.9와 0.998로 설정하였다. 또한, 본 실험에서 SLRT와 SLTT의 최적 손실 가중치는 Sign2Gloss2Text와 동일하게 5.0과 1.0으로 설정하였다. 본 실험의 수어 번역에 관한 성능 평가는 기계 번역에서 가장 대표적으로 사용되는 지표인 BLEU score [28]가 사용되었다. 해당 실험은 NVIDIA RTX A6000 (48GB) GPU 1개를 사용하여 실행되었다.

Table 1. Data description of PHOENIX-2014 T.

Corpus / Element	Sign Language Gloss Annotation			German Spoken Language		
	Train	Dev	Test	Train	Dev	Test
Frames	827,354	55,775	64,627	827,354	55,775	64,627
Vocabulary	1,066	393	411	2,887	951	1,001
Words	67,781	3,745	4,257	99,081	6,820	7,816

4.3 베이스라인과 STKC-Sign2Gloss2Text의 성능 비교

우리는 키포인트들의 시공간적 구조 정보가 수어 번역에 미치는 영향을 분석하고자 베이스라인과 베이스라인에 spatial keypoint embedding만 적용한 모델(SKC-Sign2Gloss2Text) 및 STKC-Sign2Gloss2Text의 성능을 동일한 벤치마크 데이터셋에서 비교하였다. Table 2는 전반적인 성능을 비교한 표이다. Table 2에서 STKC-Sign2Gloss2Text는 인코더와 디코더에서 모두 베이스라인보다 1/2 적은 4개의 multi-heads self-attention만으로도 기존 성능을 넘어섰음을 보여주고 있다. 제안 모델은 DEV와 TEST의 모든 BLEU score에서 베이스라인의 성능을 능가하고 있으며, 그중 가장 중요한 평가 지표인 TEST BLEU-4에서는 23.19로 베이스라인보다 1.87 향상되었다. 이를 통해 우리는 Sign2GlossText 구조에서 수어 동작에 관한 키포인트들의 이전 및 이후 시점의 순서 정보가 모델의 수어 번역 향상에 일조하고 있음을 알 수 있다. 또한, 키포인트들의 시공간적 구조 정보를 모두 반영한 STKC-Sign2Gloss2Text의 경우, 공간적 구조 정보만 반영한 SKC-Sign2Gloss2Text보다 DEV BLEU-4 score 기준 0.31, TEST BLEU-4 score 기준 0.38 높은 성능을 보였다. 따라서, 수어 이미지 전체의 정보만을 전달해주는 것보다 수어 동작의 키포인트들을 중심으로 시공간적 구조 정보를 강조해 전달해주는 것이 수어 인식 및 번역 성능 개선에 중요하다는 것을 보여준다.

4.4 비교 실험 (Ablation study)

본 논문에서는 키포인트들의 시간적 구조 정보를 반영하기 위해 최적의 RNN 계열의 순차 모델링 층

Table 2. Performance comparison table on PHOENIX-2014 T dataset.

		DEV				TEST			
Model	Multi-heads	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Baseline Model	8	47.26	34.40	27.05	**22.38**	46.61	33.73	26.19	**21.32**
STC-Sign2 Gloss2Text	4	47.22	**34.95**	**27.52**	**22.61**	**47.65**	**35.32**	**27.71**	**22.81**
STKC-Sign2 Gloss2Text	4	**47.88**	**35.42**	**27.96**	**22.92**	**48.03**	**35.57**	**28.02**	**23.19**

Table 3. Performance comparison of varying the temporal embedding layers. If the performance is better than the baseline, it is displayed in bold, and the best is in red.

Keypoints RNN Embedding	Label Smoothing	DEV BLEU-4	TEST BLEU-4	Keypoints RNN Embedding	Label Smoothing	DEV BLEU-4	TEST BLEU-4
LSTM 1 layer	0.0	21.23	**22.10**	LSTM 2 layer	0.0	21.21	20.61
	0.1	21.93	**21.81**		0.1	22.07	**22.88**
	0.2	21.93	**22.68**		0.2	21.81	**21.65**
Bi-LSTM 1 layer	0.0	21.30	20.32	Bi-LSTM 2 layer	0.0	20.72	21.28
	0.1	**22.40**	**21.58**		**0.1**	**23.14**	**22.62**
	0.2	21.50	21.17		**0.2**	**23.18**	**22.38**
GRU 1 layer	0.0	20.93	20.77	GRU 2 layer	0.0	21.00	20.44
	0.1	21.37	**22.17**		**0.1**	**22.33**	**22.89**
	0.2	**22.96**	**21.94**		**0.2**	**23.06**	**21.93**
Bi-GRU 1 layer	0.0	21.40	20.55	Bi-GRU 2 layer	0.0	20.54	21.10
	0.1	**22.92**	**23.19**		0.1	22.26	**21.97**
	0.2	**22.57**	**21.76**		**0.2**	**22.86**	**22.07**

Table 4. Performance comparison of varying the number of multi-heads in encoder and decoder.

Keypoints Embedding	Enc_layer	Enc_heads	Dec_layer	Dec_heads	DEV BLEU-4	TEST BLEU-4
OpenPose+ Bi-GRU 1 layer	3	**4**	3	**4**	**22.92**	**23.19**
		8		4	23.10	21.72
		8		8	22.12	21.98

을 찾기 위한 추가 실험을 진행하였다. Table 3은 4개의 multi-heads로 이루어진 인코더 및 디코더 구조에서 각 RNN 계열이 수어 번역에 미친 결과를 보여주고 있다. 실험에서는 수어 동작의 시퀀스에 따른 순서 정보의 손실을 방지하고자 RNN의 장기 의존성(long-term dependency) 문제를 개선한 LSTM과 GRU를 사용하였다. 우리는 LSTM과 GRU의 층의 수와 양방향(bidirectional) 여부를 변경해가며 실험하였다. 또한, 수어 번역에 키포인트 중심의 추가 정보를 보강해줌으로써 훈련 시 디코더의 구어 예측 단계에서 과잉 확신(overconfidence)되는 문제를 방지하고자 cross-entropy 손실 함수에 label smoothing [29, 30]을 0.0. 0.1과 0.2로 적용하여 진행되었다. Table 3에서 볼 수 있듯 LSTM보다 적은 학습 파라미터를 가진 1개 층의 Bi-GRU가 가장 좋은 성능을 보임으로써 제안 모델의 temporal keypoints embedding 층으로 선정되었다. Table 3은 LSTM, Bi-LSTM, GRU가 층의 개수와 상관없이 0.1의 label smoothing에서 TEST 기준 가장 향상된 성능이 나타남을 보여주고 있다. 그러나 Bi-GRU를 통해 키포인트의 시간적 구조 정보를 임베딩한 경우에는 층의 개수에 따라 label smoothing 값에 차이가 있었다. Bi-GRU를 2개의 층으로 쌓아 올린 경우, 1개 층의 Bi-GRU와 달리 label smoothing을 0.2로 설정한 경우 가장 좋은 성능을 보여주었다. 이에 대해 우리는 양방향 순차 모델링을 통해 키포인트 중심의 이전 및 이후 시점의 순서 정보를 추가로 반영하고 있는 Bi-GRU가 층이 더해질수록 모델의 구어 번역 시 과잉 확신을 상승시켜 예측 정확도가 떨어지는 데에 기여하고 있으며, 그 결과 층이 적은 경우보다 더 많은 label smoothing이 적용되어야 한다는 것으로 해석할 수 있었다. 여기서 나아가 우리는 제안 모델의 성능을 최적화하는 인코더와 디코더의 multi-heads 수를 찾기 위한 추가 실험을 진행하였다. Table 4는 수어 동작의 키포인트들을 바탕으로 시공간적 구조 정보가 강조된 경우, 8개의 multi-heads로 이루어진 베이스라인과 달리 인코더, 디코더 모두에서 그의 절반인 4개의 multi-heads attention 연산만으로도 가장 좋은 성능을 보이는 것을 알 수 있었다. 이를 통해 우리는 베이스라인보다 적은 multi-heads attention에도 불구하고 키포인트 중심의 시공간적 구조 정보가 모델이 수어를 인식 및 번역하는 데에 충분한 정보를 전달해주고 있음을 유추할 수 있었다.

5. 결 론

수어는 시각적 특성을 통해 의미를 전달한다는 점에서 이를 모델에 반영하는 것은 수어 번역 향상에 주요한 영향을 미친다. 이에 따라 우리는 spatial-temporal keypoints embedding 모듈을 통해 수어 동작의 키포인트들을 추출한 후 시간적 구조 정보를 보완해줌으로써 수어를 인식하는 데에 핵심적인 시각적 특성을 반영해주고자 하였다. 제안 모델 STKC-Sign2Gloss2Text Translation은 OpenPose을 통해 수어 동작의 상반신, 양손 및 얼굴 부위에 관한 121개의 키포인트들을 추출하였으며, 실험을 통해 가장 높은 성능을 보여준 Bi-GRU를 사용해 수어 동작 키포인트들의 순서 및 의미론적 정보를 보존하였다. 그 결과, STKC-Sign2Gloss2Text Translation은 인코더와 디코더 전부에서 베이스라인 절반의 multi-heads 수만으로도 PHOENIX-2014 T 데이터셋의 DEV와 TEST 전부에서 베이스라인보다 향상된 결과를 보여주었으며 BLEU-4 기준 각각 22.92와 23.19를 달성하였다. 이를 통해 본 연구에서는 수어 동작의 키포인트들로부터 추출한 시공간적 구조 정보를 모델에 반영하여 의미 있는 글로스 표현을 생성하는 것이 수어 번역 성능 향상에 도움을 준다는 것을 확인하였다. 이를 기반으로 후속 연구에서는 Sign2Gloss2Text Translation의 파라미터 수를 개선하면서도 수어 동작의 시공간적 구조 정보를 효과적으로

학습시키는 방법을 고안함으로써 수어 번역의 성능을 향상할 예정이다.

REFERENCE

[1] The Story on the National Institute of Korean Language, https://www.korean.go.kr/front/page/pageView.do?page_id=P000300&mn_id=202 (accessed September 20, 2022).

[2] M.G. Grif and A.V. Kugaevskikh, "Recognition of Deaf Gestures Based on a Bio-Inspired Neural Network," *Journal of Physics: Conference Series*, Vol. 1661, No. 1, p. 012038, 2020.

[3] Korean Sign Language Dictionary, https://sldict.korean.go.kr/front/sign/signContentsView.do?current_pos_index=8&origin_no=6046&searchWay=&top_category=&category=CTE001005&detailCategory=005&searchKeyword=&pageIndex=1&pageJumpIndex= (accessed September 27, 2022).

[4] U. Bellugi and S. Fischer, "A Comparison of Sign Language and Spoken Language," *Cognition*, Vol. 1, No. 2, pp. 173-200, 1972.

[5] A. Voskou, K.P. Panousis, D. Kosmopoulos, D.N. Metaxas, and S. Chatzis, "Stochastic Transformer Networks with Linear Competing Units: Application to end-to-end SL Translation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11946-11955, 2021.

[6] M. De Coster, K. D'Oosterlinck, M. Pizurica, P. Rabaey, S. Verlinden, M. Van Herreweghe, et al., "Frozen Pretrained Transformers for Neural Sign Language Translation," *18th Biennial Machine Translation Summit (MT Summit 2021)*, Association for Machine Translation in the Americas, pp. 88-97, 2021.

[7] N.C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural Sign Language Translation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7784-7793, 2018.

[8] S. Stoll, N.C. Camgoz, S. Hadfield, and R. Bowden, "Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks," *International Journal of Computer Vision*, Vol. 128, No. 4, pp. 891-908, 2020.

[9] Z. Niu and B. Mak, "Stochastic Fine-Grained Labeling of Multi-state Sign Glosses for Continuous Sign Language Recognition," *European Conference on Computer Vision*, pp. 172-186, 2020.

[10] N.C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation," *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition*, pp. 10023-10033, 2020.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., "Attention is All You Need," *Advances in Neural Information Processing Systems*, Vol. 30, 2017.

[12] S.K. Ko, C.J. Kim, H. Jung, and C. Cho, "Neural Sign Language Translation Based on Human Keypoint Estimation," *Applied Science*, Vol. 9, Issue 13, 2683, 2019.

[13] K. Yin and J. Read, "Better Sign Language Translation with STMC-Transformer," *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5975-5989, 2020.

[14] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," *arxiv Preprint*, arXiv: 1402.1128, 2014.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv Preprint*, arXiv:1412.3555, 2014.

[16] L. Pigou, S. Dieleman, P.J. Kindermans, and B. Schrauwen, "Sign Language Recognition Using Convolutional Neural Networks," *European Conference on Computer Vision*, pp.

572-578, 2014.

[17] R. Cui, H. Liu, and C. Zhang, "Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7361-7369, 2017.

[18] Korean Sign Language Maintenance Project: Korean Sign Language Dictionary (Record of Korean Sign Language), https://www.korean.go.kr/nkview/nklife/2017_2/27_0204.pdf (accessed September 29, 2022).

[19] Z. Cao, T. Simon, S.E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291-7299, 2017.

[20] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1145-1153, 2017.

[21] S.E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724-4732, 2016.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks." *ProceedIngs of the 23rd International Conference on Machine Learning*, pp. 369-376, 2006.

[23] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13009-13016, 2020.

[24] M.D. Zeiler, D. Krishnan, G.W. Taylor, and R. Fergus, "Deconvolutional Networks," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 2528-2535, 2010.

[25] O. Chapelle and M. Wu, "Gradient Descent Optimization of Smoothed Information Retrieval Metrics," *Information Retrieval*, Vol. 13, No. 3, pp. 216-235, 2010.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT press, 2016.

[27] D.P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv Preprint,* arXiv: 1412.6980, 2014.

[28] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "Bleu: A Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* pp. 311-318, 2002.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2818-2826, 2016.

[30] R. Müller, S. Kornblith, and G.E. Hinton, "When Does Label Smoothing Help?" *Advances in Neural Information Processing Systems*, Vol. 32, 2019.

[31] B. Subramanian, B. Olimov, and J. Kim, "Fast Convergence GRU Model for Sign Language Recognition," *Journal of Korea Multimedia Society*, Vol. 25, No 9, pp. 1257-1265, 2022.

김 민 채

2018년 숙명여자대학교 법학부 학사 졸업
2022년~현재 연세대학교 정보대학원 비즈니스 빅데이터 분석 석사과정

관심분야 : 머신러닝(Machine Learning), 딥러닝(Deep Learning), 영상처리(Image Processing), 멀티모달(Multi-Modal), 이미지 및 음악 검색 시스템(Image and Music Retrieval System)

김 정 은

2020년 숙명여자대학교 수학과 학사 졸업
2020년~현재 연세대학교 인공지능대학원 박사과정

관심분야 : 인공지능(Artificial Intelligence), 딥러닝(Deep Learning), 딥러닝 생성 모델(Deep Learning Generative Model), 객체 검출(Object Detection), 모델 경량화(Model Compression)

김 하 영

2000년 경희대학교 수학과 학사
2007년 퍼듀대학교 수학과 석사
2010년 퍼듀대학교 수학과 박사
2011년~2016년 삼성전자 종합기술원 전문연구원
2016년~2019년 아주대학교 금융공학과 조교수
현재 연세대학교 정보대학원 부교수

관심분야 : 머신러닝(Machine Learning), 딥러닝(Deep Learning), 수학 및 계산금융(Computational and Mathematical Finance), 확률 과정(Stochastic Process), 확률 이론(Probability Theory)