재난 상황 전달을 위한 수어 ↔ 한국어 양방향 번역 시스템

20th Yeji Lee, 21st Dongwook Kim, Jiyeop Kim, Jiwon Kim, Juneon Lee
KUBIG(Korea University Data Science & AI Society), Korea University
(ye_ji, ppuppu1200, ljykjy1309, jinnyk42, jel0206)@korea.ac.kr

KUBIG DATA SCIENCE & AI 8th CONFERENCE

Introduction

• Motivation(프로젝트의 필요성)

재난 상황에서의 정보 접근성은 생존과 직결되는 문제이다. 그러나 청각장애인을 위한 정보 전달 수단은 부족하며, 특히 수어 영상 기반의 재난 정보 전달 시스템은 거의 전무한 상황이다. 본 프로젝트는 AI 기술을 기반으로 수어, 한국어 번역 시스템을 구축함으로써, 청각장애인의 재난 대응 능력 향상에 기여하고자 한다.

또한 영상 기반 수어 인식에는 데이터 크기, 배경 노이즈 등 현실적인 한계가 존재한다. 따라서 영상의 keypoint를 기반으로 수어 ↔ 한국어 양방향 번역을 통해 경량화된 효율적인 수어 번역 모델을 구현하고자 한다.

• Key Idea(아이디어의 핵심)

- 수어 ↔ 한국어 양방향 번역 시스템을 목표로 함
- 수어 → 한국어 : Sign2Text, 수어 영상의 keypoint를 추출하여, sequence-to-sequence 모델로 자연어 번역
- 한국어 → 수어 : Text2Sign, 입력 문장을 수어 gloss로 변환한 후, 사전 제작된 수어 영상 keypoint 애니메이션을 연결

• 데이터셋 설명

AI Hub의 재난 안전 수어 영상 데이터셋 활용

- 총 201,026개의 수어 영상, text 데이터셋 중, 사회재난-화재 데이터 사용
  - Train set : 고유 문장 1,828개, 수어 영상 3,595개(동일 문장에 대해 여러 Signer 및 Camera Angle 존재)
  - Test set : 수어 영상 448개
- 각 영상에 대한 번역(자연어)/keypoint(동영상) 정보가 포함, 이를 그대로 사용함
- 재난 상황에 특화된 어휘 및 표현으로 구성되어 있어, 실제 응용 가능성이 높음

• 아이디어 구체화 및 논리적 과정 설명

(1) Sign2Text

- keypoint → 한국어 문장 : Sign Language Translation(SLT)

CV 기반으로 Sign2Text를 수행하는 방식은 수어 영상 모델의 복잡한 시공간적 특성으로 인한 모델 규모 증가를 고려하여 배제하였다. 또한, keypoint → gloss → 한국어 문장의 단계를 거치는 방식은 gloss annotation이 자연어 번역이라는 목표에 부적절하여 배제되었다. keypoint → 한국어 문장으로 바로 번역하는 방식이 효율적인 학습과 우수한 일반화 성능을 확보할 수 있어 선택되었다.

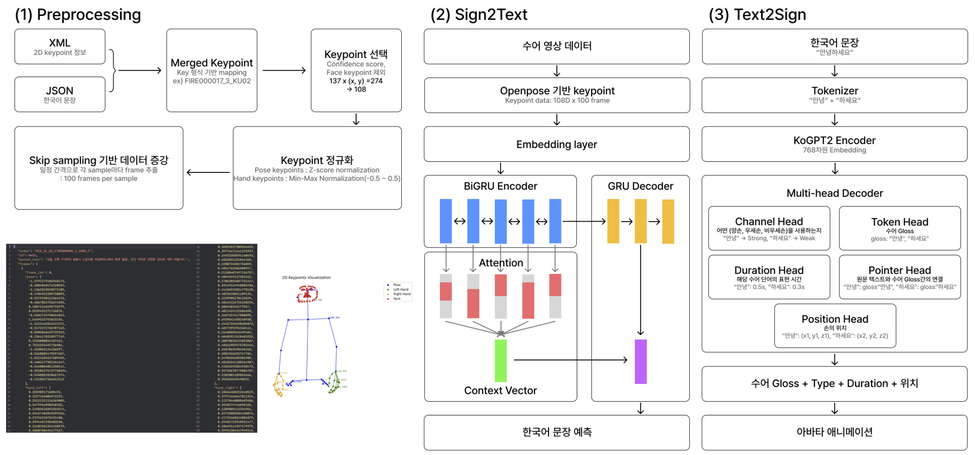
(2) Text2Sign

- 한국어 문장 → gloss → keypoint

keypoint를 직접 추정하는 방식에는 기술적 한계가 존재하고, 수어의 연속성과 자연스러움이 훼손될 수 있다. 따라서, 한국어 문장을 먼저 gloss 형태로 번역한 뒤, gloss를 기반으로 keypoint를 직접 mapping하여 가상 아바타의 움직임을 설계하는 방식이 선택되었다.

Method

• 모델 구조



• 전체적인 과정 간단하게 설명(Input & Output)

(1) Preprocessing

전처리 과정에서는 XML 형식의 2D keypoint 정보와 JSON 형식의 한국어 문장을 FIRE000017_3_KU02와 같은 고유 키를 기준으로 매핑하여, 하나의 샘플로 결합한다. 총 137개의 keypoint 중 얼굴 및 confidence 정보를 제외한 손·팔 중심의 108개의 좌표만을 사용하여 입력 차원을 줄인다. 정규화는 상체와 몸통 keypoint에는 Z-score, 손 keypoint에는 Min-Max 정규화를 적용하여 프레임 간 상대적 위치를 유지하고 민감도를 조절하였다. 또한, 영상 길이와 속도 차이에 따른 과적합을 방지하고 일반화 성능을 높이기 위해, 일정 간격으로 100프레임씩 추출하는 Skip Sampling 기반의 데이터 증강을 수행하였다.

(2) Sign2Text

Sign2Text 모델은 수어 keypoint 시퀀스를 입력으로 받아 한국어 문장으로 생성하는 구조로, BiGRU 인코더와 GRU 디코더, 그리고 Bahdanau 어텐션으로 구성된다. 인코더는 108차원 keypoint sequence를 양방향으로 처리하여 문맥 정보를 풍부하게 반영하며, 디코더는 이전에 생성된 단어와 인코더의 hidden state를 기반으로 다음 단어를 순차적으로 예측한다. 각 디코딩 단계에서는 어텐션 메커니즘을 통해 입력 sequence 중 의미 있는 프레임에 집중하여 문장을 생성하며, 특히 Bahdanau 어텐션을 사용했을 때 Luong 방식보다 더 정밀하고 안정적인 성능을 보였다. 이 모델은 중간 gloss 없이도 직접 자연어 생성이 가능하며, 학습 효율성과 실용성을 모두 갖춘 수어 번역 구조로 설계되었다.

초기 모델은 BiGRU 인코더와 GRU 디코더 기반의 Seq2Seq 구조에 Bahdanau 어텐션을 적용하였으며, gloss 없이 직접 자연어 문장을 생성한다. 하지만 표현 다양성과 고유명사의 영향으로 BLEU 점수가 매우 낮고, 학습이 memorization에 치우치는 결과를 보였다. 따라서 개선된 모델에서는 문장 정규화 및 masking을 사용하였다. 다양한 표현을 동일한 형태로 정규화하고, 고유명사(시간, 지역, 주소 등)를 <슬롯> 토큰으로 마스킹하여 입력 문장의 의미 구조에 집중할 수 있도록 했다. 이로 인해 BLEU 및 METEOR 지표가 향상되었고, 모델의 일반화 성능이 크게 개선되었다. 최종적으로는, 샘플링 품질을 높이기 위해 top-k 및 top-p 기반의 확률 제어를 적용하고, 반복 표현 억제를 위한 repetition penalty 및 과도한 확산을 억제하는 label smoothing을 함께 적용하였다. 또한 Teacher Forcing Ratio(TFR)를 조절하여 학습 안정성과 디코딩 일반화를 균형 있게 달성하였다.

(3) Text2Sign

입력된 한국어 문장은 KoGPT2의 토크나이저에서 Token sequence로 분해된 뒤, KoGPT2를 통해 768차원의 임베딩 벡터로 변환된다. 이후 Multi-head decoder에서 여러 Head를 통해 동시에 다양한 수어 정보를 추론한다. Channel head에서는 양손, 우세손, 비우세손, 비수지 등 어떤 손 동작으로 표현될지를, Token head에서는 실제로 사용될 수어 gloss를 예측한다. Duration head에서는 각 수어 단어가 얼마동안 표현되어야 하는지를, Pointer head에서는 원문 텍스트의 각 token과 수어 단어 사이의 정렬 관계(어텐션)를 계산한다. Position head는 각 수어 단어의 3차원 공간상에서 위치를 예측한다. 이렇게 각 Head에서 예측된 결과는 하나의 수어 sequence로 조합된 뒤, 사전 제작된 수어 영상 keypoint와 연결되어, 3D 아바타가 실제 사람이 수어를 구사하는 것처럼 동작하게 된다.

Result

• 실험 및 프로젝트 결과

(1) Sign2Text

[Pred] <시간> <지역> <지역> <지역> 용암교회 인근 부경사로 화재 발생. 이 지역을 우회하여 주시고 인근 주민은 안전사고 발생에 유의 바랍니다.
[True] 오늘 <시간> <지역> <지역> <지역> 용암교회 인근 부경사로 화재 발생. 이 지역을 우회하여 주시고 인근 주민은 안전사고 발생에 유의 바랍니다.

• Quantitative Results

(1) Sign2Text

실험 단계	설정 요약	BLEU	METEOR	ROUGE	Val Loss
Baseline	Aug×50, No Masking, 기본 Decoder	≈ 0	≈ 0	≈ 0	high
문장 정규화 + 마스킹	TFR=0.3	0.2565	0.3856	0.0000	8.6721
	TFR=0.1	0.2427	0.3675	0.0000	9.5380
Repetition Penalty + Top-k/p + Label Smoothing 적용	Top-k=10, Top-p=0.85, RP=1.2	0.2612	0.4159	0.0000	**3.5317**
	Top-k=5, Top-p=0.9, RP=1.2	0.2909	0.4462	0.0000	3.5741
	Top-k=5, Top-p=0.9, RP=0.9, TFR=0.5	**0.3254**	**0.4900**	**0.0016**	3.7001

(2) Text2Sign

항목	수치
Validation Loss	0.0807
Mean of Mode Loss	0.0427
Gloss Prediction Loss	0.2273
Timing Prediction Loss	0.0430
Pointing (Alignment) Loss	0.0103
BLEU (pre-train)	0.0
BLEU (post-train)	5.88e-156

• Qualitative Results

(1) Sign2Text

모델이 생성한 문장은 대부분 재난 경고 문구의 핵심 의미를 정확히 담고 있었지만, 고유명사나 시간 표현이 많은 문장에서 반복 또는 불필요한 표현이 포함되는 경우가 있었다. 이를 해결하기 위해 고유명사 표현을 <지역>, <시간> 등의 slot으로 마스킹하고, 문장 구조를 정규화함으로써 의미 표현의 일관성과 문장 품질이 크게 향상하였다.

(2) Text2Sign

Text2Sign 모델은 입력된 문장을 기반으로 수어 gloss, channel, duration, pointing, position 등의 정보를 예측하였고, 해당 결과를 기반으로 한 애니메이션은 기술적으로 자연스러운 동작을 보여주었다. 그러나, 실제 문장의 의미와 매핑된 수어 표현 간에는 차이가 있었으며, 이는 학습에 사용된 데이터셋이 사회재난 중 화재 관련 문장에 한정되어 있어 언어 표현 다양성이 부족하고 범용성이 낮았기 때문이다.

Conclusion

• Summary of Findings(요약)

본 연구는 재난 상황에서 수어 사용자에게 실시간 정보를 전달하기 위한 수어 ↔ 한국어 양방향 번역 시스템을 구축하는 데 초점을 맞추었다. 특히 OpenPose 기반의 2D keypoint 시퀀스로 처리된 수어 데이터를 이용해 이를 통해 경량화된 입력으로도 효과적인 번역 모델이 가능함을 확인하였다.

Sign2Text 모델에서는 gloss 없이도 자연어 문장을 직접 생성할 수 있도록 BiGRU-Attention 기반 seq2seq 구조를 적용하였으며, 정규화 및 slot 마스킹 전략을 도입하여 BLEU 점수를 0.32까지 향상시키는 데 성공하였다. Text2Sign 모델은 KoGPT 기반 임베딩과 multi-head 디코더 구조를 통해 손 종류, 수어 gloss, 표현 시간, 공간적 위치 등 수어 구성 요소를 분리 학습하고, 이를 기반으로 애니메이션을 생성하였다.

• Implications(의의 및 기대효과)

본 연구는 Keypoint 기반 접근을 통해 기존 영상 기반 수어 번역 모델의 한계를 극복하고, 실시간성과 경량화를 동시에 확보할 수 있었다. 특히 gloss 없이도 의미 단위로 문장을 생성하고, slot-based 전처리 기법을 통해 고유명사로 인한 성능 저하를 완화함으로써 재난 알림 시스템 등 실용적 응용 가능성을 입증하였다. 또한 Text2Sign 모델은 향후 3D 수어 아바타 제어, AR 수어 출력 시스템 등으로 확장될 수 있다.

• Limitation & Future Work

현재 데이터셋이 화재 관련 문장에 국한되어 있어 의미적 다양성이 부족하며, 실제 수어 표현과 문장 간 매핑 정확도 또한 제한적이다. 특히 Text2Sign 모델은 형태적 표현은 가능하나 의미 정합성은 아직 미흡하다. 향후에는 더 다양한 재난 유형과 일반 일상 문장을 포함한 대규모 수어 데이터셋 확보, 고도화된 slot tagging, 의미 기반 평가 지표 개발 등을 통해 모델의 범용성과 실용성을 높일 예정이다.