

# 내 이야기를 들어줘

---

청설 | 20기 이예지; 21기 김동욱, 김지엽, 김지원, 이준언

**KUBIG**  
DATA SCIENCE & AI

# CONTENTS

---

01

## Introduction

- 주제 선정 배경
- 수어의 문법적 특성

02

## Objective

- 활용한 데이터셋 소개
- 파이프라인 개요

03

## Results

- 주요 결과 및 모델 성능
- 실제 시연

04

## Conclusions

- 프로젝트 한계 및 의의





# 01. Introduction

# 01. 주제 선정 배경

## 청설 (聽說 내 이야기를 들어줘)



- 청각 장애인들이 겪는 일상 속 소통의 불편함
  - 긴급 상황에서의 정보 접근성과 의사소통의 한계
- 수어-자연어 간 자동 번역 기술의 필요성 → '내 이야기를 들어줘' 프로젝트 기획



# 01. 수어의 문법적 특성



여름아 / 좋아해!

(영화 '청설' 중)



'여름'의 얼굴 이름: 미소

(영화 '청설' 중)

💋 용준아 / 친구 하자고 / 말해줘서 / 고마워

👉 용준 / 친구 / 말하다 / 고맙다

- 의미 단위(gloss) 중심의 간결하고 직관적인 표현
- 문법적 제약이 적은 자유로운 문장 구조
- 표정, 시선, 고개 움직임 등 비수지 표현이 핵심적으로 기능



## 02. Objective

## 02. 활용한 데이터셋 소개

*'화재 사고' 관련 뉴스 문장과 이에 대응하는 수어 영상을 포함하는 데이터셋*

### AI-Hub 재난 안전 정보 전달을 위한 수어 영상 데이터

#### - Train Set

- 고유 문장 수: 1,828개
- 수어 영상 수: 3,595개 (동일 문장에 대해 다양한 Signer 및 Gloss 표현 존재)

#### - Validation/Test Set

- 수어 영상 수: 448개

#### 수어 영상 데이터

- 수어 영상 (동영상 파일)
- OpenPose로 추출한 keypoint 정보와 메타데이터가 **json**, **xml** 파일 형태로 분류되어 저장



## 02. 활용한 데이터셋 소개

### 데이터셋 문장 예시

01. 화재관련, 중구 북성동에서 화재가 발생하여 진압중이니 인근 주민들께서는 안전에 유의하시기 바랍니다.  
02. 7.8 07:37 구포동 516번지 일원 단독주택 화재 발생으로 일대가 혼잡하오니 주민들께서는 외출을 자제하시기 바랍니다.  
03. 오늘 09:30 고양시 일산서구 덕이동 177-19번지 인근에 화재 발생. 이 지역을 우회하여 주시고 인근 주민은 안전사고 발생에 유의 바랍니다.

- xml 데이터
  - OpenPose 기반 2D keypoint (x, y, confidence) 정보
  - 총 137개 keypoint (얼굴 70 + 양 손 21 \* 2 + 몸 25)
- json 데이터
  - **korean\_text**, gloss, start/end frame, 비수지 정보
  - 137개 3D keypoint 정보 (Triangulation 또는 DL-based로 추정)



## 02. 파이프라인 개요

### 왜 *keypoint* 기반인가?

- CNN 기반 방식의 한계
  - 영상 데이터를 3D CNN 등으로 직접 처리
  - 영상의 과도한 크기로 인한 학습 부하
  - 배경, 화질 등 불필요한 noise의 학습
- *keypoint* 기반 방식의 장점
  - 움직임의 핵심 정보를 *keypoint*(관절 위치)로 추출
  - 입력 차원 축소를 통한 학습 속도 및 효율 향상
  - 손, 팔, 얼굴 등 주요 부위에 집중하여 수어 표현의 핵심 요소를 효과적으로 반영



OpenPose를 통해  
keypoint 추출이 완료된  
xml, json 파일 이용

## 02. 파이프라인 개요

### 양방향 수어 번역을 위한 데이터 구조 구축

- **Sign2Text**: keypoint → 한국어 문장 [SLT]
  - gloss를 거치지 않고 직접적으로 자연어 생성
  - 제공된 gloss annotation이 품질과 수량 측면에서 제한적
- **Text2Sign**: 한국어 문장 → **gloss** → keypoint [역방향]
  - 한국어 문장에 대응되는 gloss를 생성, 사전 등록된 gloss-keypoint mapping을 활용
  - 아바타 등으로 직접 수어 생성 가능성

## 02. 파이프라인 개요

### 전처리(Preprocessing) 파이프라인

- 데이터 매핑: **keypoint(xml)** ↔ **한국어 문장(json)** 병합
- keypoint 선택: face keypoint 및 confidence score 제외
- keypoint 정규화: Z-score Normalization(**Pose**), Min-Max Normalization(**Hand**)
- Skip Sampling 기반 데이터 증강(Augmentation)
- 저장 및 학습 준비

\* Neural Sign Language Translation Based on Human Keypoint Estimation (2019) 참고



## 03. Results



## 03. 주요 결과 및 모델 성능

**Sign2Text**: BiGRU Encoder, Attention, GRU Decoder

- 기본 모델: 성능 0에 수렴, 세팅 변경에도 성능 불변
  - 모델 성능의 병목은 데이터 자체에 있음으로 판단
  - (1) 문장마다 고유명사(지명, 건물명, 시간 등)가 과반수 이상 존재
  - (2) 의미 동일성 대비 표현 다양성에 따른 모델 일반화 난이도 증가
  - 해결 방향: 정규화 + 마스킹
  - (1) “주의 바랍니다” → “유의 바랍니다”로 정규화
  - (2) 고유명사 표현을 slot으로 치환, 이후 mapping을 통하여 번역

ex) <시간>: "10:10", "<지역>: "고잔동", "<주소>: "102-57번지", "<건물>: "삼성화재 광주상무사옥"

## 03. 주요 결과 및 모델 성능

**Sign2Text**: BiGRU Encoder, Attention, GRU Decoder

Model	Teacher Forcing Ratio (TFR)	Val Loss	BLEU	METEOR	ROUGE
1	0.3	8.6721	0.2565	0.3856	0.0000
2	0.1	9.5380	0.2427	0.3675	0.0000

→ 성능 향상, 그러나 여전히 특정 문장들이 과도하게 반복되고, Train Set에 과적합하는 문제 발생

## 03. 주요 결과 및 모델 성능

문장의 과도한 반복 방지 → repetition penalty 사용

고유명사 등 noise에 대한 확신 방지 → label smoothing 및 top\_k, top\_p 사용

Model	Top-k	Top-p	Repetition Penalty	Label Smoothing	TFR	BLEU	METEOR	ROUGE
1	10	0.85	1.2	0.1	0.3	0.2612	0.4159	0.0000
2	5	0.9	1.2	0.1	0.3	0.2909	0.4462	0.0000
3	5	0.9	0.9	0.1	0.5	0.3254	0.4900	0.0016
4	5	0.9	1.2	0.1	0.5	0.2936	0.4689	0.0016

## 03. 주요 결과 및 모델 성능

**Text2Sign**: Tokenizer, KoGPT2 Encoder, Multi-head Decoder

- 정규화 기준
  - [Epoch 4] Val Loss: 0.0807, mean of mode: 0.0427, gloss: 0.2273, timing: 0.0430, pointing: 0.0103
  - BLEU (pre-train): 0.0000
  - BLEU (post-train): 5.877799068880829e-156



## 03. 실제 시연

### 파이프라인

1. 영상에서 프레임 추출 (30fps 고정)
2. MediaPipe로 2D keypoint 추출 (OpenPose 대용)
3. 해상도 보정 (MediaPipe: 0~1의 값) → 정규화
4. 모델 예측 (Seq2Seq + Attention)
5. 결과는?

```
output_sentence = predict_from_keypoint(model, kpt_norm, tokenizer, device)
print("🔍 예측 결과:", output_sentence) <
```

예측 결과 : ?? < 시간 > 발생으로 < 도로 > < 도로 > < 도로 > 발생으로 발생으로 < 도로 > < 도로 >  
> 일대가 일대가 대피 일대 대피 발생으로 대피 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가  
일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가  
일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대가 일대



# 04. Conclusions

## 04. 프로젝트 한계 및 의의

### 프로젝트 한계

- OpenPose → MediaPipe 전환에 따른 구조 불일치
- Keypoint 추출 자체의 높은 난이도 (15% 이상이 유실)
- 데이터의 크기가 커 화재 관련 데이터만 사용 → 문맥 다양성의 부족
- Transformer 등 대형 모델을 위한 GPU 메모리/시간 부족
- Masking에도 불구하고 고유명사 및 희소 표현이 많아 학습에 차질
- 웹 기반 시연 시스템의 구현 미완료



## 04. 프로젝트 한계 및 의의

### 프로젝트 의의

- 관절 keypoint 기반 수어 ↔ 한국어 양방향 번역 구조 제안
- 어려운 task임에도 Sign2text 에서 BLEU 32 달성 (Baseline Model은 일반적으로 15~22)

### 프로젝트 발전 방향

- Masking 확대로 희소 고유명사에 과적합되지 않는 일반화 성능 확보
- 데이터 확대로 Transformer 기반 대형 모델 학습, Text2Sign 성능 향상
- 아바타 및 앱 구현으로 양방향 번역 실용화, 그러나 성능 향상이 선결 조건



**Thank You**