LangCon 2021

# kosp2e – 공개 가능한 한국어 음성 번역 코퍼스 구축기

서울대학교 전기정보공학부 조원익

# 목차

Language Conference

1. 음성 번역이란?
    1. 음성 번역 및 그 활용
    2. 음성 번역 데이터셋 현황
2. 음성 번역 코퍼스 구축
    1. ‘자유롭게 사용 가능한’ 음성 번역 코퍼스 만들기
    2. 음성 인식 코퍼스의 활용
    3. 기타 코퍼스의 활용
    4. Raw text만 있는 경우의 특징
3. kosp2e [코스피]
    1. 코퍼스 세부 사항
    2. 배포 현황
    3. 베이스라인 모델 성능 검증
4. 마치며...


# 음성 번역이란?

# 음성 번역이란?

Language Conference

- 음성 번역 및 그 활용
  - 음성 번역: Source language의 음성을 target language의 텍스트 혹은 음성으로 변환하는 것
    - Speech to text (S2T) translation
    - Speech to speech (S2S) translation
  - 음성 번역 방법론
    - Pipeline
      - 음성인식 + 기계 번역 (S2T)
        - +음성 합성 (S2S)
    - End-to-end
      - Fully end-to-end training
      - Pretrained models (ASR module, MT decoder etc.)

# 음성 번역이란?

Language Conference

- 음성 번역 및 그 활용
  - 음성 번역의 활용
    - From-Englsih translation이 중요할 때
      - 국제 학회의 실시간 통역
    - To-English tranlsation이 중요할 때
      - 특정 언어가 있는 영상의 번역 (혹은 자막 달기)
      - 여행 시 영어로의 동시 통역
    - 두 방향이 모두 중요할 때
      - 영어를 공용어로 한 국제 회의
  - `한국어' 음성 번역은?
    - 상대적으로 less studied
    - End-to-end에서만 캐치할 수 있는 영역이 존재할 수 있음
    - 한국어를 메인으로 제작한 영상 (VLive 등) 에 좀 더 적극적으로 활용 가능!

# 음성 번역이란?



- 음성 번역 및 음성인식/번역 데이터셋 현황
  - End-to-end
    - MuST-C (Di Gangi et al., 2019)
      - English speech of TED talks, its transcript, and the translation to other IE languages
    - CoVoST (Wang et al., 2020)
      - Enables multilingual speech translation that bases on Common Voice (CV) data
    - Europarl-ST (Koehn, 2005)
      - Various translations for debates in European Parliament
  - ASR benchmarks
    - Librispeech (Panayotov et al., 2015) / TedLium
  - MT benchmarks
    - WMT datasets (Bojar t al., 2016, 2018; Barrault et al., 2020) / Open subtitles
  - For more information
    - https://github.com/kahne/SpeechTransProgress

# 음성 번역이란?

Language Conference

- 음성 번역 및 음성인식/번역 데이터셋 현황
  - 인구어 혹은 중국어에 집중되어 있음
  - 한국어를 source speech로 가지는 데이터셋의 부재
  - 음성 번역의 다양한 목적 달성을 위해서는 Korean을 source speech로 하는 데이터셋 역시 필요
  - 다른 음성 번역 데이터셋처럼 annotate on하고 version up할 수 있도록, 공개 가능하고 remix/redistribution이 가능한 데이터셋이 필요

- kosp2e [코스피]
  - Korean speech to English translation corpus
  - In corporation with NAVER Papago
  - Accepted at Interspeech 2021

SEOUL NATIONAL UNIVERSITY

Human Interface Laboratory

NAVER papago


# 음성 번역 코퍼스 구축

LangCon 2021

# 음성 번역 코퍼스 구축

Language Conference

- ‘자유롭게 사용 가능한’ 음성 번역 코퍼스 만들기
  - 자유롭게 사용 가능한 것이란?
    - Community contribution에 열려 있는
      - 재배포의 자유
      - 수정의 자유
    - Accessible한
      - 별도의 등록 없이도 접근 가능한
    - Research에, 혹은 research와 industry 모두에 활용 가능한
  - 어떻게 가능한가?
    - 완전히 새롭게 구축하기
    - License가 자유로운 corpus 기반으로 구축하기

# 음성 번역 코퍼스 구축

Language Conference

- ‘자유롭게 사용 가능한’ 음성 번역 코퍼스 만들기
  - 완전히 새롭게 구축하기
    - Three steps
      - Script 만들기
      - 녹음하기
      - 번역하기
    - 굉장히 품이 많이 든다
  - License가 자유로운 코퍼스 기반으로 구축하기
    - 활용 가능한 선택지
      - 음성 인식 코퍼스 (음성 확보)
        - KSS, Zeroth
      - 기타 annotated / raw corpus (스크립트 확보)
        - StyleKQC, Covid-ED

# 음성 번역 코퍼스 구축



- 음성 인식 코퍼스의 활용
  - KSS **(thanks to Kyubyong Park)**
    - Korean Single Speaker Speech corpus
    - 약 13,000개의 single speaker 녹음 / 스크립트 / 번역
    - Textbook domain
  - Zeroth **(thanks to Taeyoung Jo)**
    - 3,000여개의 unique한 문장들로 구성된 22,000여 문장의 녹음
    - 115명의 발화자
    - News domain의 문장들이 대부분
    - 한국어 ASR 학습 및 평가에 널리 사용

# 음성 번역 코퍼스 구축

Language Conference

- 음성 인식 코퍼스의 활용
  - KSS 코퍼스의 활용
    - Single speaker로만 제공되는 것은 다양한 목소리를 포괄하지 못할 수 있기에, 같은 문장으로 새로이 녹음
    - Translation이 있기에 이를 적극 활용
      - 어떻게 읽어야 할지 애매한 경우
      - 원본 음성 역시 활용 가능
  - Zeroth 코퍼스의 활용
    - 음성이 이미 주어져 있기에, 이에 따른 번역만을 진행
      - 뉴스 script이기에, 기자나 아나운서, 앵커 등의 이름이 호칭으로써 등장하는 경우가 존재 – 이 때 comma를 가정하고 번역할 것을 요청
      - 인명, 장소, 사건 등의 named entity 등에 대한 번역 방식의 통일 – 공식 명칭이 있을 때는 그를 사용하고, 없다면 revised romanization 할 것

# 음성 번역 코퍼스 구축



- 기타 코퍼스의 활용
  - 음성인식 코퍼스는 아니지만, 공개된 텍스트를 활용 가능
  - StyleKQC
    - Provided by **Human Interface Laboratory @SNU**
    - 인공지능 스피커 도메인의 30,000 문장 및 해당 문장의 메타데이터 포함
    - 6가지의 topic, 4가지의 speech act가 기재
  - Covid-ED
    - Covid-19 Emotion Diary with Empathy and Theory-of-Mind Ground Truths Dataset)
    - Provided by **Human Factors Psychology Lab @SNU**
    - 코로나 상황에서의 일기 데이터를 크라우드소싱을 위해 수집
    - 일기 별로 ground truth 감정을 최대 두 개씩 기재

# 음성 번역 코퍼스 구축



- 기타 코퍼스의 활용
  - StyleKQC
    - 특징
      - 도치, 주어 생략, 간투사 등이 포함된 구어 스크립트
      - 원문이 문장 부호를 포함하지 않음
      - 같은 intent에 대해 각 5개의 informal한 문장 및 formal한 문장 (10문장 1set) 이 존재
    - 녹음
      - Speech act를 녹음 과정에서 반영하여 intonation을 줄 것
      - 도치 및 간투사 등이 텍스트 상 어색해도, 이를 duration 등을 통해 조절해 가면서 읽을 것
      - 문장이 informal한가 formal한가에 따라 읽는 tone을 다르게 할것 (동료/친구에게 이야기하는 것 / 손윗사람에게 이야기하는것)
    - 번역
      - Formality가 번역 과정에서 반영될 수 있게 할것 (사용되는 단어의 formality나 문장 구조 등)
      - 도치 및 간투사 역시 번역 과정에서 반영될 수 있게 할 것

- 기타 코퍼스의 활용
  - StyleKQC

| | id | act | style | sentence | translation |
|---|---|---|---|---|---|
| 0 | 0_0 | altQ | formal | 에이에스 서비스가 삼성이랑 애플 중에 어디가 더 좋은가요 | Which one offers better warranty services, Samsung or Apple? |
| 1 | 0_1 | altQ | formal | 삼성 애플 중에 에이에스 잘 해주는 곳 좀 알려주세요 | Please find out which one has better warranty services, Samsung or Apple? |
| 2 | 0_2 | altQ | formal | 어디가 더 에이에스 잘해주나요 삼성이랑 애플 중에 | Which one provides better warranty services? Samsung or Apple? |
| 3 | 0_3 | altQ | formal | 삼성이랑 애플 어디가 더 에이에스가 좋은지 좀 확인 부탁드립니다 | Please check if it is Samsung or Apple that has better warranty services. |
| 4 | 0_4 | altQ | formal | 삼성 애플 어디가 더 에이에스 서비스 잘해주는지 좀 알아봐 주시겠어요 | Could you please find out which one offers better warranty services between Samsung and Apple? |
| 5 | 0_5 | altQ | informal | 에이에스 삼성이랑 애플 중 어디가 더 낫냐 | What's the better one in terms of warranty services, Samsung or Apple? |
| 6 | 0_6 | altQ | informal | 삼성이랑 애플 중에 에이에스 잘 해주는 곳 좀 알려줘봐봐 | Tell me which one has better warranty services, Samsung or Apple? |
| 7 | 0_7 | altQ | informal | 어디가 더 에이에스 잘해 삼성 애플 중에서 | Which one has better warranty services? Samsung or Apple? |
| 8 | 0_8 | altQ | informal | 삼성 애플 어디가 더 에이에스 서비스가 더 좋은지 확인 좀 | Check which one has better warranty services between Samsung and Apple. |
| 9 | 0_9 | altQ | informal | 삼성이랑 애플 어디가 더 에이에스 잘해주는지 좀 알아봐 줘 | Please check if it is Samsung or Apple that provides better warranty services. |

# 음성 번역 코퍼스 구축



- 기타 코퍼스의 활용
  - Covid-ED
    - 특징
      - 작가 한 명 당 5개의 일기 작성
      - 각 일기는 5~15여개의 문장으로 구성
      - Web text이기 때문에, 다른 코퍼스들과 달리 text에서만 활용될 수 있는 표현 (영단어, 특수기호, leetspeak 등) 들이 포함됨
    - 녹음
      - 작가 한 명의 일기는 한 사람이 전부 녹음할 수 있도록 할 것
      - 작가의 나이/성별과 녹음자의 나이/성별이 어느 정도 일치할 수 있도록 할 것
      - 제공된 감정을 모든 문장에 반영하지는 않아도, 각 일기에서 relevant한 문장에는 반영할 것
      - 영단어/숫자는 독법에 맞게, 특수기호는 확실한 경우는 읽고 그렇지 않으면 implicit하게 반영
    - 번역
      - 일기의 특징 상 주어 생략이 빈번할 수 있지만, 최대한 각 문장이 독립적으로 활용될 수 있도록 번역할 것
      - 영단어는 그대로 번역할 것
      - 숫자 및 특수기호의 경우 녹음과 동일한 컨벤션을 제공

# 음성 번역 코퍼스 구축

Language Conference

- 기타 코퍼스의 활용
  - Covid-ED

| | id | username | age | gender | doc# | sen# | sentence | translation | emotion1 | emotion2 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | author680_59_female_1_1 | author680 | 59 | female | 1 | 1 | 올해 과일 야채 가격은 정말 어마어마하다. | This year, the prices of fruits and vegetables are incredible. | 화남 | 불쾌함 |
| 1 | author680_59_female_1_2 | author680 | 59 | female | 1 | 2 | 작년보다 3배 4배는 보통이다. | It's normal for them to jump up three or four times. | 화남 | 불쾌함 |
| 2 | author680_59_female_1_3 | author680 | 59 | female | 1 | 3 | 코로나 19로 힘들 뿐만 아니라, 태풍, 장마 온갖 자연재해로 그동안 살아온 세상과는 다른 세상을 사는 기분이 들 때가 있다. | Sometimes, it feels like I'm living in a different world because of difficulties from COVID-19, as well as various natural disasters, such as typhoons and monsoons. | 화남 | 불쾌함 |
| 3 | author680_59_female_1_4 | author680 | 59 | female | 1 | 4 | 멀리 떨어진 마트에서 몇 박스 한정 고구마 10kg을 세일을 한다는 소식을 듣고 꼭 사고 싶다는 생각으로 마음이 조급해졌다 | I heard the news that a supermarket far from my place will have limited sales of 10 kg boxes of sweet potatoes, and I got anxious to buy it. | 화남 | 불쾌함 |
| 4 | author680_59_female_1_5 | author680 | 59 | female | 1 | 5 | 전날 밤부터 멀리까지 갔는데 사지 못하고 허탕 치면 어떡하지 하는 걱정으로 자는 중간에도 생각이 났다. | Even the night before, I was worried about what I must do if I don't get it, even if I've traveled far for it. | 화남 | 불쾌함 |
| 5 | author680_59_female_1_6 | author680 | 59 | female | 1 | 6 | 마트 개점시간에 맞추어 가는 것은 좀 자존심이 상해서 적당한 시간을 찾아서 그래도 평소보다는 일찍 마트로 갔다. | It hurts my pride to get there at the opening hour, so I visited during an appropriate time but earlier than usual. | 화남 | 불쾌함 |
| 6 | author680_59_female_1_7 | author680 | 59 | female | 1 | 7 | 도착하니 다행스럽게도 고구마 박스들이 마트 문밖에 쌓여 있었다. | When I arrived, thankfully, boxes of sweet potatoes were stacked in front of the entrance of the supermarket. | 화남 | 불쾌함 |
| 7 | author680_59_female_1_8 | author680 | 59 | female | 1 | 8 | 우르르 몰린 아줌마들이 직원들이 눈치 해도 상관하지 않고 고구마 상태를 살피며 이 박스 저 박스 헤집어 놓고 있었다. | A crowd of middle-aged women were going through each box, despite employees' glares. | 화남 | 불쾌함 |
| 8 | author680_59_female_1_9 | author680 | 59 | female | 1 | 9 | 그렇게 극성스럽게는 하고 않아 다른 사람이 펼쳐놓은 것 중 하나를 집어 카트에 넣고 다른 물건을 사고 계산하고 나왔다. | I didn't want to make such a fuss, so I just picked one box that someone has opened already, put it in the cart, and paid for it along with some other items. | 화남 | 불쾌함 |
| 9 | author680_59_female_1_10 | author680 | 59 | female | 1 | 10 | 물건들을 마트 카트에서 내가 가져간 작은 카트에 옮겨 싣고 있는데 마트 직원이 고구마 계산했냐고 물었다. | One of the supermarket employees asked me whether I paid for it as I was unloading the items from the cart to the small cart that I've brought with me. | 화남 | 불쾌함 |
| 10 | author680_59_female_1_11 | author680 | 59 | female | 1 | 11 | 반사적으로 영수증 보여 드릴까요? 하는 말이 나왔고 기분이 확 나빠졌다. | Instinctively, I said, "Do you want to see the receipt?" And it made me turn green all of a sudden. | 화남 | 불쾌함 |
| 11 | author680_59_female_1_12 | author680 | 59 | female | 1 | 12 | 내가 있는 곳이 고구마가 있는 곳과 멀지 않은 곳이어서 그랬나 보다 하고 아무리 좋게 생각하려고 해도 의심을 받았다는 것이 기분 좋을 리 없다. | Although I've tried to be understanding, thinking that perhaps it's because I was standing not too far from the boxes of sweet potatoes, but it's still unpleasant to have been looked upon with suspicious eyes. | 화남 | 불쾌함 |
| 12 | author680_59_female_1_13 | author680 | 59 | female | 1 | 13 | 고구마를 볼 때마다 찝찝하게 생각이 난다. | It reminds me of that unpleasant incident whenever I look at sweet potatoes. | 화남 | 불쾌함 |
| 13 | author680_59_female_1_14 | author680 | 59 | female | 1 | 14 | 그래서인지 고구마 맛도 별로다. | Perhaps it's because of that, but they don't taste that good either. | 화남 | 불쾌함 |
| 14 | author680_59_female_1_15 | author680 | 59 | female | 1 | 15 | 몇 푼 아끼자고 거기까지 가서 좋은 소리도 못 듣고..... | Getting that kind of treatment for traveling that far to save a few pennies. | 화남 | 불쾌함 |

# 음성 번역 코퍼스 구축

Language Conference

- Raw text만 있는 경우의 특징
  - 한국어 script의 정제가 완벽할 필요까지는 없음
    - 음성과 번역을 pairing하는 것이기 때문에, intermediate한 한국어 script가 완벽할 필요는 없음 (ASR 성능을 체크하는 것이 아니기에)
    - 번역 convention을 통일하는 것이 더 중요함
  - 음성과 번역이 완전히 align되지 못할 가능성도 존재
    - 영어보다 더 prosody에 민감하기 때문
    - StyleKQC에서는 speech act를, Covid-ED에서는 emotion을 metadata로 제공함으로써 그럴 가능성을 최소화
  - 웬만하면 음성인식/번역 코퍼스 사용하자!
    - 자유롭게 사용 가능한, 퀄리티 좋은 번역 코퍼스 확보의 필요성


# kosp2e [코스피]

LangCon 2021

- 코퍼스 세부 사항

| Dataset | License | Domain | Characteristics | Volume (Train / Dev / Test) | Tokens (ko / en) | Speakers (Total) |
|---|---|---|---|---|---|---|
| **Zeroth** | CC-BY 4.0 | News / newspaper | DB originally for speech recognition | 22,247 utterances (3,004 unique scripts) (21,589 / 197 / 461) | 72K / 120K | 115 |
| **KSS** | CC-BY-NC-SA 4.0 | Textbook (colloquial descriptions) | Originally recorded by a single speaker (multi-speaker recording augmented) | 25,708 utterances = 12,854 * 2 (recording augmented) (24,940 / 256 / 512) | 128K / 190K | 17 |
| **StyleKQC** | CC-BY-SA 4.0 | AI agent (commands) | Speech act (4) and topic (6) labels are included | 30,000 utterances (28,800 / 480 / 720) | 237K / 391K | 60 |
| **Covid-ED** | CC-BY-NC-SA 4.0 | Diary (monologue) | Sentences are in document level; emotion tags included | 32,284 utterances (31,324 / 333 / 627) | 358K / 571K | 71 |

- 각 subcorpus의 라이센스는 원 코퍼스의 라이센스를 따름

# kosp2e [코스피]



- 코퍼스 세부 사항
  - Utterances: 110,239 / Hours: 198H
  - Source language (tokens): Korean (795K)
  - Target language (tokens): English (1,272K)
  - Speakers: 263 (in total)
  - Domains: News, Textbook, AI agent command, Diary

- 배포 현황
  - Github repository를 통해 이루어짐
    https://github.com/warnikchow/kosp2e
    - Speech files 및 Train/Dev/Test 파일 리스트/영문 번역 freely downloadable
    - Provided under request
      - Korean scripts
      - Other metadata (for StyleKQC and Covid-ED)

# kosp2e [코스피]

Language Conference

- 배포 현황

## kosp2e

Korean Speech to English Translation Corpus

## Dataset

### Freely available

- Speech files
- Train/Dev/Test filenames' list their English translation

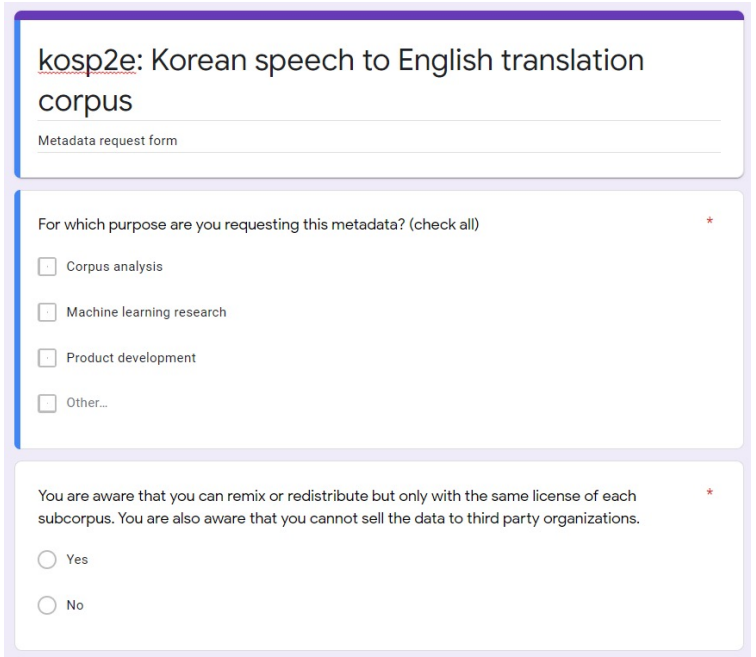
### Provided under request (in this link)

- Korean scripts
- Other metadata (for StyleKQC and Covid-ED)

## Howto

```
git clone https://github.com/warnikchow/kosp2e
cd kosp2e
cd data
wget https://www.dropbox.com/s/y74ew1c1evdoxs1/data.zip
unzip data.zip
```

Then you get the folder with speech files (*data* and subfolders) and split files' list (*split* and .xlsx files).



kosp2e: Korean speech to English translation corpus

Metadata request form

For which purpose are you requesting this metadata? (check all) *

- Corpus analysis
- Machine learning research
- Product development
- Other...

You are aware that you can remix or redistribute but only with the same license of each subcorpus. You are also aware that you cannot sell the data to third party organizations. *

- Yes
- No

- 베이스라인 모델 성능 검증
  - ASR-MT pipeline
    - SpeechRecognition 음성인식 툴킷
    - 활용 가능한 MT API들
      - Pororo
      - PAPAGO
  - End-to-end models: fairseq S2T의 MuST-C recipe를 기반
    - Vanilla transformer architecture
    - ASR pretraining 후 decoder part를 fine-tuning
      - ASR은 AI HUB의 한국어 음성인식 코퍼스 활용
    - Synthetic한 음성 번역 데이터로 warm-up 후 fine-tuning하기
      - AI HUB 한국어 음성인식 코퍼스를 PAPAGO API를 통해 번역하여 사용

# kosp2e [코스피]

Language Conference

- 베이스라인 모델 성능 검증
  - 대체적으로 ASR-MT pipeline 성능이 end-to-end model 보다 높음
    - En-De 등에서 보여준 경향과 다르게, 한국어 syntax가 영어와 판이한 점도 원인으로 분석됨
  - Pretraining 방법론의 적용은 매우 효과적
    - ASR pretraining보다 warm-up 이 더 효과적이라는 것은, ko-en MT 역시 쉽지 않음을 암시
  - 고려해야 할 점들
    - Pretraining corpus와의 overlap
    - Subcorpus 별 경향성

| Model | BLEU | WER (ASR) | BLEU (MT/ST) |
|---|---|---|---|
| ASR-MT (Pororo) | 16.6 | 34.0 | 18.5 (MT) |
| ASR-MT (PAPAGO) | 21.3 | 34.0 | 25.0 (MT) |
| Transformer (Vanilla) | 2.6 | - | - |
| ASR pretraining | 5.9 | 24.0* | - |
| Transformer + Warm-up | *8.7* | - | 35.7 (ST)* |
| + Fine-tuning | *18.3* | - | - |

# kosp2e [코스피]

Language Conference

- 베이스라인 모델 성능 검증
  - Vanilla transformer model에 대한 fairseq 기반 레시피 제공
    - 특정 버젼 fairseq 설치 필요
  - Advanced model 들의 구현에 관해선 paper에 서술
    - 현재 별도 배포 x
    - 추후 논의 예정

## Recipe

- Fairseq is required for the basic recipe. You may install specific fairseq version for replication.

```
wget https://github.com/pytorch/fairseq/archive/148327d8c1e3a5f9d17a11bbb1973a7cf3f955d3.zip
unzip 148327d8c1e3a5f9d17a11bbb1973a7cf3f955d3.zip
pip install -e ./fairseq-148327d8c1e3a5f9d17a11bbb1973a7cf3f955d3/

pip install -r requirements.txt
```

- First, you preprocess the data, and then prepare them in a format that fit with transformer. Other part follows fairseq S2T translation recipe with MuST-C.
- This recipe leads you to the *Vanilla* model (the most basic end-to-end version). For the advanced training, refer to the paper below.

```
python preprocessing.py

python prep_data.py --data-root dataset/ --task st --vocab-type unigram --vocab-size 8000

fairseq-train dataset/kr-en  --config-yaml config_st.yaml \
--train-subset train_st --valid-subset dev_st --save-dir result --num-workers 4 \
--max-tokens 40000 --max-update 50000 --task speech_to_text \
--criterion label_smoothed_cross_entropy --report-accuracy \
--arch s2t_transformer_s --optimizer adam --lr 2e-3 --lr-scheduler inverse_sqrt \
--warmup-updates 10000 --clip-norm 10.0 --seed 1 --update-freq 8 --fp16
```


# 마치며...

# 마치며...

Language Conference

- 의의 및 활용 방안
  - 한국어 종단형 음성 번역 연구를 위한 첫걸음
  - CC-BY 라이센스로 자유롭게 이용 가능하며, 수정 재배포 가능한 형태로 community contribution을 장려하는 데이터
  - 한국어와 영어가 판이한 syntax를 보여주지만, 종단형 음성 번역이 다양한 방법으로 가능하다는 것을 시사
  - 여러 가지 pretrained ST module에 적용하여, fine-tuning으로 도메인 별 성능을 높일 수 있는 데이터로 활용 가능


  - 네이버 파파고, 그린웹, 에버영, 그리고 렉스코드에 깊은 감사를 드립니다


greenweb service

EVERYOUNG KOREA

LEXCODE
언어는 장벽이 아니라 기회입니다


감사합니다!