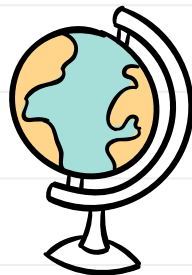
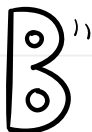
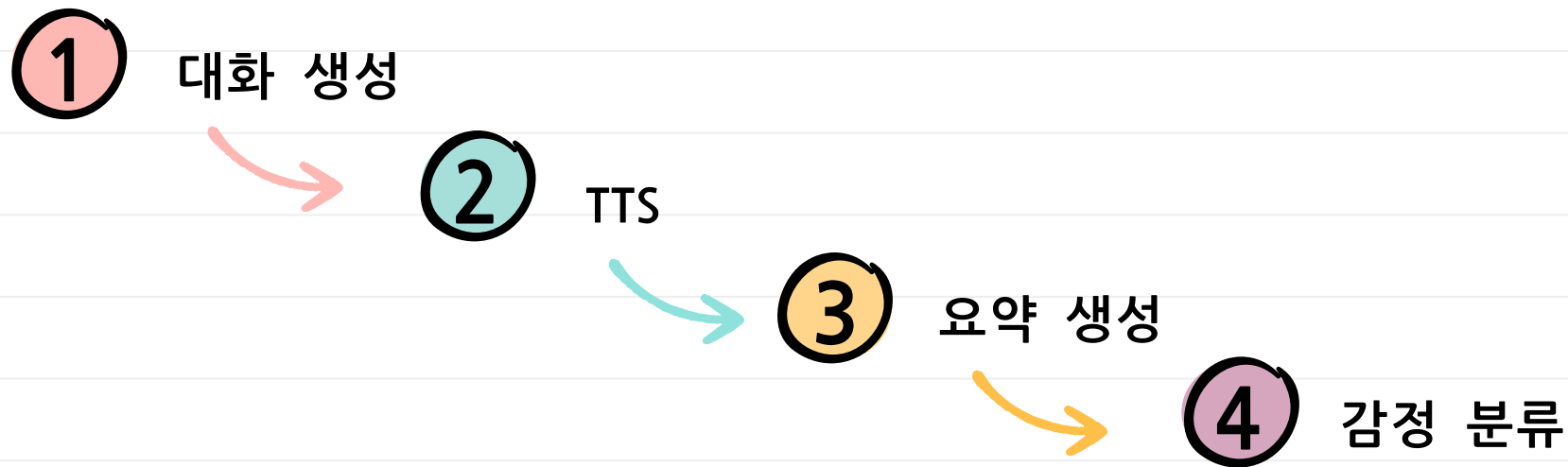


초등학생을 위한 한영 챗봇,  
코잉봇  
(KOR-ENG BOT)



## 2. 서비스 개요



# 대화 예시 (실제 저희 모델의 답변으로 구성된 내용입니다.)

## 1 대화 생성

2022년 5월 28일 토요일

유튜브가 너무 재밌어서 고민이야. 공부보다 더 재밌어.  
I'm worried because YouTube is so fun. It's more fun than studying.

유튜브가 너무 재밌어서 고민이시군요. 유튜브는 공부하는데 어떤 도움이 될까요?

You are worried because YouTube is so fun. How can YouTube help you study?

유튜브를 보면서 스트레스를 풀 수 있어.  
You can relieve stress by watching YouTube.

유튜브로 스트레스를 푸시는군요.  
YouTube is a way to relieve stress.

엄마는 내가 공부를 안하고 유튜브만 볼까봐 걱정하는 거 같아.  
I think my mom is worried that I won't study and only watch YouTube.

엄마가 내가 공부를 안하고 유튜브만 볼까봐 걱정하는 거 같군요. 조금 더 자세히 말씀해 주시겠어요?

## 2 TTS



## 2. 서비스 개요

It seems that your mother is worried that you will not study and only watch YouTube. Could you please elaborate a little more?

내 학교 점수가 떨어질까봐... 나는 유튜브 조금만 보고, 공부 열심히 할 수 있는데.  
I'm afraid my school grades will drop... I watch a bit of YouTube I can study hard.

성적이 떨어질까 봐 두려우시군요.  
You are afraid of your grades going down.

## 3 요약 생성

2022년 5월 28일

## 4 감정 분류

승민이의 오늘의 감정은 **불안** 입니다.

유튜브를 보면서 스트레스를 풀 수 있는데 엄마는 공부를 안 하고 유튜브만 볼까 봐 걱정하는 것 같다.



**AI 모델 구현**

**AI MODEL DEVELOPMENT**

### 3. 모델 구현

#### 1. 대화 생성

#### 2. TTS 음성

#### 3. 요약 생성

#### 4. 감정 분류

모델

KoBART

GLow-TTS(한글)  
& Google tts api  
(영어)

KoBART

KoBERT

데이터

AI HUB  
웰니스 대화  
한국어 대화  
감성대화 말뭉치  
  
송영숙 챗봇데이터  
  
총 69,354개

네이버 클로바  
하준이 목소리  
  
총 3,900개

AI HUB  
한국어 대화  
요약  
  
총 314,997개

AI HUB  
감성대화  
말뭉치  
  
총 46,009개

# 1

## 대화 생성 - 데이터 선정

## 3. 모델 구현

(1)

(2)

(3)

(4)

### 감성 대화 말뭉치 소개

소개

다운로드

데이터셋명	감성 대화 말뭉치	데이터 출처
데이터 분야	음성/자연어	데이터 사용처
구축기관	미디어젠	데이터 제공처
가공기관	미디어젠	데이터 가공처
검수기관	미디어젠	데이터 검수처
구축 데이터량	10,000문장 (음성), 27만문장 (코퍼스)	구축년도
버전	1.0	최종수정
소개	크라우드 소싱 수행으로 일반인 1,500명을 대상으로 하여 음성 통해 감성 대화 연진을 개발하여 세대별 감성 대화 서비스 제공	
주요 키워드	코퍼스(Corpus), 감성대화(Emotional dialogue), 감성 챗봇	

### 웰니스 대화 스크립트 데이터셋

① 외부 데이터는 해당 기관의 이용정책과 다운로드 절차를 따라야 하며 시 허브와 관련이 있음

#### 구축량

정신건강 상담 주제의 359개 대화의도에 대한 5,232개의 사용자 발화 및 1,000개의 상담사 응답

#### 대표 도면

대분류	소분류	사용자 발화
감정	감정조절 이상	더 이상 내 감정을 내기 컨트롤 못 하겠어.
감정	걱정	계속 이렇기 봐 너무 걱정돼.
감정	자괴감	저는 왜 이렇게 못난 인간일까요...
배경	남편	남편이랑 한달에 몇번은 보는 것 같아요.

### Chatbot\_data.

Chatbot\_data\_for\_Korean v1.0

#### Data description.

인공데이터입니다. 일부 이별과 관련된 질문에서 다음카페 "사랑보다 인연"에서 자주 나오는 질문을 기반으로 생성되었습니다. (예: "별한 지 얼마(또는 100일) 되었어요"라는 질문에 "찾고 있어"라는 답변이 생성된 것)

1. 챗봇 트레이닝용 문장 페어 11,876개
2. 일상대화 0, 이별(부정) 1, 사랑(긍정) 2로 레이블링

#### Quick peek.

ChatbotData	
Q	A
12시 땀	하루가 또 가네요.
1지방 학교 떨어졌어	위험해 드립니다.
3억4월 놀라가고 싶다	여행은 언제나 좋죠.
3억4월 정도 놀라가고 싶다	여행은 언제나 좋죠.
PPL 심하네	눈살이 찌푸라지.
SD카드 망가졌어	다시 새로 사는 게 마음 편해요.

### 한국어 대화 데이터셋

① 외부 데이터는 해당 기관의 이용정책과 다운로드 절차를 따라야 하며 시 허브와 관련이 있음

#### 필요성

- 딥러닝 기반의 한국어 대화 모델의 개발을 위한 한국어 및 상황에 맞는 맞춤 데이터
- 도메인 특화된 모델을 위해서는 상황에 특화된 데이터가 요구됨
- 응급 상황 및 오픈스 task 지원을 위한 인공지능 대화 모델 개발에 활용하기 위한 데이터

#### 구축 내용



주제에 맞는 동영상 (해와 포함)을 참조하여 한국어 및 황에 적절하도록 각색하여 구축



초등학생을 위한  
심리상담 및 일상대화 데이터 사용!

## 1

## 대화 생성 - 데이터 전처리

## 3. 모델 구현

## 정규식 처리

ㅋㅋ, ㅌㅌ 등 제거 &  
공백 여러 개는  
한 개의 공백으로 처리



## 품사 태깅

KoNLPy의 OKT를 이용  
품사 태깅 진행 &  
훈련 데이터로 사용



## 특정 데이터 제거

[ 출근, 퇴사, 술벗, 술,  
음주운전, 허니문, 산부인과,  
불륜, 치맥, 남편, 아내, 대출... ]  
등의 어린이에게 적절하지 않은  
내용의 데이터 제거



나 오늘 길 가다가 넘어졌어 ㅍㅍ  
사람들이 날 다 바라보더라.. ㅎㅎ..



나 오늘 길 가다가 넘어졌어  
사람들이 날 다 바라보더라.

나 오늘 길 가다가 넘어졌어  
사람들이 날 다 바라보더라.



'나', '오늘', '길', '가다가',  
'넘어지다', '사람', '이',  
'날', '바라보다', '.'

오늘 같은 날씨에는 치맥이지!

아 퇴사하고 싶다.

어디 내 술벗 해줄 사람 없나?

# 1

## 대화 생성 - 모델 선정

### 3. 모델 구현

나 오늘 친구랑 싸웠어.

S-BERT

신경 쓰이는 일이 있으시군요. 그럴 땐 자면서 긴장을 푸는 게 최고죠.

KoGPT2

친구랑 싸우셨군요. 무슨 일이 있으셨나요.

KoBART

속상하셨겠어요. 왜 싸우셨는지 여쭙봐도 될까요?

엄마가 이 썩는다고 사탕  
조금만 먹으래.

S-BERT

아침부터 혼구멍이 나서 기분이 안 좋아 보이네요.

KoGPT2

조금만 참으시면 나아지실 거예요.

KoBART

엄마가 이 썩는다고 사탕을 조금만 먹으라고 하셨군요. 조금 더 자세히 말씀해 주시겠어요?

나 반에서 시험성적 일등이야.

S-BERT

이별을 받아들이는 것도 중요해요.

KoGPT2

좋은 부모님이 계셔서 정말 좋으시겠어요.

KoBART

시험성적 일등이시군요. 기쁘시겠어요.



# 1

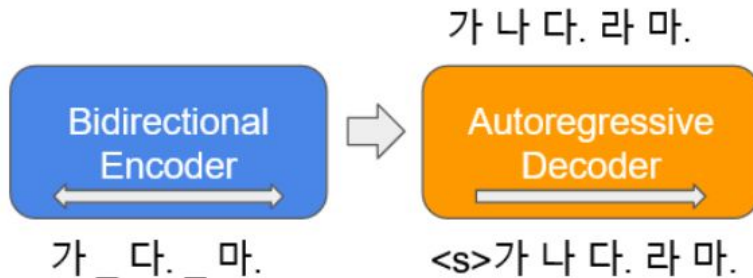
## 대화 생성 - 최종 모델 선정

### 3. 모델 구현

#### KoBART



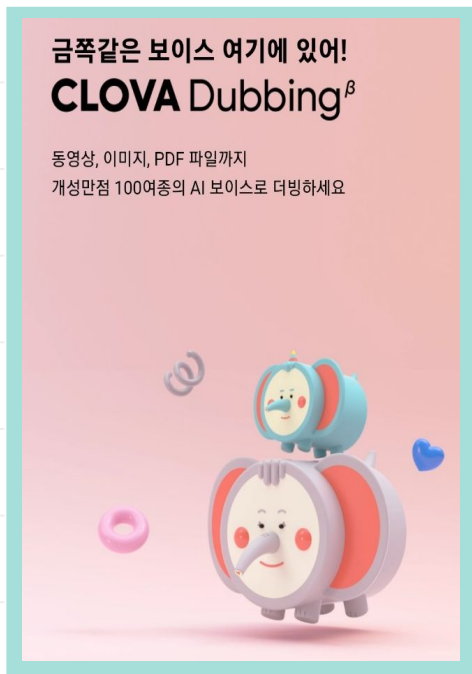
SKT에서 공개한  
문서 요약 모델로,  
40GB 이상의 한국어  
텍스트에 대해 학습한  
한국어 encoder-decoder  
언어 모델



## 2

## TTS - 음성

## 3. 모델 구현



네이버 클로바 무상 제공 - 하준이  
3900개 음성 파일 저장

Glow-TTS

카카오 엔터프라이즈  
AI-lab와 서울대학교에서  
제안한 TTS 모델로  
입력된 텍스트의 순서에  
따라 발화를  
차례대로 정렬



TTS

# 3

## 요약 생성

## 3. 모델 구현

### 한국어 대화 요약 소개

소개

다운로드

데이터셋명	한국어 대화 요약
데이터 분야	음성/자연어
구축기관	바이브컴퍼니
가공기관	바이브컴퍼니
검수기관	바이브컴퍼니
구축 데이터량	35만건
버전	1.3
소개	일상 대화, 토론 등 다양한 유형의 한국어 대화 원 텍스트 데이터
주요 키워드	한국어 대화 요약, 자연어처리, 원문, 요약문, 구0

AI HUB 한국어  
대화 요약

### KoBART

SKT에서 공개한  
문서 요약 모델로,  
40GB 이상의 한국어  
텍스트에 대해 학습한  
한국어 encoder-decoder  
언어 모델

KoBART

(2020. 10)



encoder-decoder model

# 4

## 감정 분류

### 3. 모델 구현

AI HUB 감성대화  
말뭉치

채택!

KoBERT

SKT-Brain에서 만든  
한국어 버전의 BERT  
자연어 처리 모델

감성 대화 말뭉치 소개

소개

다운로드

데이터셋명 감성 대화 말뭉치

데이터 분야 음성/자연어

데이터

분노

상처

슬픔

당황

불안

기쁨

70%의 정확도

# 서비스 조직도

## 3. 모델 구현

