청년 AI Bigdata 교육

Machine translation & Speech

How?

토큰화

```
from transformers import BertTokenizer

tokenizer = BertTokenizer.from_pretrained("bert-base-cased")
```

```
tokenizer("Using a Transformer network is simple")
```

```
{'input_ids': [101, 7993, 170, 11303, 1200, 2443, 1110, 3014, 102],
 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0],
 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1]}
```

임베딩

- 희소 표현

  벡터의 유의미한 유사성을 표현할 수 없음

  ex) one-hot 방식

- 밀집 표현

  벡터의 차원을 단어 집합의 크기로 상정하지 않음.

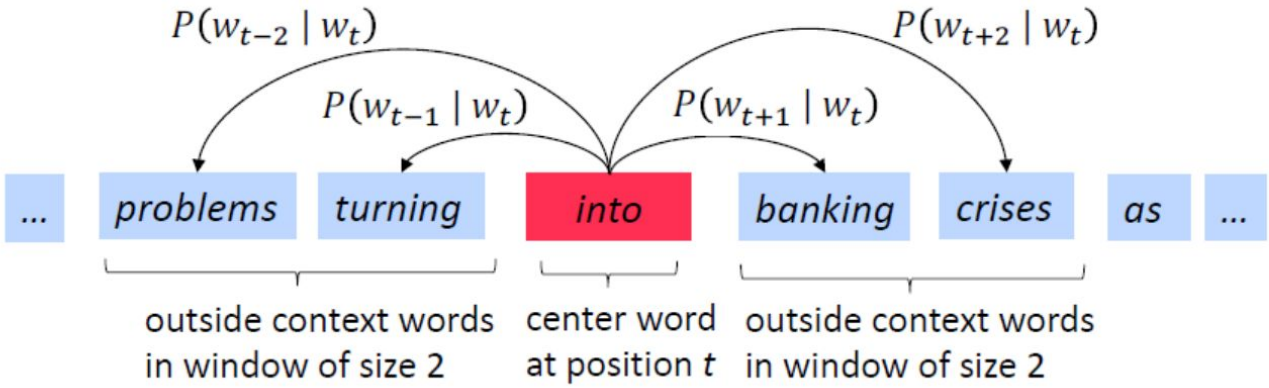
  사용자가 설정한 값으로 모든 벡터표현의 차원을 맞춤

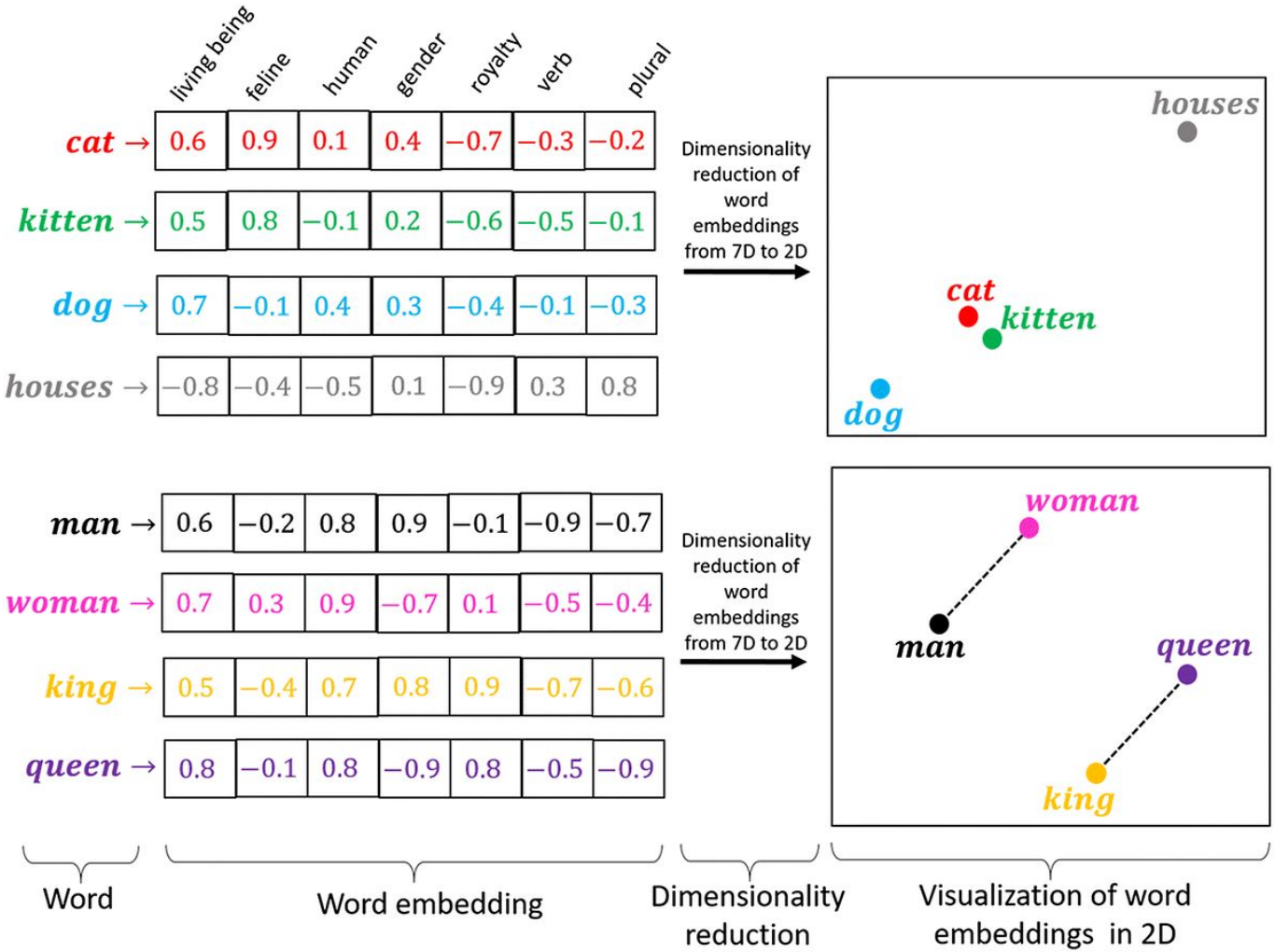
  ex) embedding 방식

Word Vector

- Distributional semantics : A word's meaning is given by the words that frequently appear close-by
- *"You shall know a word by the company it keeps"*
- Word2vec objective function (skip-grams)

$$J(\theta) = -\frac{1}{T}\log L(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m\le j\le m \\ j\neq 0}} \log P(w_{t+j} \mid w_t;\theta)$$



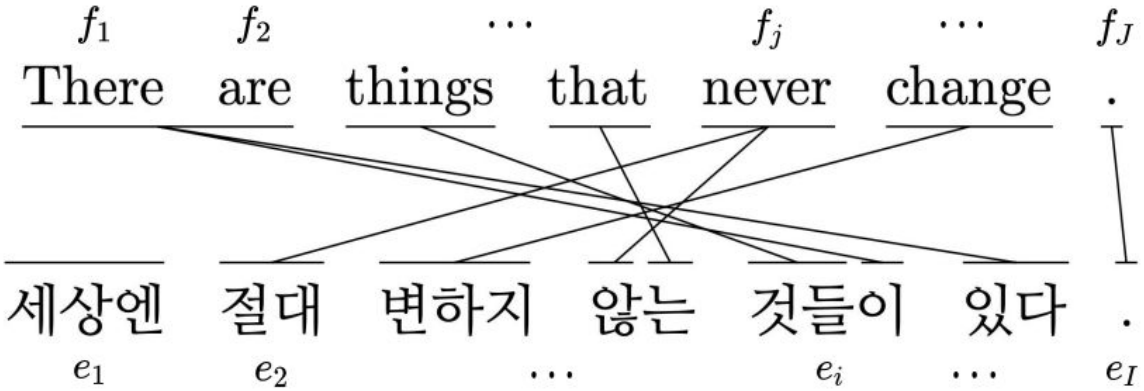

living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

https://www.youtube.com/watch?v=0RacJ0MQcDA

NMT(Neural Machine Translation)

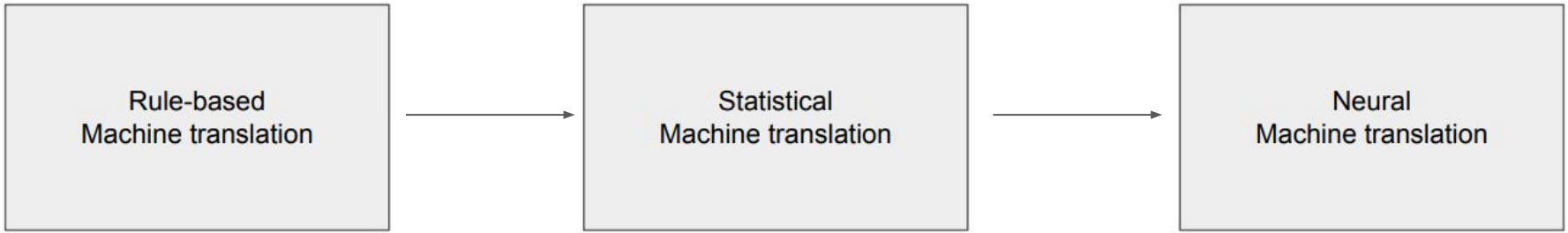
Machine Translation

- 기계번역은 A언어의 문장을 B언어의 문장으로 바꾸어주는 Task



- Source length ≠ Target length

Machine Translation



Machine Translation

- 규칙 기반 기계번역
  - 개별 단어로 접근한 뒤, 문법과 같은 규칙을 적용하여 번역
  - 문법의 용례를 파악하는데에는 좋으나 동음이의어를 분류하지 못함

- 통계 기반 기계번역
  - 방대한 데이터를 이용하여 확률을 기반으로 학습
  - 다른 언어권의 고유한 단어나 관용적 표현은 잘 번역하지 못함

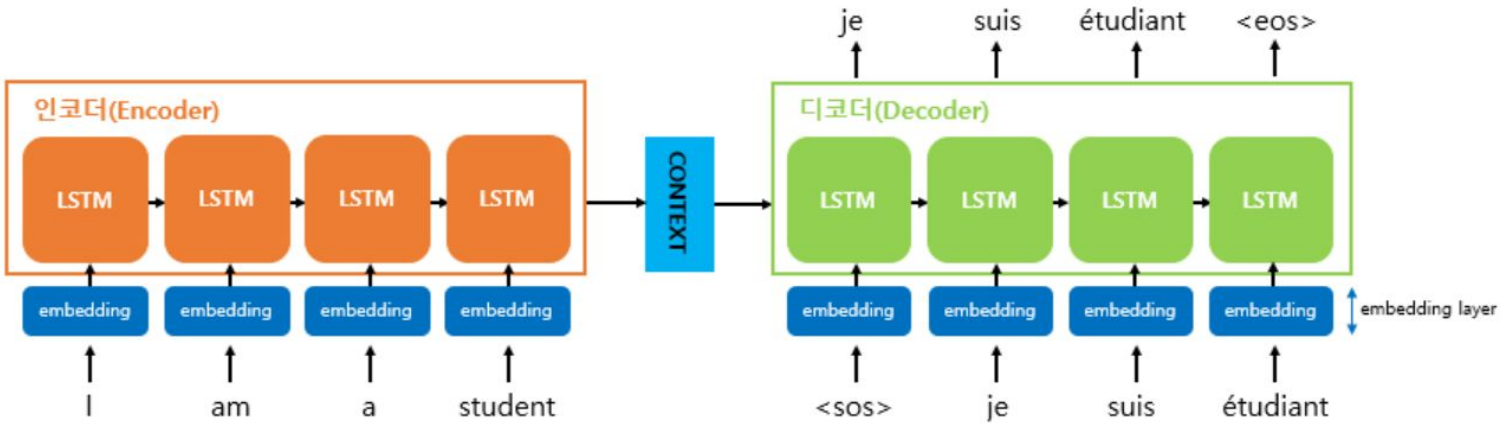
- 신경만 기반 기계번역
  - 신경망을 이용하여 전체 문장을 통째로 인식하면서 의미를 파악하여 학습
  - 주변의 맥락을 이해하여 상황에 맞는 번역이 가능함

How?

Machine Translation

- 입력 시퀀스의 길이는 패딩을 이용하여 맞출 수 있으나 출력 시퀀스는 맞추기 어려움.

  -> RNN 구조를 활용한 seq2seq 구조를 주로 사용함.

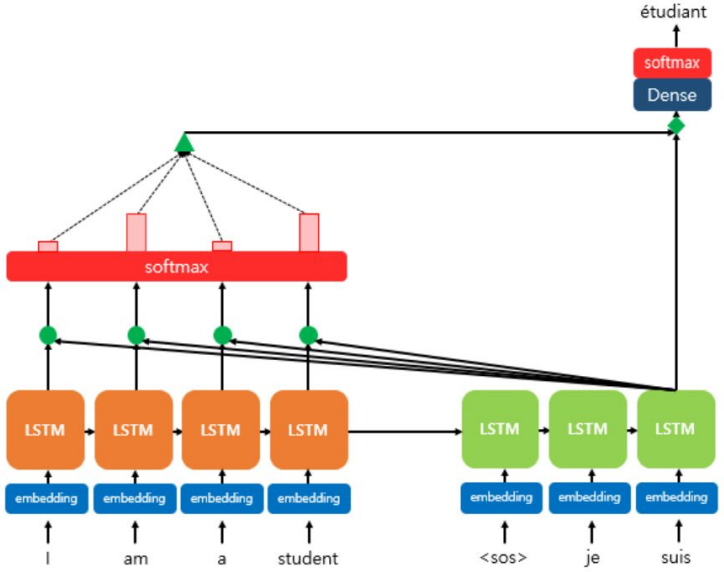


Machine Translation

- RNN을 사용했기에 문장이 길어질수록 정보가 손실되는 문제가 발생

  -> LSTM(long short term memory) 기법 사용 어느정도 해결되었지만 여전히 힘듬

  -> 출력 단계에서 입력 단계의 각 부분을 참고하는 **Attention** 기법을 사용

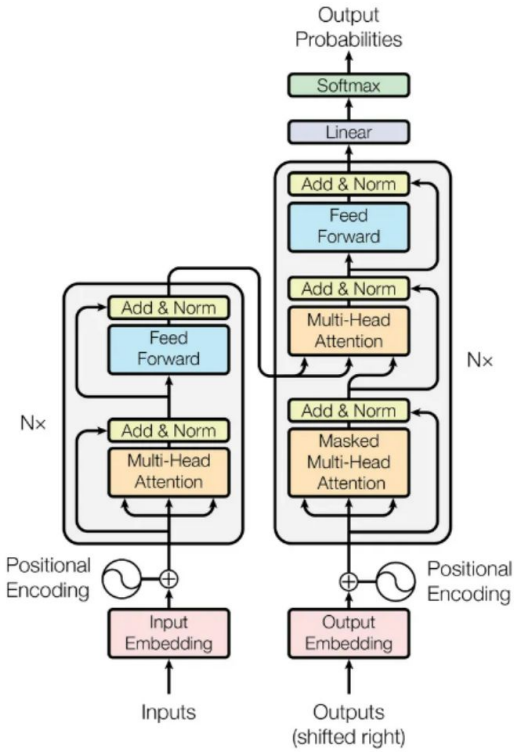


Machine Translation

- 이후 Attention만을 고려하여

  Transformer 모델이 개발됨.

- 이후의 자연어처리 분야에서 혁명을 일으킴
  (BERT & GPT 등)



MT evaluation

번역 평가는 왜 어려운가?

1. 맞다 / 틀리다로 단순하게 판단하기 어려움. => 정답이 굉장히 많음

2. 비슷한 번역 결과는 비슷한 점수를 얻어야 함. => 정형화된 수식을 구하기 어려움

MT evaluation

번역 평가의 기준

1. 정확도
    - 번역의 의미가 얼마나 정확한가?
    - 더해지거나 없어지거나 대체된 부분이 있는가?

2. 유창성
    - 얼마나 자연스러운가?

MT evaluation

번역 평가 방식

1. Human evaluation (사람이 직접 평가)
    - 주관적이다. 평가자에 따라 점수가 바뀔 수 있다.
    - 전문가를 고용해야하기에 비싸고 느림

2. Auto evaluation (컴퓨터로 계산)
    - 항상 동일한 결과값을 보여준다.
    - 빠르고 저렴하다.

MT evaluation

Auto evaluation - Precision and Recall

$$\text{Precision} = \text{Recall} = \text{F1} = 1.0$$

- reordering에 대한 penalty가 전혀 없음.

Evaluation

Auto evaluation - BLEU(Bilingual Evaluation Study)

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

Evaluation

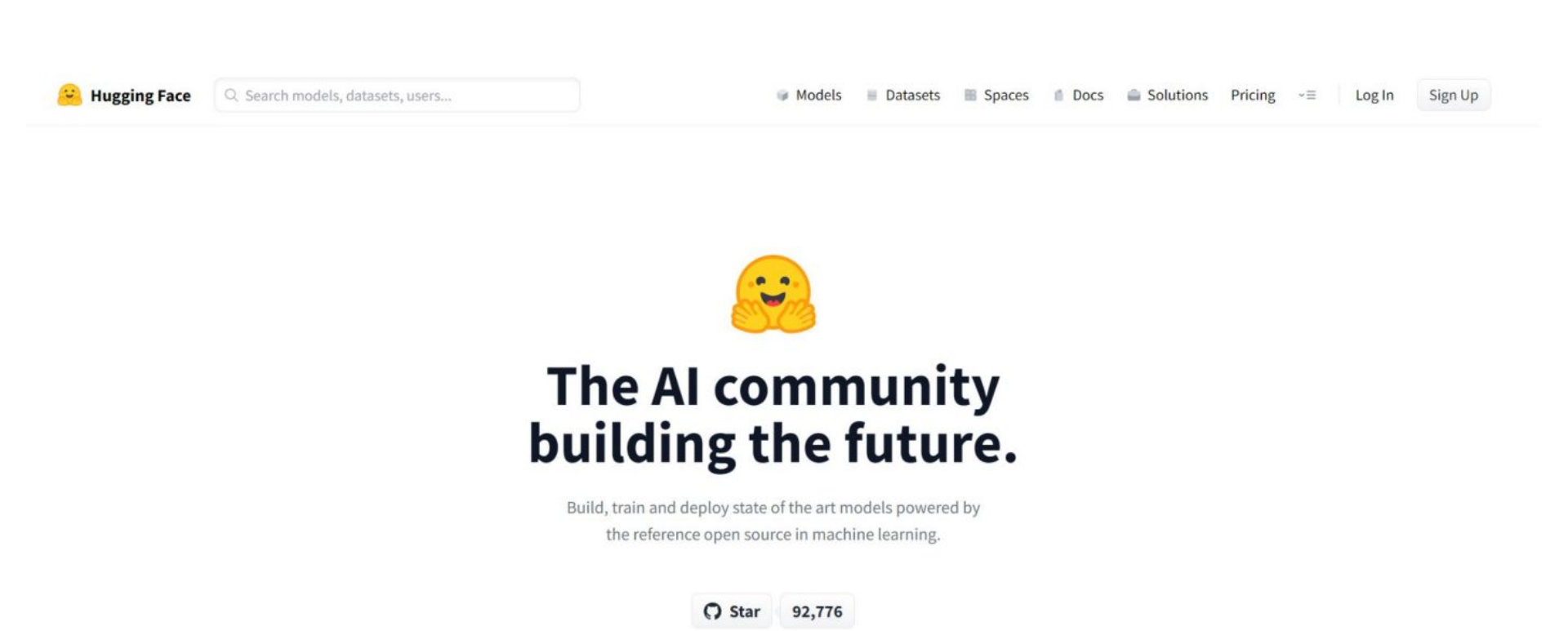
Auto evaluation - WER(Word Error Rate)

Minimum number of **edit**ing steps to transform output to reference

- Match: words match, no cost
- Substitution: replace one word with another
- Insertion: add a word
- Deletion: drop a word

Levenshtein distance, a.k.a. edit distance:

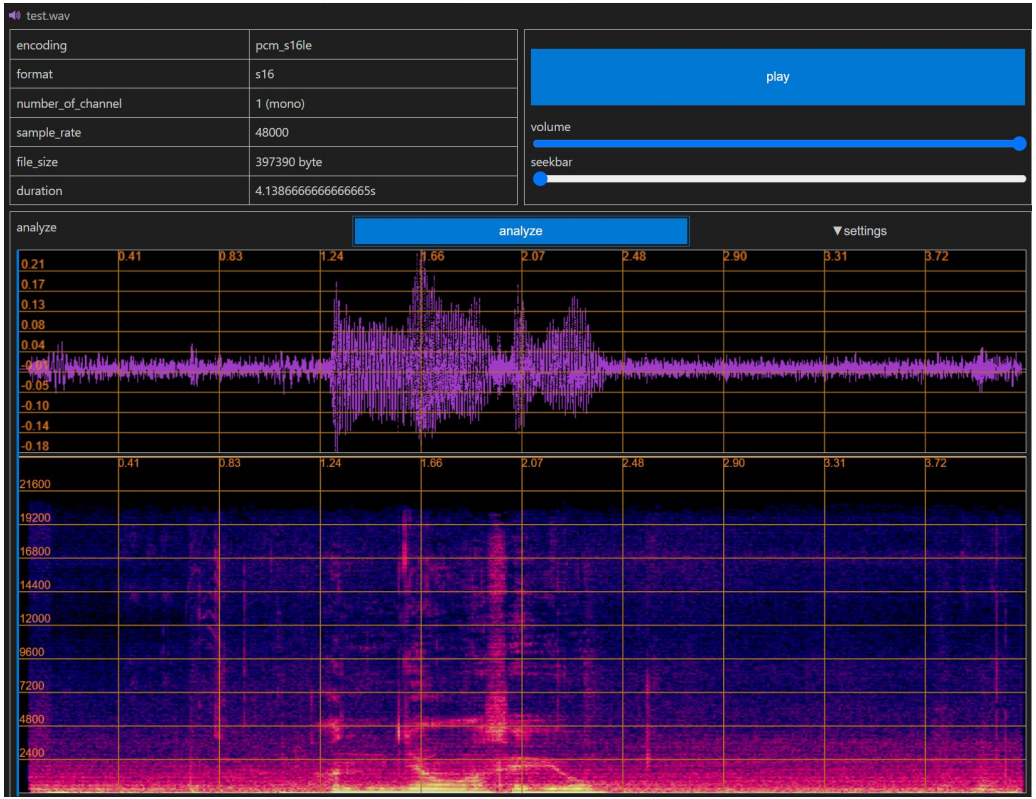
$$\mathrm{WER}\left(\hat{e}_1^{\hat{I}}, \dot{e}_1^{\dot{I}}\right) = \frac{\#\text{substitutions} + \#\text{insertions} + \#\text{deletions}}{\dot{I}}$$



https://wikidocs.net/166832

Speech(Recognition & Synthesis)

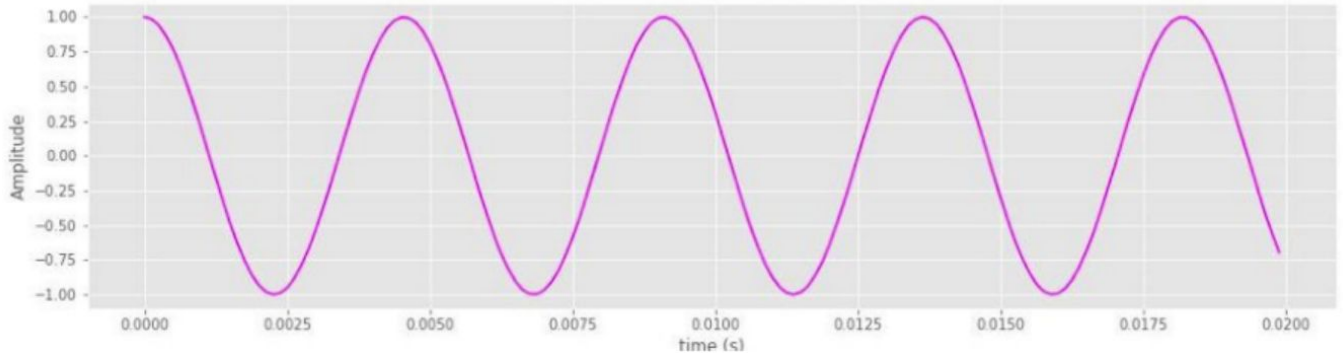
Speech Processing



Speech Processing

음성은 기본적으로 연속된 analog 신호

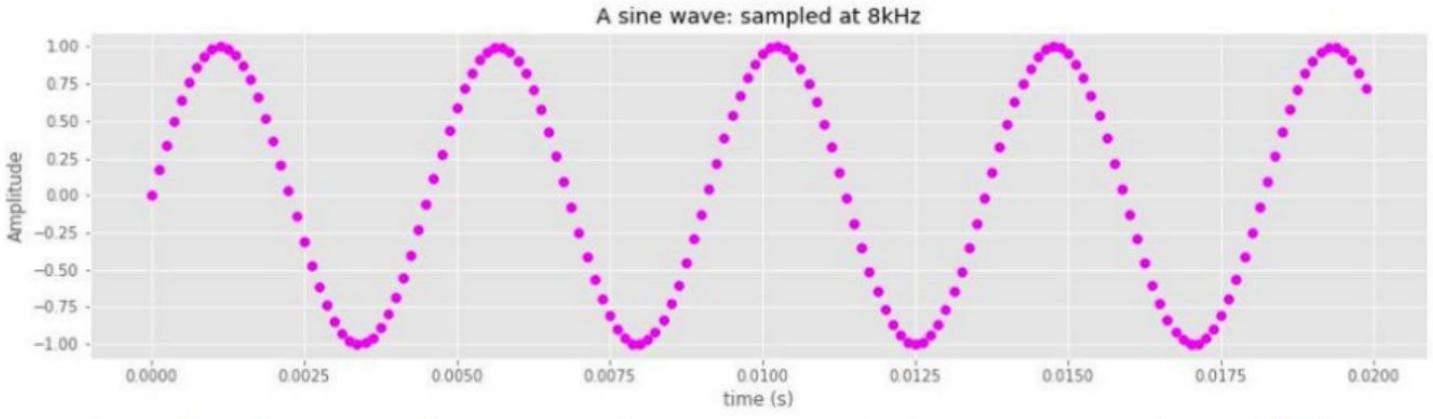
-> 연속적인 신호는 컴퓨터가 인식할 수 없기에 따로 처리해주어야함



Speech Processing

음성은 기본적으로 연속된 analog 신호

-> 연속적인 신호는 컴퓨터가 인식할 수 없기에 따로 처리해주어야함



음성을 sampling rate 수치에 따라 세밀하게 쪼개서 계산에 이용

Speech Processing

Automatic Speech Recognition (STT : Speech to Text)

- 음성을 입력받아 해당하는 언어의 텍스트를 출력하는 기능

Speech Synthesis (TTS : Text to Speech)

- **Text**를 입력받아 학습된 목소리로 음성을 출력하는 기능

Automatic Speech Recognition

음성인식의 목표

- 주어진 음성을 텍스트로 변환하는 것

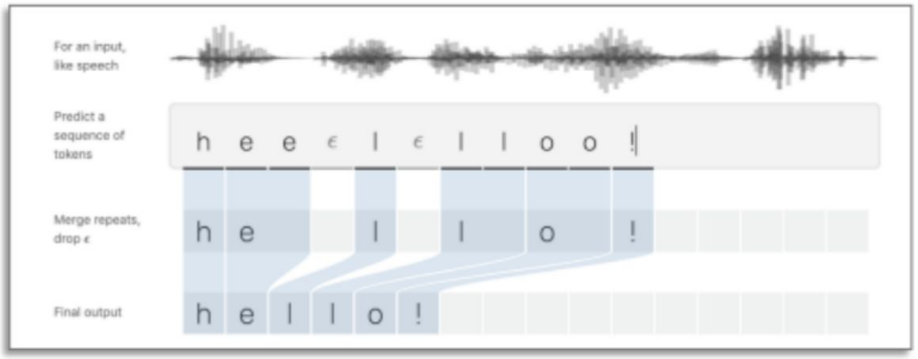
음성인식의 사용 예

1. 휴대폰의 비서(Siri, 빅스비)
2. AI 스피커
3. Youtube 자막생성

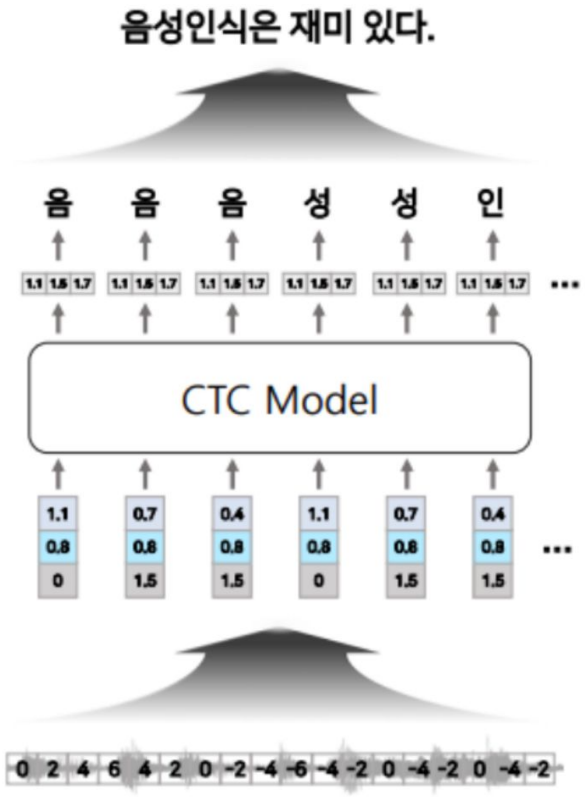
Automatic Speech Recognition

음성인식의 특징

1. 입력과 출력의 길이가 일정하지 않음
2. 각 텍스트에 대응되는 발화의 길이가 항상 다름.
   -> 음성의 길이와 위치에 상관없는 구조가 필요함



CTC 구조



ASR Evaluation

Auto evaluation - WER(Word Error Rate), CER(Character Error Rate)

- 변환된 text와 정답 text를 비교하여 대체, 삭제, 추가된 토큰을 고려한 평가기준
- WER은 단어를 하나의 토큰으로 CER은 문자를 하나의 토큰으로 취급

$$WER = \frac{S + D + I}{N}$$

where...
S = number of substitutions
D = number of deletions
I = number of insertions
N = number of words in the reference

Speech Synthesis

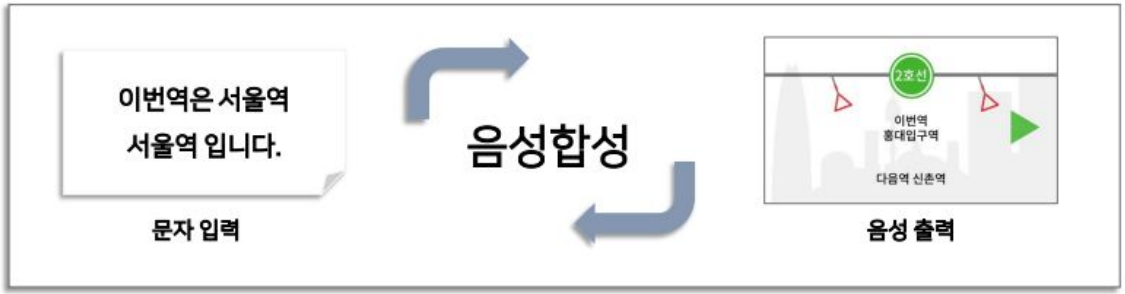
https://youtube.com/shorts/ceJv9vWgOrk?si=JRi2cK71aGAw1NY5

음성합성의 목표

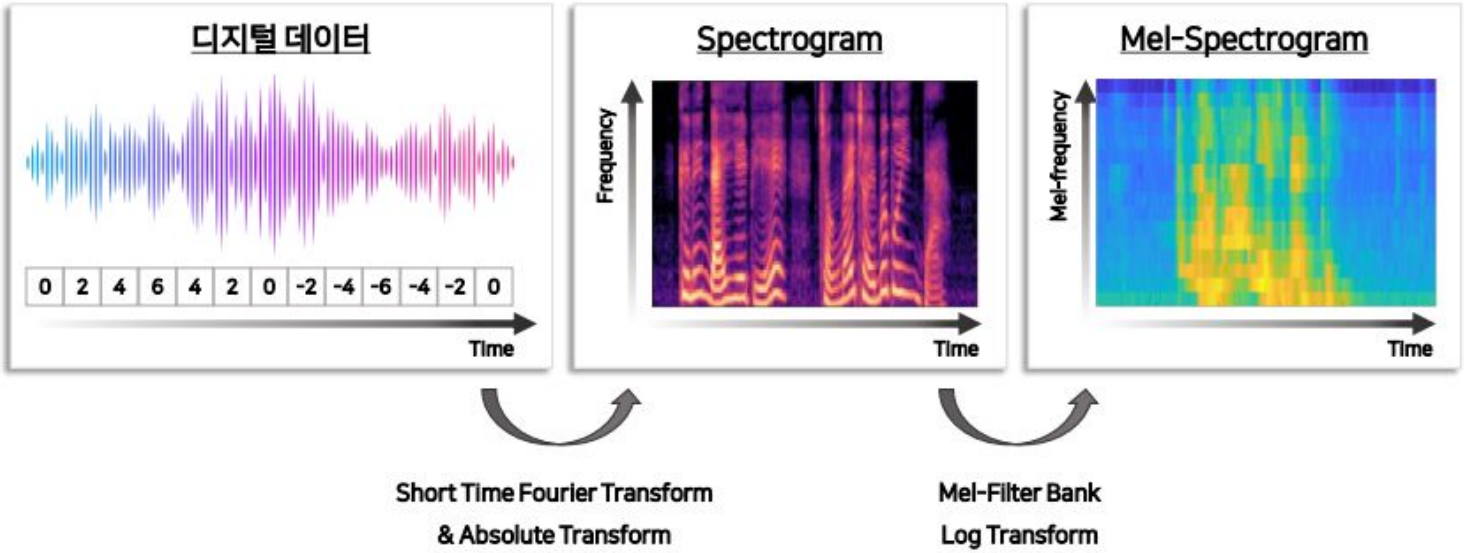
- 주어진 문자를 음성으로 변환하는 것

음성합성의 사용 예

1. 휴대폰의 비서(Siri, 빅스비)
2. AI 스피커
3. 도네이션 (인터넷 방송 / 유튜브 등)
4. 오디오북(유인나, 박명수 오디오북)



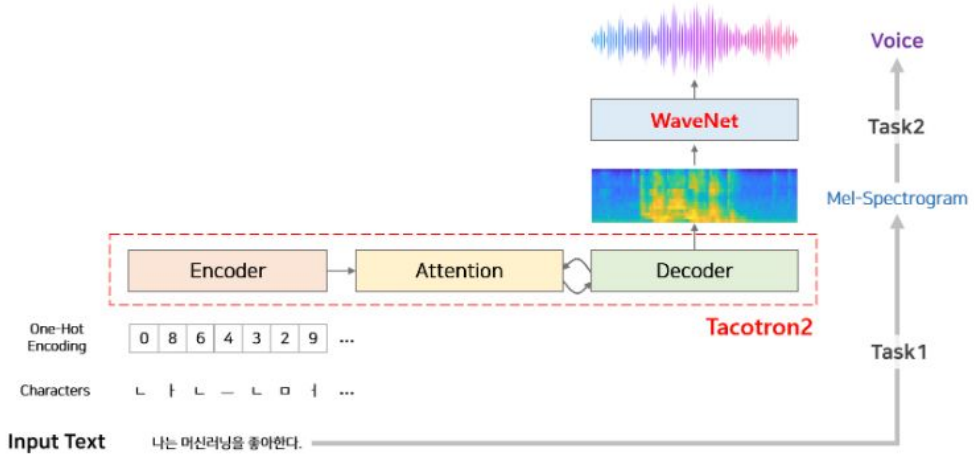
Speech Synthesis (TTS)



Speech Synthesis

음성합성 모델의 종류 (2-stage model)

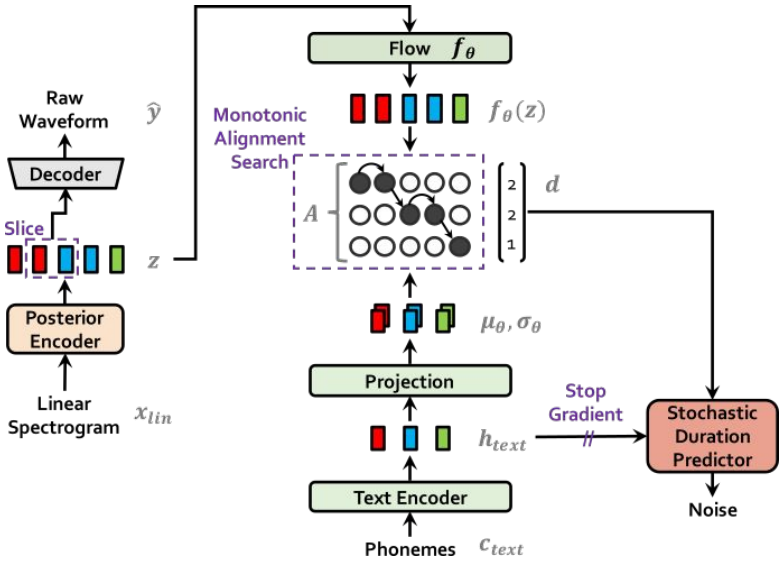
- Text -> model -> Spectrogram -> vocoder -> waveform



Speech Synthesis

음성합성 모델의 종류 (End-to-End Model)

- Text -> model -> waveform



Speech Synthesis Evaluation

Human evaluation - MOS(Mean Opinion Score)

- 합성된 음성에는 정답이 없음 -> 고려해야 할 사항이 너무 많음

  -> 자연스러움(naturalness), 강세(stress), 억양(intonation) 등

- 사람이 직접 음성의 여러가지 측면을 점수로 매기는 MOS가 주로 사용됨

  -> 종합적으로 1 ~ 5점(bad ~ good)으로 점수를 매기는 방식

Speech Synthesis Evaluation

Auto evaluation - CER(Character Error Rate) & MCD(Mel Cepstral Distortion)

- 정확한 음성의 품질을 대변하기에는 어려움이 있어 주로 사용되지는 않음

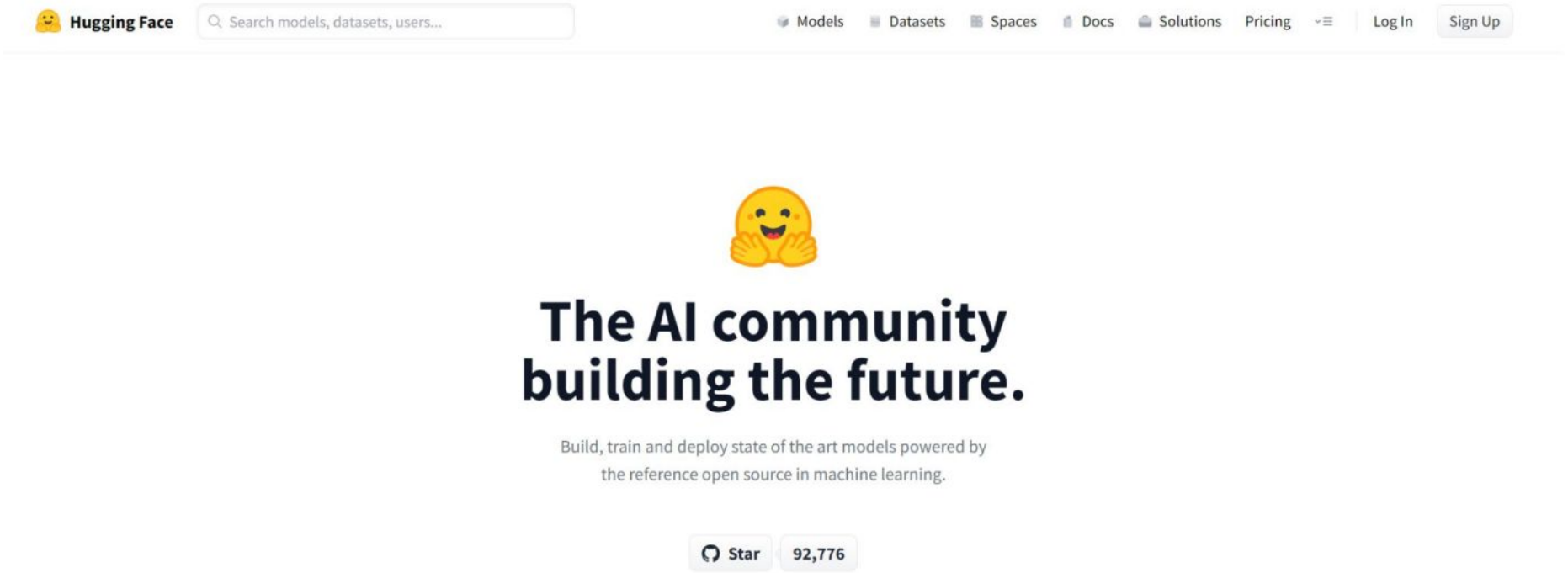
1. CER or WER 평가 : 만들어진 모델로 합성한 음성을 음성인식 모델에 돌려서 정답 텍스트와 비교

-> 사용하는 음성인식 모델에 의존한다는 문제점이 생김

2. MCD(Mel Cepstral Distortion) : 음성 자체의 왜곡을 측정하는 평가방법

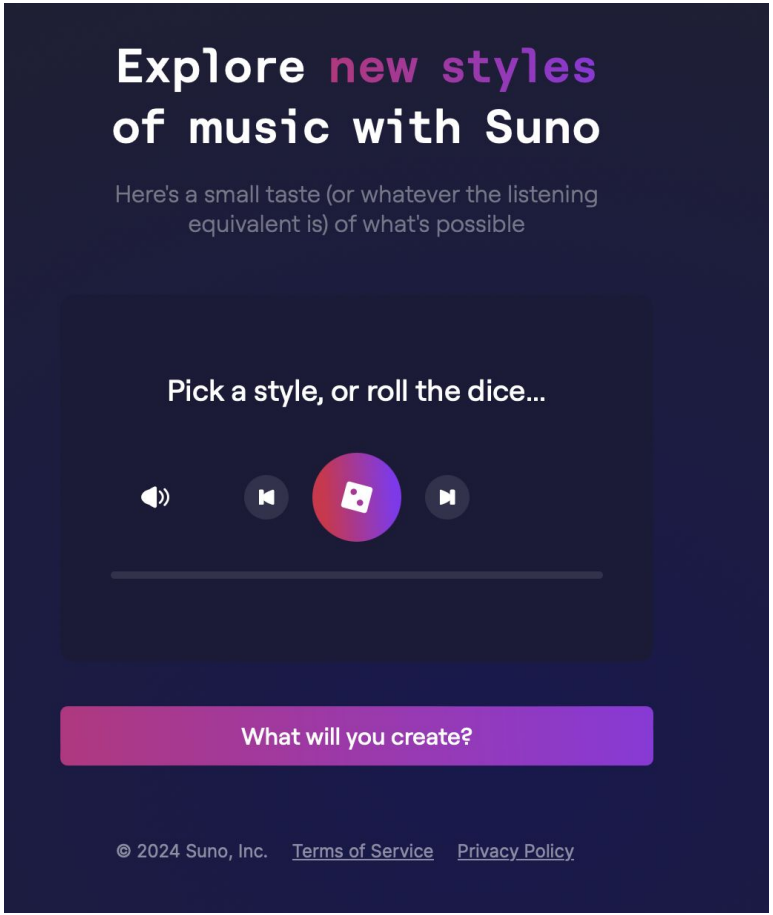
-> 주로 Voice Conversion에 주로 이용됨

https://suno.com/create



와 이게 인공지능 목소리라고?! 진짜 가수와 함께하는 🎤AI 클론싱어 10CM 권정열편

나만의 노래만들기 실습



https://suno.com/create

Tips & Tricks (Chirp v1) | Notion

e.g.

https://suno.com/song/f81ca68a-1509-4f34-afc9-22194d2d7e92

https://suno.com/song/66ebcb51-74a5-4637-91a1-c4c8e49ccc1f