

Capston Festival

# 한국어의 특수성을 반영한 한영 번역 성능 향상

- 지도교수 최희열 교수님
- 팀 허재무, 김정희, 김주환
- 자문 AITRICS

# CONTENTS

01

소개

한영 번역기

한국어의 특성과 번역

02

자모 단위 변환

03

높임말 반말 변환

기본 원리

예외처리

04

Train 결과

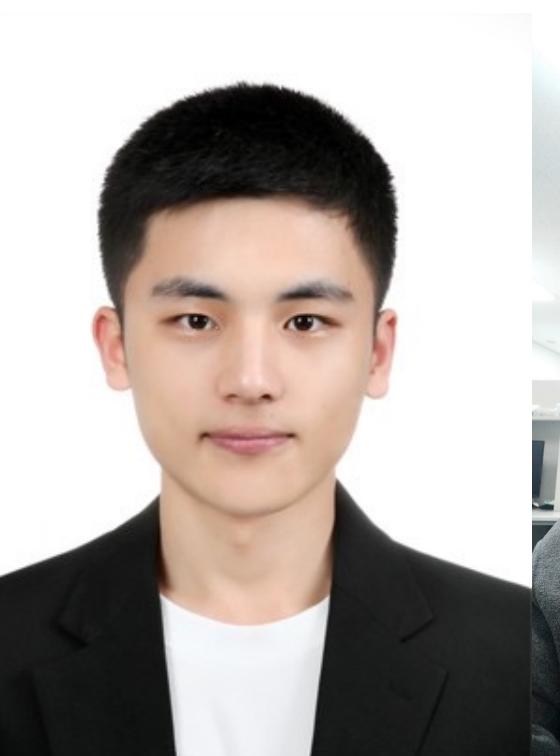
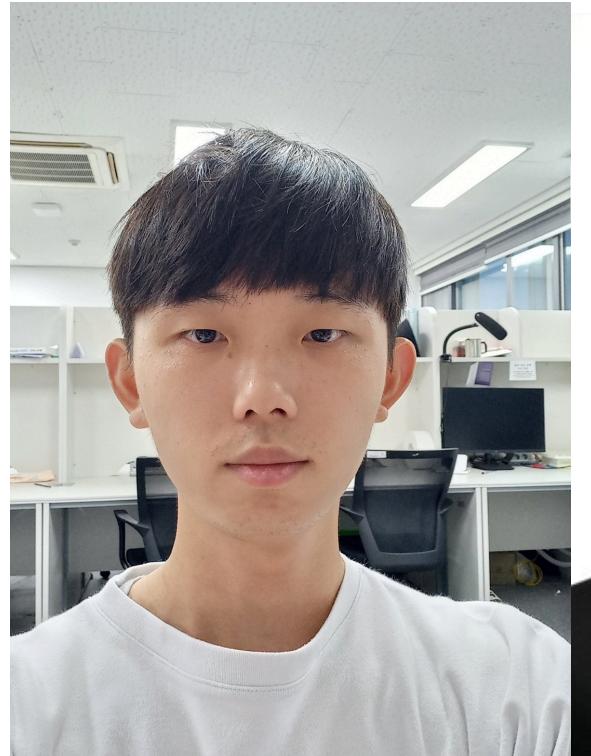
01

## 소개(개요)

한영 번역  
한국어의 특성

# 프로젝트 구성원 소개

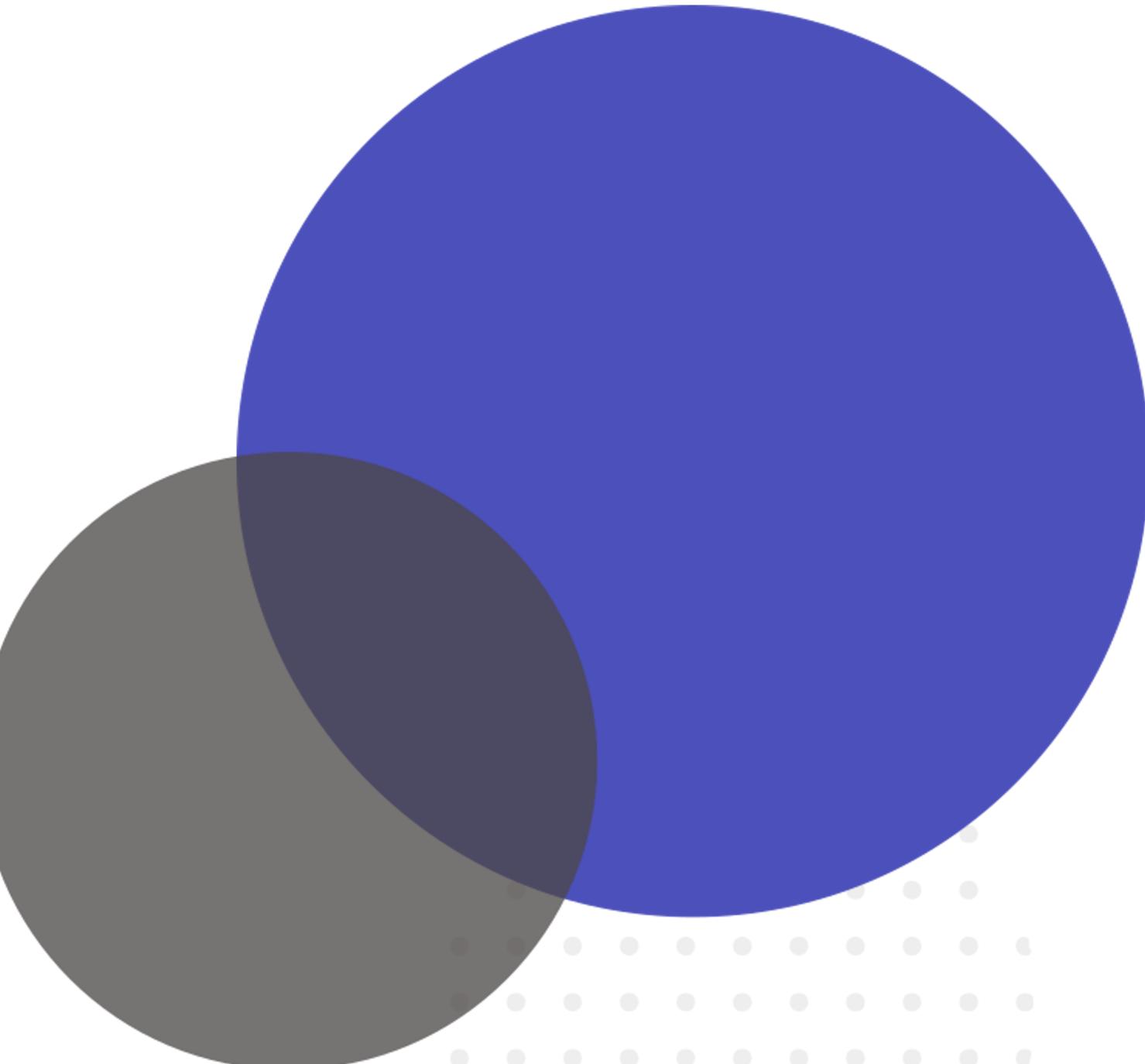
한국어의 특성을 반영한 한영 번역 성능 향상



21700784 허재무

21700156 김정희

21700165 김주환

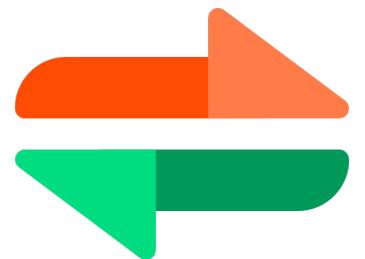


# 프로젝트 소개

## 한국어의 특성을 반영한 한영 번역 성능 향상



한국어



English



# 한 영 번역기



# 문제점 1 자모 단위 결합



한국어 ▾ ↔

영흰 집에 간다.  
yeonghuin jib-e ganda.

영어 ▾ ↔

Young white goes  
home

한국어 ▾ ↔

엄만 아무것도 몰라!  
eomman amugeosdo mollal

영어 ▾ ↔

I don't know anything!



# 문제점 1

## 자모 단위 결합

영환(영희는) ➞ ○ ㅋ ○ ㅎ \_ ㄴ

엄만(엄마는) ➞ ○ ㅏ ○ ㅁ ○ ㅏ ㄴ

한국어의 자음과 모음 단위의 구성

# 문제점 2 높임말, 반말 혼용



I bought a pen  
yesterday. But it  
is broken.

×

나는 어제 펜을 샀다. 하  
지만 고장났습니다.  
naneun eoje pen-eul sassda. hajiman  
gojangnassseubnida.



영어

한국어

I want to eat  
something. How  
about you?

뭔가 먹고 싶어 당신은 어  
떤가요?  
mwonga meoggo sip-eo dangs-in-eun  
eotteongayo?



## 문제점 2 높임말, 반말 혼용

샀다. > 반말 ➤➤➤

샀습니다. > 높임말

싶어? > 반말 ➤➤➤

싶어요? > 높임말

한국어의 높임말, 반말 간 구분

# 문제 해결 방안

## 1. 자모 단위 BPE(Byte Pair Encoding)



# BPE란?

예시

원본 문장의 정보를 압축하기 위한 자연어처리의 전처리 방법

1. 원본 데이터: abcdef

- (a, b, c, d, e, f)의 크기 6인 사전으로 기억 가능

2. 원본 데이터: XabcabcerXazc

- (Xa, bc, a, e, r, z, c)로 분해하여 사전에 저장

등장 빈도수에 따른 단위 생성  
및 원본 문장의 압축 가능

## 자모단위 BPE

모델의 기존 BPE 방식의 subword 단위로 자르는 것이 아닌  
자모단위로 자르는 것.

기대 효과

### subword

- 영흰 -> 영@@흰
- 엄만 -> 엄@@만

### 자모

- 영흰 -> ㅇ ㅎ ㅇ@@ㅎ ㄴ@@ㄴ
- 엄만 -> ㅇ ㅓ ㅁ@@ㅁ ㅏ@@ㅏ



받침 형태의 조사를 분리

# 문제 해결 방안

## 2. 데이터 문장들의 높임말, 반말 통일



## 실행 방안

데이터의 한국어 문장을 높임말, 반말로 각각 통일해 준 다음 train을 진행한다.

## 예시

### 1. 높임말로 통일

문장: 나는 아침마다 출근한다.

-> 저는 아침마다 출근합니다.

### 2. 반말로 통일

문장: 어제 당신과 함께 영화를 봤습니다.

-> 어제 너와 함께 영화를 봤다.



데이터의 한국어 문장을  
**높임말**로 통일,  
**반말**로 통일한 뒤 각각에  
대하여 train을 진행한다.

# BLEU Score

## Bilingual Evaluation Understudy Score

자연어 처리 모델로부터 생성되는 결과물의 정확성에 대한 측도

기대 효과

- 데이터의 통일로 인한 결과 문장의 어체의 일관성.
- 데이터의 통일로 인한 BLEU Score의 향상.

I love you.



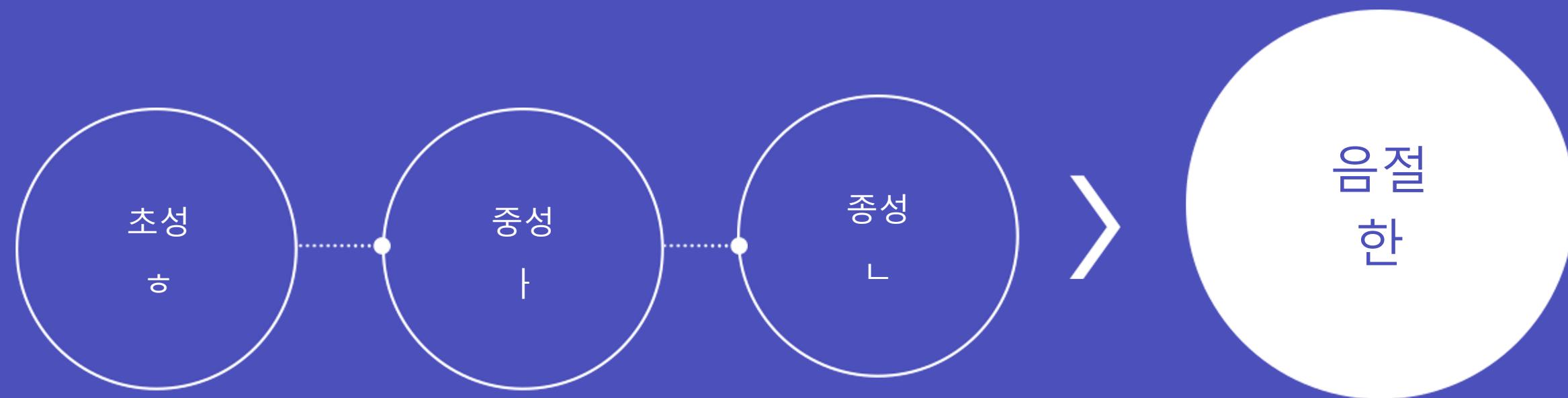
나는 당신을 사랑합니다.  
사랑해요.  
사랑해.

02

## 자모 단위 변환

자모 단위 변환을 통한 모델 학습

# 자모 단위 변환 자모란?



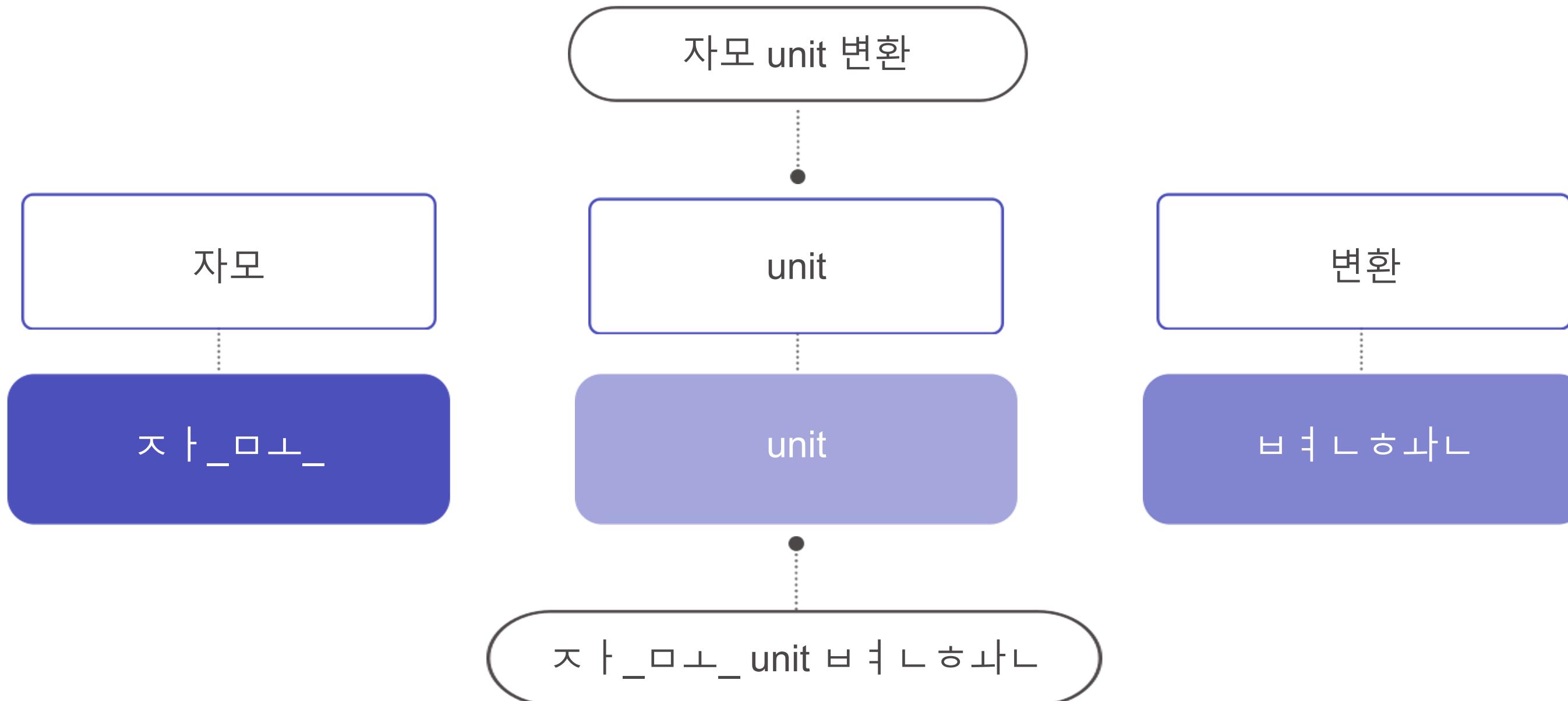
Only Korean  
오직 한국어에만 적용



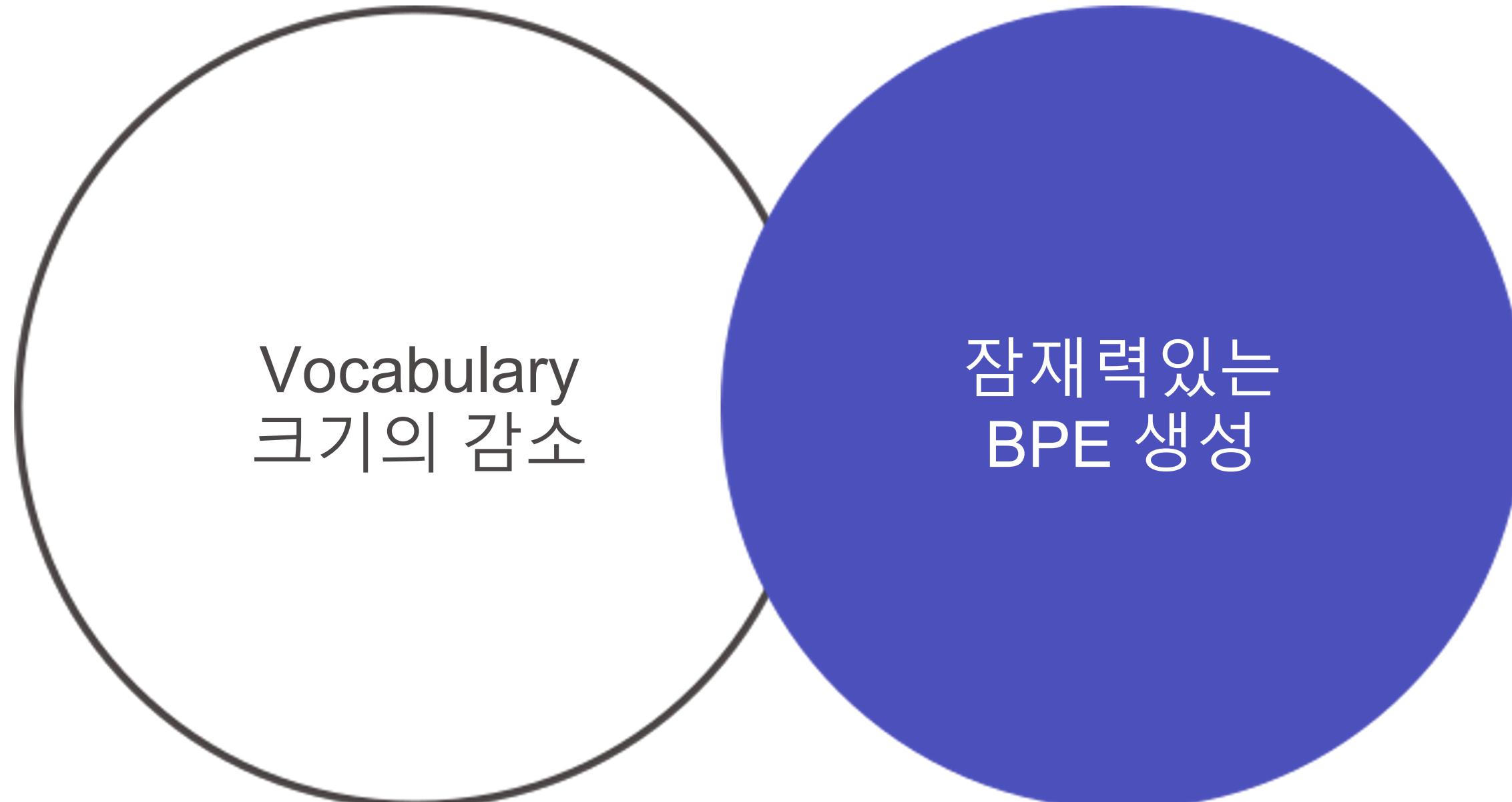
structure  
종성이 없을 경우 '\_'로 대체

# 자모 단위 변환

## 변환 예시



# 자모 단위 변환 장점



# 자모 단위 변환 Vocabulary 크기

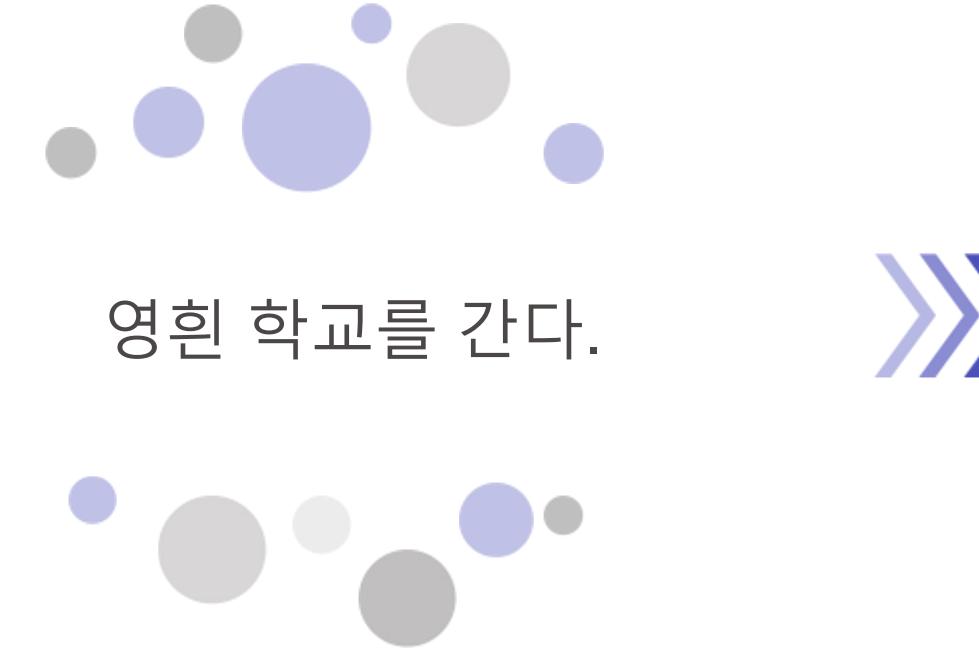
```
INFO:preprocess 16340 unique words in 1596418 sentences with a total of 41063283 words.  
INFO:preprocess:Total: 16340 unique words in 1596418 sentences with a total of 41063283 words.  
INFO:preprocess:Creating dictionary of 12000 most common words, covering 99.9% of the text.  
INFO:preprocess:Saving to /home/nmt21/hgu_data/processed_high/aihub.vocab.kr.tok.sym.10000sub.safe.pkl.
```

**Subword**

**자모**

```
INFO:preprocess 13068 unique words in 1596418 sentences with a total of 42480936 words.  
INFO:preprocess:Total: 13068 unique words in 1596418 sentences with a total of 42480936 words.  
INFO:preprocess:Creating dictionary of 12000 most common words, covering 100.0% of the text.  
INFO:preprocess:Saving to /home/nmt21/hgu_data/processed_high/aihub.vocab.jamo.tok.sym.10000sub.safe.pkl.
```

# 자모 단위 변환 잠재력 있는 BPE



# 자모 변환 BPE 예시

## Subword

- Input : 우리는 영훨 좋아한다.
- BPE(Subword) : 우리는 영@@ 훨 좋아@@ 한다 .  
Output(Subword) : We like Young-Som.
- Input : 교환 지금 축제의 분위기다.
- BPE(Subword) : 교@@ 환 지금 축@@ 제의 분위기@@ 다 .  
Output(Subword) : Kyoto, the atmosphere of the festival is now.

## 자모

- Input : 우리는 영훨 좋아한다.
- BPE(Jamo) : ㅇ ㅜ \_ ㄹ | \_ ㄴ — ㄴ ㅇ ㅋ ㅇ @ @ ㅎ — ㅣ @ @ ㄹ ㅈ ㅗ ㅎ ㅇ ㅏ @ @ ㅎ ㅏ ㄴ ㄷ ㅓ ㅏ \_ .  
Output(Jamo) : We like Younghee.
- Input : 교환 지금 축제의 분위기다.
- BPE(Jamo) : ㄱ ㅍ \_ @ @ ㅎ ㄱ @ @ ㄴ ㅈ | \_ ㄱ — ㅁ ㅊ ㄱ @ @ ㅈ ㅔ \_ ㅇ ㅓ \_ ㅂ ㄴ ㄷ ㅇ ㅌ |\_ ㄱ ㅣ \_ @ @ ㄷ ㅓ \_ .  
Output(Jamo) : The church is in the mood of the festival now.

# 대중적인 번역기와의 비교



구글 번역

- 교회 지금 축제의 분위기다.
- It is a festive mood right now.



파파고 번역

- 교회 지금 축제의 분위기다.
- The church is in a festive mood.



카카오 번역

- 교회 지금 축제의 분위기다.
- It is the atmosphere of the festival now.

## 기존 번역기

- 교회 지금 축제의 분위기다.
- Kyoto, the atmosphere of the festival is now.

## 자모 단위 번역기

- 교회 지금 축제의 분위기다.
- The church is in the mood of the festival now.

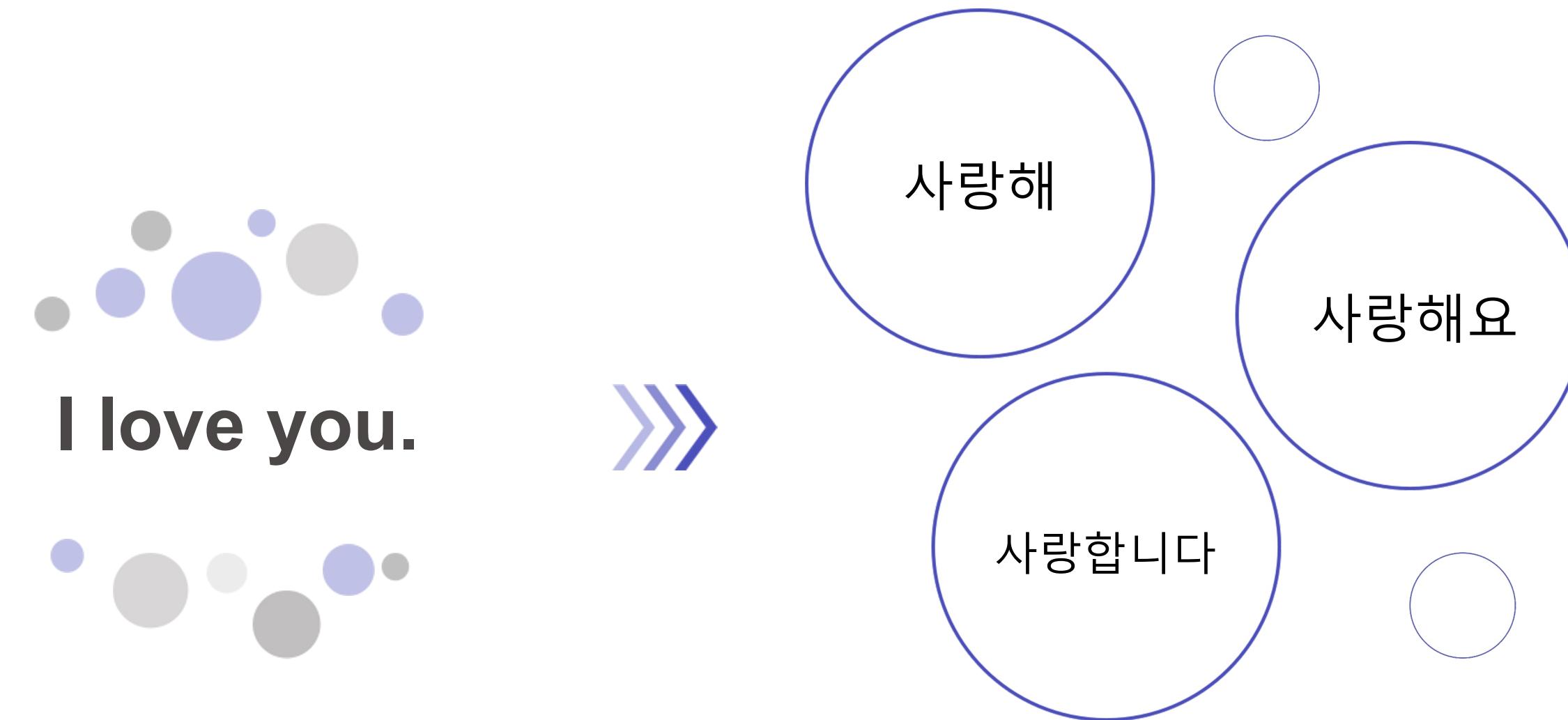
03

## 높임말 반말 변환

한국어 어체 변환 전처리 모듈 개발

# 한국어 표현의 다양성

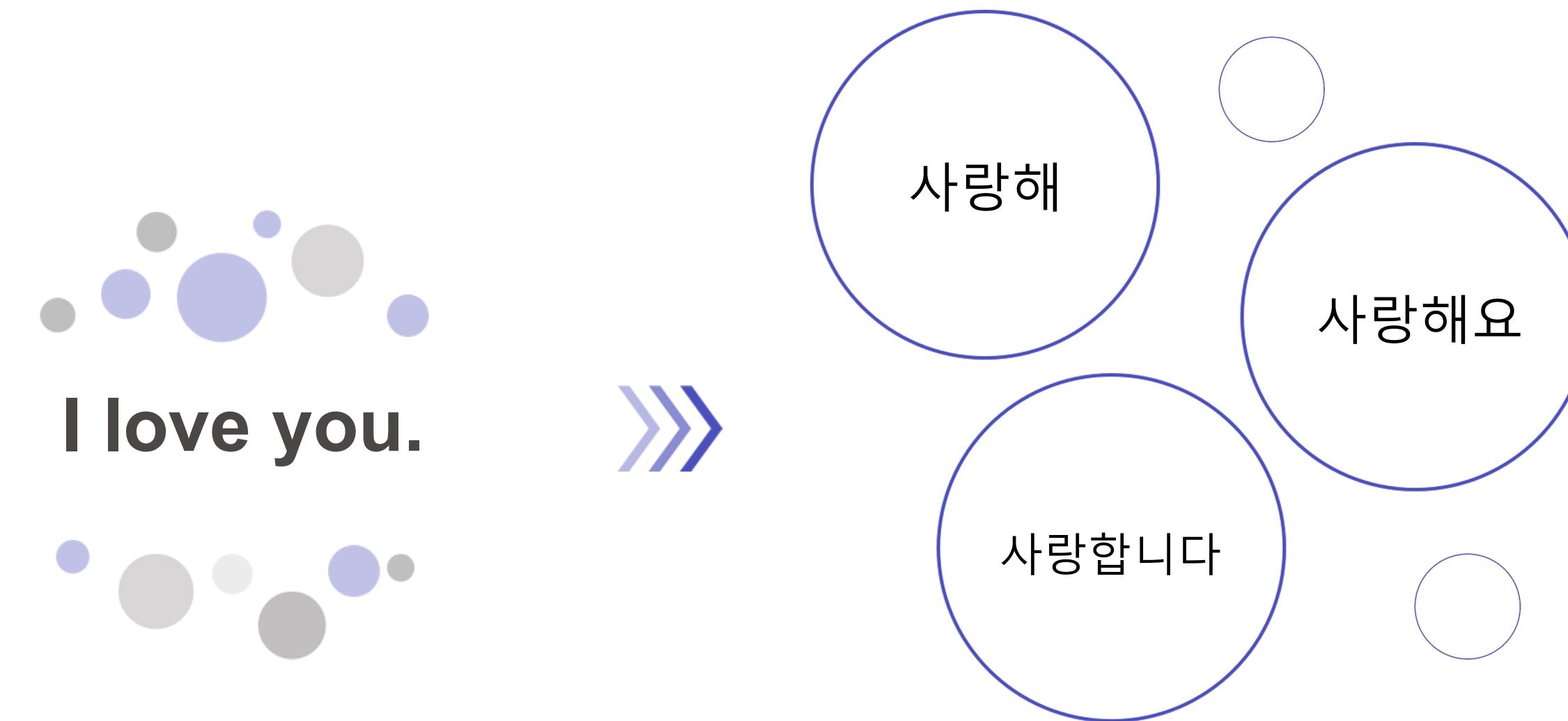
하나의 의미를 여러 문장으로 나타낸다.



무엇이 맞는 표현인가?

# 한국어 표현의 다양성

하나의 의미를 여러 문장으로 나타낸다.



표현의 다양성  > BLEU Score 

# 상대높임법을 반영한 변환 종결어미 기준 어체 간 변환

	격식체				비격식체	
	하십시오	하오	하게	해라	해요	해
평서형	-(ㅂ니)다	-오	-네	-(는/ㄴ)다	-어요	-어
감탄형		-(는)구려	-(는)구먼	-(는)구나	-어요	-어
의문형	-(ㅂ니)까	-오	-(느)ㄴ가	-(느)냐	-어요	-어
명령형	-(ㅂ)시오	-오	-게	-어라/아라	-어요	-어
청유형	-(ㅂ)시다		-세	-자	-어요	-어

하십시오체



해라체

# 상대높임법을 반영한 변환 종결어미 기준

	격식체				비격식체	
	하십시오	하오	하게	해라	해요	해
평서형	-(ㅂ니)다	-오	-네	-(는/ㄴ)다	-어요	-어
감탄형		-(는)구려	-(는)구먼	-(는)구나	-어요	-어
의문형	-(ㅂ니)까	-오	-(느)ㄴ가	-(느)냐	-어요	-어
명령형	-(ㅂ)시오	-오	-게	-어라/아라	-어요	-어
청유형	-(ㅂ)시다		-세	-자	-어요	-어

해요체



해체

# 변환 기본 원리 높임표현으로 변환 예시

Level1.

입력: 나는 책을 읽었다.

Level2.

‘인칭대명사’(NP)  
‘나’ / ‘는’

명사 + 조사  
‘책’ / ‘을’

종결어미(EF) ‘해체’  
‘읽’ / ‘었’ / ‘다.’

Level3.

나 저는  
> 저

책을  
변동없음

읽었  
변동없음

다 습니다  
> 습니다

자모 단위 분리  
후 변환



# 예외 처리

한국어는 한국어 만의 문법적 특성이 있다.

01. 용언의 활용.

03. '-니다' 종결어미

02. 종결어미 '-아/어'

04. 현재형 동사



반말로 변경 시 고려해야 할  
한국어의 문법적 요소

# 어체를 변환할 때의 고려해야 할 문법적 요소



## 1. 용언의 활용

- 용언의 규칙 불규칙 활용을 반영.  
‘-세요’와 같이 용언의 활용이 일어나는 동사,  
형용사는 규칙에 따라 변경됨. ....

당신은 누구와 노세요?



너는 누구와 놀아?

## 2. ‘-아/어’ 종결어미

- 해체로 변경할 때 종결어미 ‘-아/어’로 끝나는  
종결어미는 종성을 보고 변경됨. ....

당신은 누구와 노세요?



너는 누구와 놀아?

# 어체를 변환할 때의 고려해야 할 문법적 요소

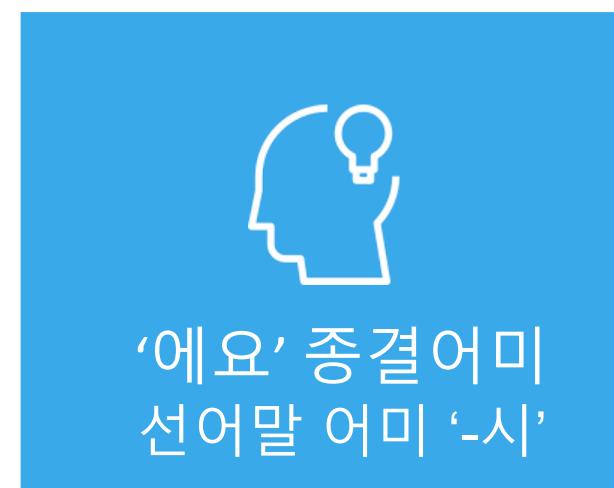


## 3. ‘-니다’로 끝나는 종결어미

- 용언의 종류에 따라 변환 시 ‘-다’로 변경되거나,  
‘다’로 변경

## 4. 현재형 동사 + 종결어미

- 현재형 동사에서는 현재형 종결어미 ‘-는다’로 변경



집에 갑니다.



집에 간다.

밥을 먹습니다.

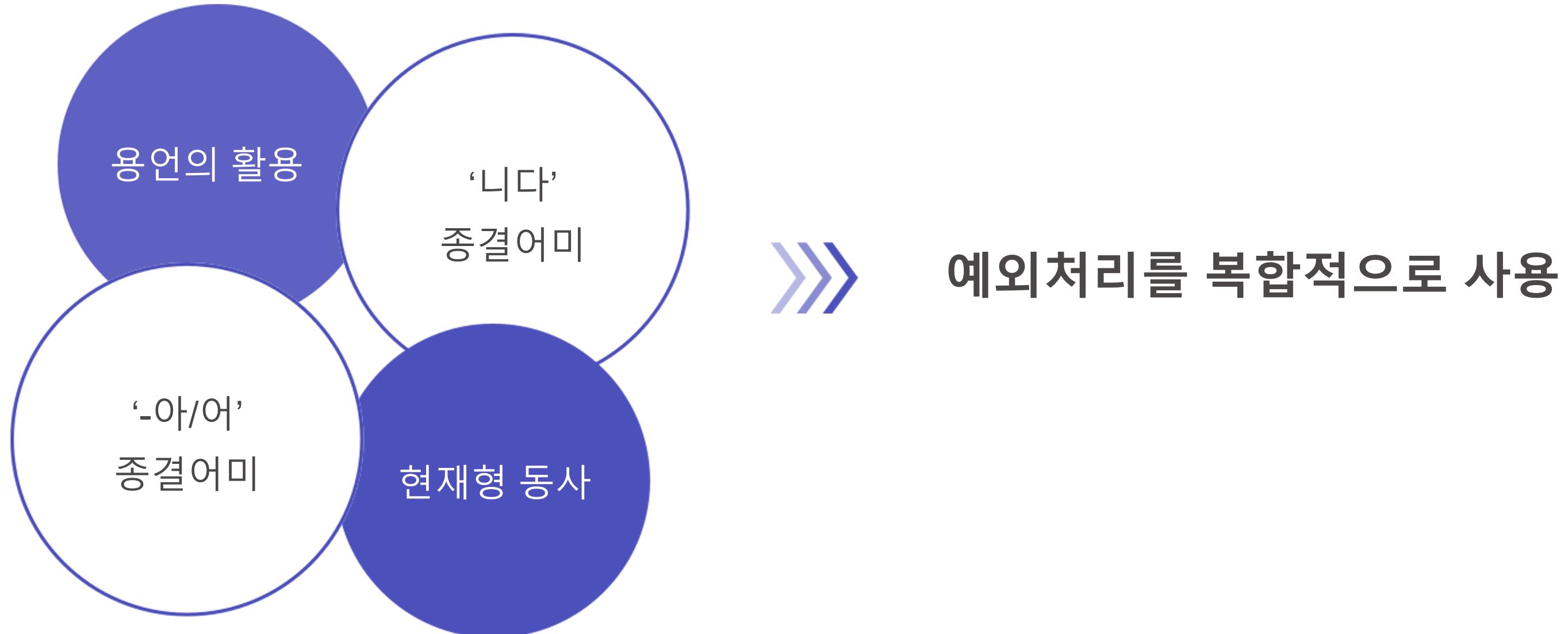


밥을 먹는다.

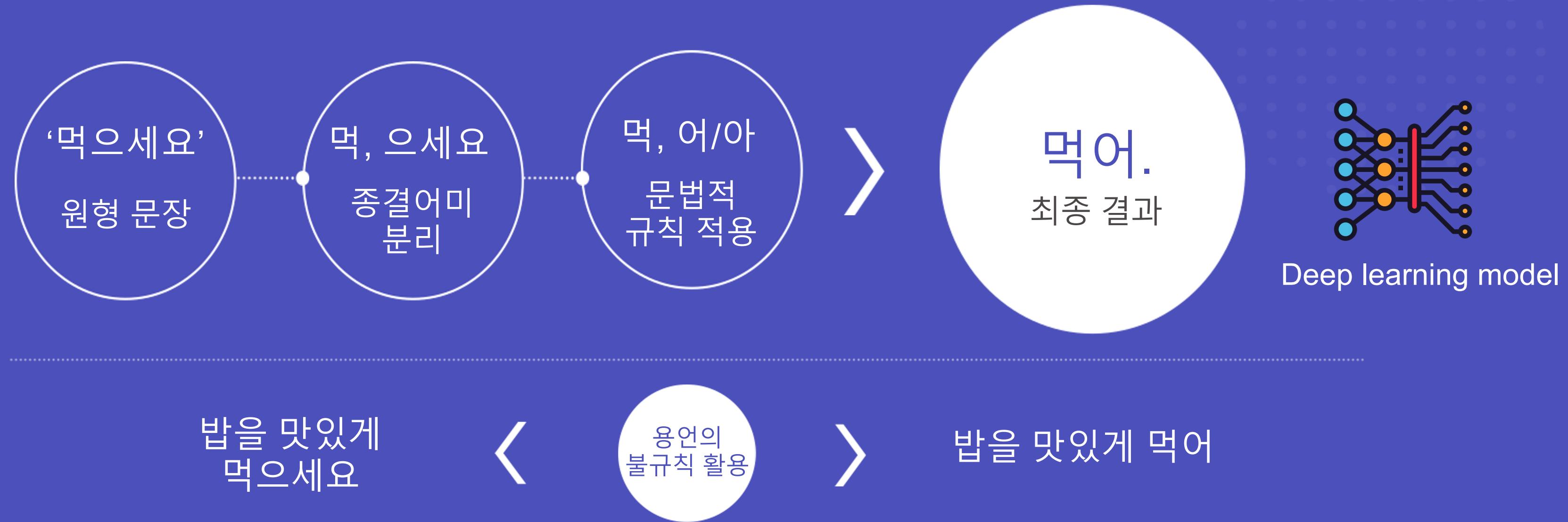
저녁에요? ➞ 저녁에?

누구십니까? ➞ 누구니?

# 변환 모듈의 예외 처리 방식 복합적으로 사용



# 반말로의 변환 과정 ‘먹으세요’ 예제

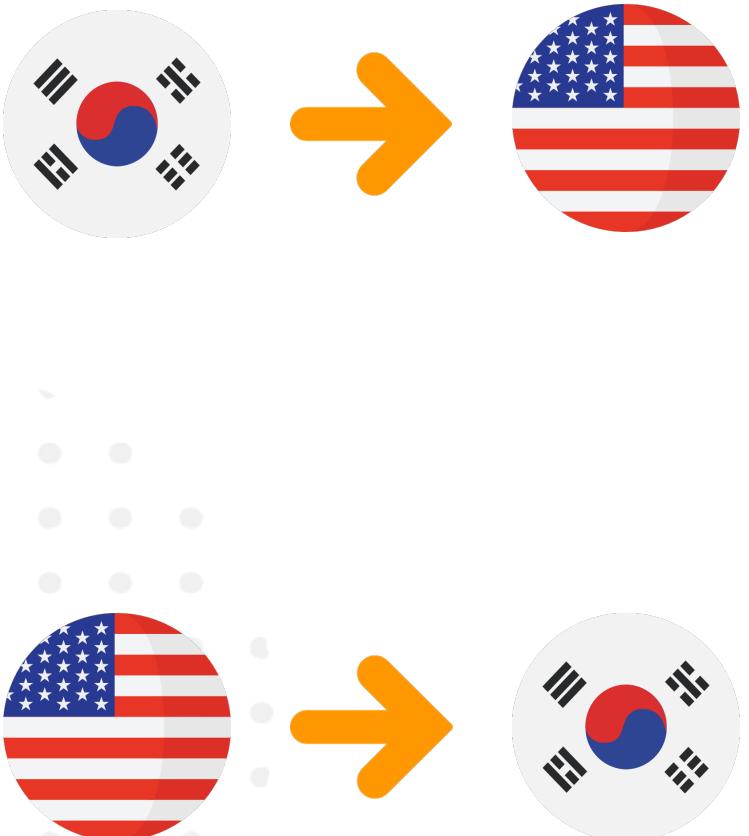


각 종결어미가 문법적 요소를 반영한 규칙에 의해 변경됨

04

Train 결과

# 학습 결과 BLEU Score



데이터 모델명	valid	Aihub_test	Hgu_test
kr2en	39.17	32.58	25.16
high2en	39.16	32.83	25.08
low2en	39.44	33.05	15.71
jamo2en	39.3	32.99	25.71
jamo_h2en	39.6	32.94	26.75
jamo_l2en	39.35	33.05	15.93
en2kr	20.52	13.30	10.62
en2high	20.54	13.27	10.50
en2low	20.77	13.41	11.34
en2jamo	20.7	13.50	10.98
en2jamo_h	21.08	13.66	11.29
en2jamo_l	20.99	13.91	10.91

## 번역기 별 실제 예시 Google Translate, Papago, Kakao i

Input: I bought a pen yesterday. But it is broken.



나는 어제 펜을 구입했습니다. 하지만 그건 부러졌어요.



어제 펜을 샀다. 하지만 고장났습니다.



어제 펜을 샀는데 고장이 났어요.

저는 어제 펜을 샀어요. 하지만 깨져있어요.



Historical, Social, Cultural

언어의 특성을 반영한 번역

2039 미래 기술 세미나

THANK  
YOU

