2022년 봄학기
캡스톤 축제
[캡스톤 디자인]

한국어의 특수성을 반영한 번역성능향상

참여기업체 : AITRICS

지도 교수님 : 최희열 교수님

팀원 : 허재무 김정희 김주환

1 필요성 및 문제 정의

▪ 과제의 필요성

딥러닝 모델을 사용한 한영 번역기가 대중적으로 많이 사용되고 있지만, 한국어의 특성을 반영하지 못해 어색한 결과물을 만들어내는 경우가 존재한다.

▪ 문제 정의

[Problem Statement]

- 한국어의 특성을 반영한 모델 학습의 필요성
  - ① 조사 등이 받침의 형태로 결합된 경우 오역하는 문제 발생
  - ② 여러 영어 문장 입력 시 높임말과 반말의 혼용

[Constraints]

- 음절 단위로 떨어지는 BPE의 생성

  *BPE란: Byte Pair Encoding의 줄임 말로, out of vocabulary 문제를 해결하기 위한 알고리즘으로, 문장을 단어보다 작은 단위의 서브 워드 단위로 분할하여 단어를 구성하는 방법론이다.

- 한국어 높임말, 반말 변환 모듈의 부재

[Objectives]

- 한국어의 특성을 반영한 번역 성능 및 가독성 향상

[Functions]

- 자모단위 BPE 생성을 통한 받침 반영(①번 해결)
- 모듈을 통한 문장의 높임말&반말 변환 (②번 해결)

2 제품 비교 분석

▪ 한 -> 영 번역

번역기	번역 결과	정확성
input	교훤 지금 축제의 분위기다.	-
	The church is now in a festive mood.	✓
	It's a festive mood right now.	✗
	It is the atmosphere of the festival now.	✗
	The church is in the mood of the festival now.	✓

▪ 영 -> 한 번역

번역기	번역 결과	가독성
input	I love you. I will marry you.	-
	사랑해요. 나는 너와 결혼할 것이다.	✗
	사랑해요. 나는 당신과 결혼할 것이다.	✗
	사랑해. 결혼하고 싶어.	✓
	사랑해요. 저는 당신과 결혼할 거예요	✓

3 핵심 내용 요약

▪ 자모 단위 BPE

- 단어의 종성에 결합된 조사 분리 가능

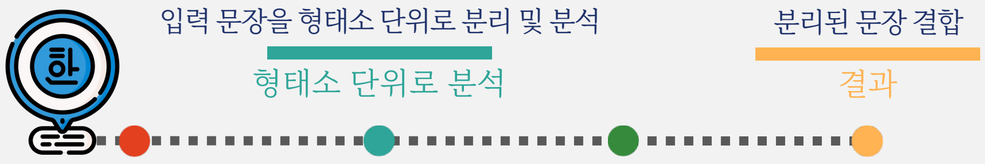
단위	분리결과	BPE 결과	조사 분리
input	교훤	-	-
기존	교, 훤	교@@훤	✗
자모	ㄱ, ㅛ, _, ㅎ, ㅚ, ㄴ	ㄱㅛ_@@ㅎㅚ@@ㄴ	✓

- Vocabulary size의 감소

문장 1596418 » 단어(자모) 13068 단어(기존) 16340

▪ 높임말, 반말의 통일

- 형태소 분석기를 활용해 한국어 어체 변환 모듈 개발



4 실험 결과 / 평가

▪ 모델 성능 비교

- 한국어 -> 영어 번역 모델 성능

단위	어체	BLEU Score		
		Valid	Aihub(Test)	HGU(Test)
음절	-	39.17	32.58	25.16
	높임말	39.16	32.83	25.08
	반말	39.44	**33.05**	15.71
자모	-	39.3	32.99	25.71
	높임말	**39.6**	32.94	**26.75**
	반말	39.35	**33.05**	15.93

- 영어 -> 한국어 번역 모델 성능

단위	어체	BLEU Score		
		Valid	Aihub(Test)	HGU(Test)
음절	-	20.52	13.30	10.62
	높임말	20.54	13.27	10.50
	반말	20.77	13.41	**11.34**
자모	-	20.7	13.50	10.98
	높임말	**21.08**	13.66	11.29
	반말	20.99	**13.91**	10.91

▪ 평가

자모 단위 그리고 어체 변환을 통해 번역 모델의 성능을 향상 할 수 있었고 왼쪽의 예시들 처럼 정성적으로도 번역의 성능이 높아짐을 확인 할 수 있다. 또한 어체 변환 버튼을 추가하여 웹상에서도 추가적인 비용 없이 어체 변환을 진행 할 수 있다.