한국어의 특수성을 반영한 한영 번역

(Kor-Eng NMT using 한국어의 특수성)

김정희† 허재무† 김주환† 최희열§

(Jeonghui Kim) (Jaemu Heo) (Joowhan Kim) (Heeyoul Choi)

**요 약** 딥러닝은 한영 번역의 성능을 향상시켰다. 그렇지만, 한국어의 ‘언어적 특성의 차이’ 를 반영하지 않아 생기는 문제들이 존재한다. 예를 들어, Google 번역은 한국어 문장의 고유명사에 종성 형태로 조사가 결합되어 오역을 하는 경우가 존재하며, 여러 영어 문장이 입력된 경우 존댓말과 반말이 혼용된 일관성 없는 한국어 문장들을 출력한다. 이는 한국어의 ‘자모단위 구성’ 과 ‘존댓말, 반말 간의 구분’이라는 특성이 반영되지 않아 생기는 문제이다. 본 논문에서는 이를 해결하기 위해 corpus 문장들의 존댓말, 반말 통일 및 자모단위 구성을 통한 세밀한 BPE(Byte Pair Encoding)를 통한 모델 학습을 제안한다. 제안된 방법의 성능을 검증하기 위해 기존 corpus와 비교한 결과 BLEU score의 향상을 확인했다.

키워드 : 신경 기계 번역, 자모 단위 번역, 높임말-반말 변환

*Abstract:*

Keywords :

\* 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2018-0-00749,인공지능 기반 가상 네트워크 관리기술 개발)

† 학생회원 : 한동대학교 전산전자공학부

§종신회원 : 한동대학교 전산전자공학부 교수

[heeyoul@gmail.com](mailto:heeyoul@gmail.com) (Corresponding author)

논문접수 : 2020년 08월 일

심사완료 : 년 월 일

CopyrightⒸ2004 한국정보과학회ː개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

1. 서론

최근 딥러닝에 대한 연구 및 개발이 활발히 진행되고 있다. 이에 따라 자연어 처리 분야와 기계 번역도 많은 발전이 이루어졌고, 음성인식, 번역과 같은 방식으로 우리의 삶에 큰 영향을 미치고 있다. 그 중 번역의 경우, 국내에서는 NAVER 파파고, Google 번역 그리고 Kakao 번역이 대중적으로 많이 사용되며, 방대한 양의 데이터와 신경 기계 학습 기반의 번역 모델을 사용해 뛰어난 한영 번역 성능을 보여준다. 하지만 ‘존댓말, 반말 간의 구분’, ‘자모단위의 구성’이라는 한국어의 특징을 반영하지 못하여 한영 번역의 성능과 품질을 낮추는 경우가 존재한다.

먼저 Google 번역의 경우 다음 예시와 같이 ‘존댓말, 반말 간의 구분'을 하지 못한다. Kakao 번역과 NAVER 파파고의 경우 ‘존댓말, 반말 간의 구분’을 적용할 수 있지만 한영 번역을 위한 번역 모델뿐만 아니라 어체 변환을 위한 별도의 모델이 필요하다(이거 확인해봐야됨). 이를 위해 높임말, 반말 학습 데이터 같은 추가적인 비용이 든다.

2022.03.21(월) 기준

입력 문장 :

I love you. I want to marry with you.

Google :

사랑해. 나는 당신과 결혼하고 싶습니다.

또한 다음 예시처럼 명사에 종성 형태로 조사가 결합된 결합형 단어의 경우 제대로 번역하지 못하는 경우가 존재한다.

2022.03.21(월) 기준

입력 문장 : 누난 너무 바쁘다.

NAVER : Noona is too busy.

Kakao : She’s too busy.

Google : Noona is too busy.

본 논문은 이와 같이 상용화된 번역기가 한국어의 특징을 반영하지 못하는 문제를 해결하기 위해 형태소 분석기를 이용한 높임말, 반말 변환과 자모단위 학습을 사용한다. 즉 형태소 분석기를 이용하여 각 단어를 분석하고 종결어미와 그 주변 단어들을 통해 적절한 높임말 또는 반말로 변환을 진행한다. 이를 통해 일관성 있는 번역을 가능하게 하며, 추가적인 번역 모델의 학습없이 적은 비용으로 효율적인 높임말, 반말의 변환이 가능하다. 또한 기존의 문장을 [ 안녕하세요. ㅇㅏㄴㄴㅕㅇㅎㅏ\_ㅅㅔ\_ㅇㅛ\_. ]와 같이 자모단위로 분리를 한 후 BPE를 진행한다. 더 정교한 BPE의 생성이 가능하며, 이를 통해 vocabulary의 크기가 상당히 줄어드는 것을 확인 할 수 있다. 또한 이는 번역성능의 향상으로 이어진다.

또한 본 논문은 Self-Attention 방법을 이용한 Transformer 모델을 사용하며, 학습에 사용된 데이터는 AIHub(www.aihub.co.kr)에서 제공하는 한국어-영어 말충치와 HGU\_clean(이거 설명 필요함)을 사용한다. 번역 모델의 성능 평가는 BLEU(Bilingual Evaluation Understudy) Score를 사용하며 기존 데이터, 높임말 데이터, 반말 데이터, 자모 데이터, 높임말+자모 데이터 그리고 반말+자모 데이터를 비교하며, 그 결과 어떤 데이터들의 경우 기존보다 성능이 향상 되는 것을 보여준다. 또한 BLEU Score로는 판단할 수 없지만 질적으로 향상된 몇가지 예시를 NAVER 파파고, Google 번역, Kakao 번역의 결과와 비교하여 제시한다.