한국어의 특수성을 반영한 한영 번역

(Kor-Eng NMT using Characteristics of Korean)

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, 구글 번역은 한국어 문장의 고유명사에 종성 형태로 조사가 결합하여 오역하는 경우가 존재하며, 여러 영어 문장이 입력되면 존댓말과 반말이 혼용된 일관성 없는 한국어 문장들을 출력한다. 이는 한국어의 ‘자모 단위 구성’과 ‘존댓말, 반말 간의 구분’이라는 특성이 반영되지 않아 생기는 문제이다. 본 논문에서는 이를 해결하기 위해 자모 단위 구성을 통한 잠재력 있는 BPE(Byte Pair Encoding) 및 코퍼스 문장들의 존댓말, 반말 통일과 이를 이용한 모델 학습을 제안한다. 제안된 방법의 성능을 검증하기 위해 기존 코퍼스와 비교한 결과 BLEU 점수의 향상을 확인했다.

키워드 : 신경 기계 번역, 자모 단위 번역, 높임말-반말 변환

*Abstract:* Deep learning has improved the performance of Korean-English translation. However, there are problems caused by not reflecting the 'difference in linguistic characteristics' of Korean. For example, In Google translation, there are cases where postposition is combined with proper nouns in Korean sentences in the form of final consonant and misinterpreted, and when multiple English sentences are entered, it outputs inconsistent Korean sentences that are mixed with honorifics and informal speech. This is a problem caused by not reflecting the characteristics of 'phoneme unit configuration' and 'the distinction between honorifics and informal speech' in Korean. In this paper, Potential Byte Pair Encoding (BPE) through phoneme unit configuration, the unity of honorifics and informal speech in corpus sentences and model learning using them are proposed to solve this problem. To verify the performance of the proposed method, we compared it with existing corpus and confirmed an improvement in the BLEU score.

Keywords : neural machine translation, phoneme unit translation, honorific-to-informal speech translation

\* 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 개발 및 연구가 활발하게 진행되고 있다. 이에 따라 자연어 처리 분야와 기계 번역도 많은 발전을 이루었다. 이는 음성인식, 번역과 같은 방식으로 우리의 삶에 큰 영향을 미치고 있다[1][2][3][4].

그중 번역의 경우, 국내에서는 네이버 파파고, 구글 번역 그리고 카카오 번역이 대중적으로 많이 사용되며, 신경 기계 학습 기반의 번역 모델과 방대한 양의 데이터를 사용해 뛰어난 한영 번역 성능을 보여준다. 하지만 ‘자모 단위 구성’과 ‘존댓말, 반말 간의 구분’이라는 한국어의 특성을 반영하지 못하여 한영 번역의 성능을 낮추는 경우가 존재한다.

먼저 ‘자모 단위 구성’을 반영하지 못한 경우 아래의 예시처럼 명사에 종성 형태로 조사가 결합한 단어는 제대로 번역하지 못한다.

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.

또한 구글 번역의 경우 아래의 예시와 같이 ‘존댓말, 반말 간의 구분’을 하지 못한다.

2022.03.21(월) 기준

입력 문장 : I love you. I want to marry with you.

Google : 사랑해. 나는 당신과 결혼하고 싶습니다.

카카오 번역의 경우 ‘존댓말, 반말 간의 구분’을 적용할 수 있지만 한영 번역을 위한 번역 모델 뿐만 아니라 어체 변환을 위한 별도의 모델이 필요하다[5]. 이에 따라 높임말, 반말 학습 데이터와 같은 추가적인 비용이 든다.

본 논문은 이처럼 상용화된 번역기가 한국어의 특성을 반영하지 못하는 문제를 해결하기 위해 자모 단위 학습과 형태소 분석기를 이용한 높임말, 반말의 변환을 제안한다. 첫째, 기존의 문장을 [좋다. ㅈㅗㅎㄷㅏ\_.]와 같이 자모 단위로 분리한 후 BPE(Byte Pair Encoding)[6]를 진행한다. 이에 따라 더 잠재력 있는 BPE의 생성이 가능하며, Vocabulary 크기도 상당히 줄어든다[7]. 둘째, 형태소 분석기를 이용하여 각 어휘를 분석하고 종결 어미와 그 주변 어휘들을 통해 적절한 높임말 또는 반말로의 변환을 가능하게 하며, 추가적인 어체 변환을 위한 별도의 모델 없이 적은 비용으로 효율적인 높임말, 반말의 변환을 가능하게 한다.

본 논문은 Self-Attention 방법을 이용한 Transformer 모델을 사용하며, AI 학습에 필요한 다양한 데이터를 제공하는 AIHub(www.aihub.or.kr)에서 가져온 한국어-영어 번역(병렬) 말뭉치와 HGU 데이터를 사용한다. 번역 모델의 성능 평가는 BLEU(Bilingual Evaluation Understudy)[8] 점수를 사용하며 표 4에 나오는 총 6개의 서로 다른 학습된 번역 모델을 비교한다. 그 결과 특정 번역 모델의 경우 기존보다 성능이 향상되는 것을 확인했으며, BLEU 점수로는 판단할 수 없는 정성적으로 향상된 몇 가지 예시를 네이버 파파고, 구글 번역, 카카오 번역의 결과와 비교하여 제시한다.

1. 배경 지식
   1. Transformer

입력 문장의 문맥을 출력 문장에 반영하는 것으로 자연어 번역의 성능을 향상할 수 있으며 문장의 문맥은 문장을 구성하는 단어 간의 관계를 이용하여 파악할 수 있다. 이에 따라 초기 RNN(Recurrent Neural Networks)을 활용한 자연어 번역 모델에 LSTM(Long Short Term Memory)을 추가한 모델도 등장하였다[1]. 하지만 이런 시도에도 불구하고 RNN은 문장의 단어들이 순서대로 입력되기 때문에 문장 내 거리가 먼 단어 간의 관계를 모델링하는 것이 힘들며, 문장이 길어질수록 학습하는 데 걸리는 시간이 길어진다는 한계가 있다[1]. 이를 개선한 것이 Transformer 모델이며, 이는 단어 간의 관계를 모델링 하는 Attention Mechanism만을 사용한다. Transformer 모델은 이전의 모델들 보다 더 좋은 번역 성능을 보여주며, 현재 번역 모델의 대부분을 차지한다[4].

본 논문의 Transformer 모델의 특성은 다음과 같다. 첫째, 문장 전체가 한 번에 입력되므로 병렬 계산이 가능하다. 따라서 RNN에 비해 학습에 소요되는 시간이 줄어든다. 둘째, Attention Mechanism만을 사용하여 구성했기 때문에 문장 내 단어 간의 거리에 따른 관계를 모델링 하는 것이 효과적이다. 해외의 BERT, GPT 그리고 국내의 카카오 번역과 네이버의 Hyper Clova 등이 Transformer를 사용 중이다[9][10][11].

* 1. BPE(Byte Pair Encoding)

NMT(Neural Machine Translation)는 Vocabulary 크기에 따라 한정된 수의 단어만을 사용하기 때문에 Vocabulary에 존재하지 않는 단어들이나 오타가 입력되는 경우 UNK(Unknown Token)로 설정되어 그 의미를 파악할 수 없게 된다. 이러한 상황을 OOV(out of Vocabulary)라고 하며, 이를 완화하기 위해 문장의 단어들을 더 작은 서브 워드 단위로 나누어 사용하는 서브 워드 Tokenization이 나오게 되었다[12]. 본 논문에서는 BPE라는 대표적인 서브 워드 Tokenization 알고리즘을 사용한다. BPE는 빈도를 기준으로 단어를 재구성하며 다음과 같은 절차를 가진다.

1. 문장들을 어절 단위로 분할
2. 어절의 끝에 띄어쓰기임을 알려주는 </w>추가
3. 어절들을 문자(character) 단위로 나누어 사전 생성
4. 가장 많이 등장하는 문자 쌍을 찾아 병합하고, 병합한 쌍을 사전에 추가하는 행위를 하이퍼 파라미터로 설정한 횟수만큼 반복
5. 사전을 사용해서 Vocabulary 생성

3번과 4번을 실행한 결과는 아래의 예시와 같이 가장 빈도가 높은 (‘대’, ‘학’)을 (‘대학’)으로 병합하게 된다.

3번 실행

Dictionary{

한 동 졸 업 생 </w> : 5

한 동 학 부 생 </w> : 2

한 국 **대 학** 교 </w> : 6

일 본 **대 학** 교 </w> : 3

}

4번 실행

Pair statistic {('한', '동'): 7, ('동', '졸'): 5, ('졸', '업'): 5, ('업', '생'): 5, ('생', '</w>'): 7, ('동', '학'): 2, ('학', '부'): 2, ('부', '생'): 2, ('한', '국'): 6, ('국', '대'): 6, **('대', '학'): 9**, ('학', '교'): 9, ('교', '</w>'): 9, ('일', '본'): 3, ('본', '대'): 3}

Updated Dictionary{

한 동 졸 업 생 </w> : 5

한 동 학 부 생 </w> : 2

한 국 **대학** 교 </w> : 6

일 본 **대학** 교 </w> : 3

}

위의 예시에 하이퍼 파라미터를 10으로 설정하여 BPE를 진행하고 입력으로 ‘한동대학교’를 넣은 결과는 아래의 예시와 같다.

Input : 한동대학교

한 동 대학 교 </w>

한 동 대학교 </w>

한 동 대학교</w>

한동 대학교</w>

이처럼 ‘한동대학교’는 ‘한동’과 ‘대학교’의 두 서브 워드로 분할된다. 따라서 BPE를 사용하면 사전에 등장하지 않는 단어가 입력되더라도 서브 워드의 조합으로 설명이 가능해진다.

* 1. 한국어의 특성

한국어와 영어는 다른 문화권의 언어이기에 서로 다른 언어적 특성들이 존재한다. 그중 본 논문에서 사용하는 한국어에서 나타나는 특징들은 다음과 같다. 첫째, 한국어는 각 음절이 초성, 중성 그리고 종성으로 이루어진 자모 단위로 구성되어 있다. 둘째, 한국어는 어미변화 등의 문법적인 존대법체계를 가지고 있지만 영어는 그렇지 않다[13].

* + 1. 자모 단위 구성

한국어는 자음과 모음을 나타내는 글자가 따로 존재하는 음소 문자이며, 자음과 모음을 합쳐서 음절 단위로 사용하는 음절 문자적 성격을 지니고 있다[14]. 일반적으로 한국어는 음절 단위로 사용되며, 음절을 자음과 모음, 즉 음소로 나눌 수 있다. 한국어는 19개의 자음 23개의 모음으로 구성되어 있다. 음절과 음소에 대한 예시는 아래와 같다.

단어: 한글 → 음절: 한,글 → 음소: ㅎ,ㅏ,ㄴ,ㄱ,ㅡ,ㄹ

* + 1. 존칭 표현(높임법)

한국어에는 상대를 높여서 말을 하는 높임법이 존재한다. 높임법은 주체 높임법과 상대 높임법으로 나눌 수 있다. 주체 높임법은 문장에 등장하는 주체를 높이는 방법이고 상대 높임법은 청자를 높이는 방법이다[15]. 이러한 높임 표현은 종결어미, 인칭 대명사, 감탄사 등 형태소 단위에서 나타나게 된다.

* + - 1. 형태소

한국어는 의미를 가진 최소 단위인 형태소로 이루어져 있으며 자립성의 유무에 따라 의존 형태소와 자립 형태소로, 실질 의미 유무에 따라 실질 형태소와 형식 형태소로 구분할 수 있다[16]. 여기서 의존 형태소는 그 자체로 자립성이 없고 다른 말에 의존해야 하는 형태소로, 주로 조사나, 어미, 접사, 용언의 어근 등이 해당하며 자립 형태소는 그 자체로 자립성이 있는 말로, 주로 체언의 어근, 감탄사, 관형사, 부사 등이 해당한다[17]. 문장을 형태소 단위로 나눈 예시는 아래와 같다.

문장: 재무가 이야기 책을 읽었다.

어절: 재무가/이야기/책을/읽었다.

형태소: 재무/가/이야기/책/을/읽/었/다.

자립 형태소: 재무가 이야기 책을 읽었다.

의존 형태소: 재무가 이야기 책을 읽 었 다.

* + - 1. 상대 높임법

한국어는 청자에 대한 화자의 태도에 따라 여러 문체가 존재하며 문장의 종결 유형별로 문체를 구분한 문체 법이 존재한다. 문장의 목적에 따라 평서법, 의문법, 명령법, 청유법, 감탄법으로 나뉘며 청자 또는 높인 정도에 따라 ‘하십시오’, ‘하오’, ‘하게’, ‘해라’, ‘해요’, ‘해’ 체로 나뉘며 그중 ‘해요’ 체와 ‘해’ 체의 경우 비격식체, 나머지는 격식체로 분류된다[17]. 또한, 이때 ‘저’와 ‘나’, ‘당신’과 ‘너’ 등의 인칭대명사 등을 고려하여 화자와 청자 간 관계를 나타낼 수 있다[16].

1. 제안 모델

본 논문은 한국어의 특수성을 반영한 한영 번역 성능의 향상을 위해 자모 단위 번역과 높임말, 반말의 변환을 제안한다. 번역 모델은 Transformer를 사용했으며, AIHub에서 가져온 한국어-영어 번역(병렬) 말뭉치 160만 개와 HGU 데이터 190만 개를 사용했다.

학습을 위해 데이터를 가공하는 전처리 과정을 거치게 되며, 이는 크게 3가지(Split, Tokenize, BPE)로 나뉜다. 먼저 데이터를 Train, Valid, Test 데이터로 나눠주는 Split을 진행한다. 이후 ‘ . ’, ‘ ! ’, ‘ ? ’ 등의 문장 부호를 기준으로 문장을 띄워주는 Tokenize를 진행한다. 마지막으로 BPE를 생성 및 적용하게 된다. 자모 단위 변환은 Tokenize 이후에 진행되며 높임말, 반말의 변환은 Split 전에 진행된다.

* 1. 자모 단위 변환

본 논문은 자모 단위에 BPE를 적용하기 위해 Tokenize와 BPE 단계 사이에 자모 단위 변환을 진행하였다. 자모 단위 변환은 오로지 한국어에만 작용하며 초성, 중성, 종성의 구조로 이루어진다. 그리고 ‘나’와 같이 종성이 없는 경우, 종성은 ‘\_’로 대체하게 된다. 예를 들어, 자모 단위 변환을 적용하면 ‘하이. Kim’이 ‘ㅎㅏ\_ㅇㅣ\_. Kim’으로 변환된다.

이러한 자모 단위 변환을 통한 BPE 생성은 기존과 다른 2가지의 차이점이 있다. 바로 Vocabulary 크기의 감소[7]와 더 잠재력 있는 BPE의 생성이다.

* + 1. Vocabulary 크기의 감소

본 논문의 자모 단위 변환 후 BPE를 생성하면 Vocabulary 크기가 줄어드는 것을 확인하였다. 기존의 경우 서브 워드 단위로 BPE를 생성하게 되고, 이들 중 빈도수가 가장 높은 12,000개의 서브 워드의 집합을 Vocabulary로 사용하게 된다. 이때 총 4,340개의 단어는 표현할 수 없게 된다. 그에 반해 자모 단위 변환을 적용하면 13,068개의 고유한 단어가 나온다. 따라서 1,068개의 단어를 표현할 수 없게 된다. 이는 자모 단위 변환이 동일한 Vocabulary 크기에 대해 잃어버리는 고유한 단어가 적다는 것을 의미한다.

* + 1. 잠재력 있는 BPE의 생성

기존의 BPE는 서브 워드 단위로 BPE를 생성하게 된다. 예를 들어 ‘영흰’이라는 단어의 경우 ‘영’과 ‘흰’을 분리하여 BPE를 진행하고 때에 따라 ‘영’과 ‘흰’ 또는 ‘영흰’으로 BPE가 생성된다. 두 가지 중 어떠한 상황에서도 ‘영흰’에 붙은 조사 ‘ㄴ’을 땔 수 없다. 하지만 자모 단위 BPE는 ‘ㅇ’, ‘ㅕ’, ‘ㅇ’, ‘ㅎ’, ‘ㅢ’, ‘ㄴ’으로 나뉘기 때문에 데이터에 따라 ‘ㄴ’이 떨어질 가능성이 생긴다. 이는 번역 결과에 유의미한 차이를 가져오며 많은 경우 조사가 잘 분리되는 것을 확인할 수 있다. 아래의 예시를 보면 서브 워드 단위 BPE와 자모 단위 BPE의 차이를 알 수 있다. 서브 워드 단위 BPE는 ‘영흴’과 ‘교횐’을 제대로 번역하지 못하지만, 자모 단위 BPE는 명사에 붙은 조사를 분리하고 제대로 번역한다.

2022.05.01(일) 기준

입력 문장 : 우리는 영흴 좋아한다.

BPE(서브 워드) : 우리는 영@@ 흴 좋아@@ 한다 .

Output(서브 워드) : We like Young-Som.

BPE(자모) : ㅇㅜ\_ㄹㅣ\_ㄴㅡㄴ ㅇㅕㅇ@@ ㅎㅢ@@ ㄹ ㅈㅗㅎㅇㅏ\_@@ ㅎㅏㄴㄷㅏ\_ .

Output(자모) : We like Younghee.

입력 문장 : 교횐 지금 축제의 분위기다.

BPE(서브 워드) : 교@@ 횐 지금 축@@ 제의 분위기@@ 다 .

Output(서브 워드) : Kyoto, the atmosphere of the festival is now.

BPE(자모) : ㄱㅛ\_@@ ㅎㅚ@@ ㄴ ㅈㅣ\_ㄱㅡㅁ ㅊㅜㄱ@@ ㅈㅔ\_ㅇㅢ\_ ㅂㅜㄴㅇㅟ\_ㄱㅣ\_@@ ㄷㅏ\_ .

Output(자모) : The church is in the mood of the festival now.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 격식 체 | | | | 비 격식체 | |
| 하십시오 | 하오 | 하게 | 해라 | 해요 | 해 |
| 평서형 | (ㅂ니)다 | -오 | -네 | -(ㄴ다) | -어요 | -어 |
| 감탄형 |  | -구려 | -구먼 | -구나 | -어요 | -어 |
| 의문형 | (ㅂ니까) | -오 | -ㄴ가 | -냐 | -어요 | -어 |
| 명령형 | -시오 | -오 | -게 | -어라 | -어요 | -어 |
| 청유형 | -시다 |  | -세 | -자 | -어요 | -어 |

* 1. 높임말, 반말의 변환

표 1. 격식체와 비격식체

한국어 문장은 ‘단어’와 같이 의미를 가진 요소들의 배열로 구성된다. 형태소는 이런 단위 중 가장 작은 단위이며[18], 같은 문장을 단어 단위로 분해하는 것보다 더 세밀하게 분해할 수 있다. 아래는 단어 단위 분해와 형태소 단위의 차이를 보여주는 예시이다.

원본 문장: 철수가 비빔밥을 맛있게 먹고 있다.

단어 단위 분해

- 철수 / 가 / 비빔밥 / 을 / 맛있게 /

먹고 / 있다.

형태소 단위 분해

- 철수 / 가 / 비빔밥 / 을 / 맛 / 있- /

-게 / 먹- / -고 / 있 - / -다.

본 논문은 형태소 중에서 어말 어미와 인칭대명사를 높임말과 반말을 구분할 수 있는 기준으로 선택했다[15]. 아래의 예시는 어말 어미에서 높임말, 반말 어휘를 찾은 것이다.

높임말: **저**는 아침마다 회사에 출근**합니다**.

- **저** / 는 / 아침 / 마다 / 회사 / 에 / 출근 / 하 / **-ㅂ니다**.

반말: **나**는 아침마다 회사에 출근**한다**.

- **나** / 는 / 아침 / 마다 / 회사 / 에 / 출근 / 하 / **-ㄴ다**.

따라서, 형태소 단위의 어휘 치환은 다음과 같은 특징을 지닌다. 첫째, 문장을 높임말, 반말로 변환할 수 있다. 둘째, 의미를 가진 가장 작은 단위를 교체하므로 원본 문장의 정보 유지에 유리하다.

본 논문에서는 높임말, 반말 변환 시 표 1를 참고한 상대 높임법을 고려했으며, 종결어미 간 변환은 표 2의 격식체와 비격식체 각각 아주 높임에서 아주 낮춤으로 변환되도록 했다. 격식체의 경우 높임 표현과 낮춤 표현에는 두 어체가 있지만, 최근 사용이 줄고 있는 ‘하오’ 체와 ‘하게’ 체 대신, ‘하십시오’ 체와 ‘해라’ 체로 변환되도록 했다.

표 2. 상대높임법

|  |  |  |  |
| --- | --- | --- | --- |
|  | 종결어미 | | 인칭대명사 |
| 격식체 | 비격식체 | 대이름씨 |
| 아주 높임 | ‘하십시오’ 체 | ‘해요’ 체 | 어르신, 당신, 어른 |
| 예사 높임 | ‘하오’ 체 | 당신, 그대 |
| 예사 낮춤 | ‘하게’ 체 | ‘해’ 체 | 자네 |
| 아주 낮춤 | ‘해라’ 체 | 너 |

* + 1. 어휘 치환 이전 과정

어휘 치환 이전 과정은 높임말, 반말 변환 이전에 입력 문장에서 치환에 필요한 정보들을 탐색하는 과정으로 다음과 같은 과정을 거친다.

1. 입력 문장을 형태소 단위로 분해 및 분석한다.

입력 문장: 저는 아침마다 회사에 출근합니다.

형태소 분석(어휘/형태소) :

저/인칭대명사 + 는/보조사 + 아침/일반명사 + 마다/보조사 + 회사/일반명사 + 에/부사격 조사 + 출근/일반명사 + 하/동사 파생 접미사 + ㅂ니다/어말 어미

1. 표 2를 참고하여 높임말 또는 반말의 특성을 나타내는 어휘와 형태소를 확인한다. 즉 어말어미와 인칭대명사를 확인하게 된다.

형태소 분석(어휘/형태소) :

**저/인칭대명사(존대형 인칭대명사)** + 는/보조사 + 아침/일반명사 + 마다/보조사 + 회사/일반명사 + 에/부사격 조사 + 출근/일반명사 + 하/동사 파생 접미사 + **ㅂ니다/어말 어미(‘하십시오’ 체)**

* + 1. 높임말로의 변환

입력 문장을 높임말로 변환하는 과정은 표 2의 ‘해라’ 체와 ‘해’ 체가 적용 대상이 되며, 이들을 각각 ‘하십시오’ 체와 ‘해요’ 체로 변환하여 문장을 높임말로 통일시킨다. 어휘 치환 이전 과정을 통해 ‘반말’이라 판단되면 다음의 과정을 따라 높임말로 변환된다.

1. 어휘 치환 이전 과정을 통해 확인한 어휘와 형태소를 확인하고 표 1과 표 2를 통해 반말의 특성을 가지는 경우 높임말에 해당하는 어휘를 확인한다.

어휘 치환 이전 과정의 결과

- ㄴ다/어말 어미(‘해라’ 체)

하십시오 체의 어말 어미

- ㅂ니다, 습니다, 입니다

1. 치환 대상 어휘의 주변 어휘, 형태소를 확인하여 치환할 존댓말 어휘를 찾는다.

어휘 치환 이전 과정의 결과

- ㄴ다/어말 어미(‘해라’ 체)

하십시오 체의 어말 어미

- ㅂ니다, 습니다, 입니다

* + 1. 반말로의 변환

반말로의 변환은 높임말로의 변환과 달리, 종결어미를 용언의 활용, 품사 등과 함께 고려해서 변경해야 한다. 이에 따라, 종결어미의 종류에 따라 1:1로 변경되지 못하는 경우가 존재하게 된다. 예를 들어, ‘주세요’라는 문장은 반말로 변경 시 어간의 종성 ‘ㅜ’가 탈락하지 않아 ‘줘’가 되지만, ‘푸세요’라는 문장은 원형이 ‘푸다’인 동사가 활용되어 종성 ‘ㅜ’가 탈락해 ‘퍼’로 변환되는 불규칙 활용이다[17]. 이처럼 높임말에서 반말로 변경할 때 문법적 요소를 고려해야 하며, 본 논문에서는 이를 표 3과 같이 6가지 예외적인 경우로 분류하고 처리하도록 하였다.

표 3. 반말로 변환 시 예외 규칙

|  |  |  |
| --- | --- | --- |
| 예외 | 예외 경우 | 이유 |
| 1 | ‘-아/어’ 종결어미 | ‘어’는 끝음절의 모음이 ‘ㅏ, ㅗ’가 아닌 용언의 어간 뒤에 붙음 |
| 2 | 용언의 활용 | 용언의 규칙 활용, 불규칙 활용에 따라 어간이나 어미의 모습이 변하거나 변하지 않는 경우가 존재 |
| 3 | ‘-니다’로 끝나는 종결어미 | 용언의 종류에 따라 변환 시 ‘ㄴ다’로 변경되거나, ‘다’로 변경 |
| 4 | 현재형 동사 변환 | 현재형 동사에서는 현재형 종결어미 ‘는다’로 변경 |
| 5 | ‘ㅂ시다’ 종결어미 | 해라체의 ‘-자’ 대신 해체의 ‘아/어’로 변경 |
| 6 | 서술격조사와 함께 쓰이는 종결어미 | 서술격조사와 종결어미 ‘요’는 ‘-야’로 변경 |

각 어휘의 치환 시 높임 표현과 반말 표현을 일대일로 치환하는 것을 원칙으로 하지만, 위의 표 3에 해당하는 예외 규칙의 경우 그 규칙에 따라 치환한다. 이때, 예외 규칙은 독립적으로만 사용되는 것이 아닌, 문법적 규칙에 따라 결합하여 적용되기도 한다. 종결어미 외에도 감탄사인 ‘예’, ‘아니요’도 알맞은 낮춤 표현으로 변경되도록 하였다.

1. 실험
   1. 데이터

전처리 단계를 통해 데이터에 높임말, 반말 그리고 자모 단위를 적용하고 학습을 진행한다. 데이터는 표 4처럼 적용 유형 별로 총 6개의 데이터로 나뉘게 된다.

표 4. 적용 유형 별 데이터

|  |  |  |  |
| --- | --- | --- | --- |
| 적용 유형  데이터 | 자모 | 높임말 | 반말 |
| 기존 | X | X | X |
| 높임말 | X | O | X |
| 반말 | X | X | O |
| 자모 | O | X | X |
| 높임말 - 자모 | O | O | X |
| 반말 - 자모 | O | X | O |

* 1. 번역 결과

번역 성능의 향상을 평가하기 위해 각 번역 모델의 BLEU 점수를 구하였고 각 번역 모델별 Valid와 Test 데이터에 대한 BLEU 점수는 표 5와 같다. 전체적으로 자모 단위 변환을 적용한 번역 모델의 BLEU 점수가 더 높다. 기존의 번역 모델을 SOTA 번역 모델과 비교한 결과 한영 번역의 경우 Valid는 0.43, Aihub\_test는 0.52, Hgu\_test는 1.59가 높아졌고 영한 번역의 경우 Valid는 0.56, Aihub\_test는 0.61, Hgu\_test는 0.72가 높아졌다. 이를 통해 자모 단위 변환을 통한 잠재력 있는 BPE의 생성이 더 높은 성능으로 이어지며 높임말, 반말의 통일 또한 성능의 향상으로 이어짐을 확인했다.

그리고 추가적인 비용 없이 가벼운 어체 변환을 위해 웹에서도 본 논문에서 구현한 어체 변환기를 사용할 수 있다.

BLEU 점수를 통한 정량적인 성능 향상 뿐만 아니라 정성적인 성능의 향상을 보이기 위해 여러 번역 예시를 대중적으로 많이 사용하는 파파고, 카카오, 구글 번역과 함께 보이고자 한다.

표 5. 각 학습 모델 별 BLEU 점수

|  |  |  |  |
| --- | --- | --- | --- |
| 데이터  모델명 | Vaild | 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 |

**Kr는 기존 한국어 데이터, en은 기존 영어 데이터, high는 한국어 높임말 데이터, low는 한국어 반말 데이터, jamo는 한국어 자모 단위 데이터, jamo\_h와 jamo\_l은 각각 한국어 자모 단위 데이터에 높임말과 반말을 적용한 데이터이다. A2B는 A를 B로 번역하는 모델을 의미하며, kr2en은 한영 번역 모델, en2high는 영한(높임말) 번역 모델이다.**

* + 1. 자모 단위 번역 비교

아래의 예시를 보면 ‘교회’ + ‘ㄴ’, ‘영희’ + ‘ㄹ’와 같이 명사에 조사가 결합한 것을 볼 수 있다. 이런 경우 자모 단위 변환을 적용한 번역기는 제대로 번역할 수 있지만 파파고를 제외한 나머지 번역기의 경우 제대로 번역하지 못한다. 이 외에도 많은 경우 자모 단위 번역기는 명사에 조사가 결합한 문장을 처리할 수 있다. 이는 BLEU 점수 뿐만 아니라 정성적으로도 성능이 향상되었음을 보여준다.

2022.04.27(일) 기준

입력 : 교횐 지금 축제의 분위기다.

자모 단위 : The church is in the mood of the festival now.

파파고 : The church is in a festive mood.

카카오 : It is the athmosphere of the festival now.

구글 : It is a festive mood right now.

입력 : 우리는 영흴 좋아한다.

자모 단위 : We like Younghee.

파파고 : We like Younghee.

카카오 : We like Young.

구글 : We like Young.

* + 1. 높임말, 반말 번역 비교

아래의 첫 번째 예시와 같이, 구글 번역과 파파고는 높임말, 반말의 혼용이 일어나는 경우가 있으며, 두 번째, 세 번째 예시에서의 카카오 번역은 두 개의 영어 문장을 하나의 한국어 문장으로 통합하여 출력하는 경우가 있다. 반면, 본 논문에서 제시한 방법을 적용한 번역기의 경우, 이러한 문제를 해결했다.

2022.05.01(일) 기준

입력 : I love you. I will marry you.

구글 번역기: 사랑해요. 나는 당신과 결혼할 것이다.

카카오 번역: 사랑해. 결혼하고 싶어.

파파고: 사랑해요. 나는 너와 결혼할 것이다.

높임말 통일: 사랑해요. 저는 당신과 결혼할 거예요

입력 : I want something to eat. How about you?

구글 번역기: 뭔가 먹고 싶어 당신은 어떤가요?

카카오 번역: 뭐 좀 먹고 싶은데 너는?

파파고: 나 뭐 먹고 싶어. 너는 어때?

높임말 통일: 저는 뭔가 먹고 싶어요. 당신은 어떠세요?

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

구글 번역기: 어제 펜을 샀다. 하지만 고장났습니다.

카카오 번역: 어제 펜을 샀는데 고장이 났어요.

파파고: 나는 어제 펜을 구입했습니다. 하지만 그건 부러졌어요.

높임말 통일: 저는 어제 펜을 샀어요. 하지만 깨져있어요.

1. 결론

본 논문은 한국어의 특수성을 반영하지 못하여 발생하는 번역 성능의 하락을 해결하기 위해 자모 단위 변환과 높임말, 반말의 변환을 제안했다. Transformer 모델 기반 학습을 통한 번역 성능을 확인하였을 때 약간의 BLEU 점수 향상을 이끌어 냈고 정성적으로도 조사가 결합한 경우의 번역 오류를 해결하는 것을 확인할 수 있다. 또한 본 논문에서는 자모 단위로 변환을 진행하였지만, 형태소 단위의 변환을 통한 번역 또한 의미가 있을 것으로 보인다. 하지만 형태소 분석기의 성능이 다르므로 정확한 형태소 분석을 가능하게 하는 것이 도전이 될 것으로 생각된다.

참고문헌

1. H. Choi, and Y. Min (2015). “Introduction to Deep Learning and Major Issues,” *Korea Information Processing Society Review*, Vol. 22, No. 1, p.7-21.
2. D. Bahdanau, K. Cho, and Y. Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate,” *ICLR* .
3. H. Choi (2019). “Understanding Neural Machine Translation,” *Communications of the Korean Institute of Information Scientists and Engineers*, Vol. 37, No. 2, p.16-24.
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin(2017). “Attention Is All You Need,” *arXiv*.
5. 카카오 I 번역 서비스에 적용된 학습 원리와 성능 개선기[웹사이트]. (2022.05.12).URL: https://tech.kakaoenterprise.com/22
6. R. Sennrich, B. Haddow, &A. Birch (2016). Neural Machine Translation of Rare Words with Subword Units.*54th Annual Meeting of the Association for Computational Linguistics.* https://arxiv.org/pdf/1508.07909v5.pdf
7. C. Lee, D. Lee, Y. Hur, K. Yang, & H. Lim (2018). Comparing Byte Pair Encoding Methods for Korean. *제30회 한글 및 한국어 정보처리 학술대회 논문집 (2018년),* 293–294. https://www.koreascience.or.kr/article/CFKO201832073078730.pdf
8. K. Papineni, S. Roukos, T. Ward, & W. J. Zhu(n.d.). BLEU: A Method for Automatic Evaluation of Machine Translation.*Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, Pp. 311-318.* https://aclanthology.org/P02-1040.pdf
9. B. Kim, H. Kim, S. W. Lee, G. Lee, D. Kwak, D. H. Jeon, ... & N. Sung(2021). What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. *arXiv preprint arXiv:2109.04650*. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *NAACL.* <https://arxiv.org/pdf/1810.04805.pdf>
10. J. Devlin, M. W. Chang, K. Lee, & K. Toutanova(2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *NAACL.* https://arxiv.org/pdf/1810.04805.pdf
11. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding with Unsupervised Learning," *Technical report, OpenAI,* 2018.
12. S. Eo, C. Park, H. Moon, & H. Lim(2021). Research on Subword Tokenization of Korean Neural Machine Translation and Proposal for Tokenization Method to Separate Jongsung from Syllables.*Journal of the Korea Convergence Society, 12(3),* 1–7. <http://koreascience.or.kr/article/JAKO202111037333482.pdf>
13. 김혜미. (2015). [한국에서 미국을, 미국에서 한국을, 새롭게 다시 보기] 존댓말과 상호존중. *새가정*, *62*, 45-49.
14. 국립국어원. *알고 싶은 한글.* (2008). *국립국어원*. https://www.korean.go.kr/hangeul/principle/001.html
15. 한길. (2002). *현대 우리말의 높임법 연구(초판).* 역락.
16. 최형용. (2016). *한국어 형태론*. 역락.
17. S. Kang, & Y. Kim (1992). A Computational Analysis Model of Irregular Verbs in Korean Morphological Analyzer. *정보과학회논문지*, 19(2), 151-164. https://www.dbpia.co.kr/journal/voisDetail?voisId=VOIS00028769

|  |
| --- |
| 김 정 희 |
| 사람, 실내이(가) 표시된 사진  자동 생성된 설명 |
| 2017~ 현재: 한동대학교 전산전자공학부 재학 |
| 관심분야: 머신러닝, 딥러닝, 인공지능 |
| 허 재 무  사람, 남자, 실내, 천장이(가) 표시된 사진  자동 생성된 설명  2017~ 현재: 한동대학교 전산전자공학부 재학  관심분야: 머신러닝, 딥러닝, 인공지능 |
| 김 주 환 |
| 사람, 실내, 천장이(가) 표시된 사진  자동 생성된 설명 |
| 2017~ 현재: 한동대학교 전산전자공학부 재학 |
| 관심분야: 머신러닝, 딥러닝, 인공지능 |
|  |
| 최 희 열 |
| A person wearing glasses  Description automatically generated with medium confidence |
| 2010, Texas A&M University, Computer Science and Engineering(PhD),  2010 ~2011 Indiana University, Cognitive Science Program(Post-Doc),  2011~2016 삼성전자 종합기술원(전문연구원),  2015 ~2016 Univ. of Montreal(방문연구원),  2016 ~현재 한동대학교 전산전자공학부(부교수) |
| 관심분야: Deep Learning, Cognitive Science |