한국어 자모 단위 구성과 높임말을 반영한 한영 번역

(Kor-Eng NMT Using Characteristic of Korean Alphabet and Honorific Expression)

김정희† 허재무† 김주환† 최희열§

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

요 약 최근 딥러닝은 기계 번역의 성능을 향상시켜왔지만, 언어별 특성을 반영하지 못하는 경우가 대부분이다. 특별히, 한국어의 경우 단어와 표현에 있어서 독특한 특징들이 오번역을 일으킬 수 있다. 예를 들어, 구글 번역은 한국어 문장의 명사에 종성 형태로 조사가 결합하여 오역하는 경우가 존재하며, 여러 영어 문장이 입력되면 존댓말과 반말이 혼용된 일관성 없는 한국어 문장들을 출력한다. 이는 한국어 단어의 ‘자모 단위 구성’과 ‘존댓말과 반말 간의 구분’이라는 특성이 반영되지 않아 생기는 문제이다. 본 논문에서는 이를 해결하기 위해 번역 모델에 사용하는 서브워드를 자모 단위로 구성하고, 코퍼스 문장들에서 존댓말과 반말을 통일하여 모델을 학습하는 것을 제안한다. 실험에서 제안된 방법의 성능을 검증하기 위해 기존 방법 및 코퍼스와 비교한 결과 위 문제들을 해결하면서 BLEU 점수에서 기존과 유사하거나 소폭 향상된 것을 확인했다.

키워드 : 신경 기계 번역, 한국어 단어, 자모, 존대어

*Abstract:* Recently, deep learning has improved the performance of machine translation, but in most cases, it does not reflect the characteristics of the languages. In particular, Korean has unique linguistic features in words and expressions which might cause mistranslation. For example, in Google Translate from Korean to English, there are mistranslations when a postposition (‘josa’) in a Korean sentence is combined with a noun in the form of a consonant. Also, in the English-Korean translation, the honorifics and casual expressions are mixed in the translated results. It is because the alphabetic characteristics and honorifics of the Korean language are not reflected. In this paper, to solve the problems, we propose to train the model with subwords composed of units of letters (‘jamo’) and unifying honorific and casual expressions in the corpus. Experiment results confirm that the proposed method resolves the problems mentioned above, and has a similar or slightly higher BLEU score compared to the existing method and the corpus.

Keywords : neural machine translation, Korean words, jamo, honorifics

\* 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2018-0-00749,인공지능 기반 가상 네트워크 관리기술 개발). 또한2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2022R1A2C1012633)

† 학생회원 : 한동대학교 전산전자공학부

§종신회원 : 한동대학교 전산전자공학부 부교수

[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 있고, 자연어 처리 분야와 기계 번역도 많은 발전을 이루었다 [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 오번역되는 경우가 존재한다.

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

(2022.03.21(월) 번역 결과)

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) 번역 결과)

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 프로세싱을 위한 모델이 추가로 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 문장의 단어들을 [‘좋다.’ ‘ㅈㅗㅎㄷㅏ\_.’]와 같이 자모 단위로 분리한 후 Byte Pair Encoding (BPE)[6]를 진행한다. 이에 따라 더 효율적인 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 하며, 높임말 혹은 반말로 통일된 데이터로 번역모델을 학습함으로써 효과적인 학습을 가능하게 한다. 이를 통해, 사용자에게 높임말/반말을 선택할 수 있게 서비스를 개선할 수 있다.

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 추가 데이터를 사용한다 [8]. 번역 모델의 정량적 성능 평가는 BLEU(Bilingual Evaluation Understudy) 점수를 사용하며, 제안된 방법이 기존 번역보다 성능이 유사하거나 소폭 향상되는 것을 확인했으며, BLEU 점수로는 판단할 수 없는 정성적 비교를 위해 네이버 파파고, 구글 번역, 카카오 번역의 결과와 함께 제시한다.

1. 배경 지식
   1. Transformer

입력 문장의 문맥을 반영하여 번역하는 모델로 초기 Recurrent Neural Networks (RNNs)를 활용한 자연어 번역 모델이 등장했다 [1]. 하지만 RNNs는 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 몇가지 방법들이 있는데, 그 중에 하나가 Transformer 모델이며, 이는 단어 간의 Attention Mechanism만을 사용하여 단어간 관계를 이해한다. Transformer 모델은 이전의 모델들 보다 더 좋은 번역 성능을 보여주며, 현재 자연어 처리분야에 활발히 사용되고 있다 [4]. 특히, BERT, GPT 그리고 국내의 카카오 번역과 네이버의 Hyper Clova 등이 Transformer를 사용 중이다 [9][10][11].

* 1. Byte Pair Encoding (BPE)

Neural Machine Translation (NMT)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 글자 (alphabet) 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

BPE를 사용하면 사전에 등장하지 않는 단어가 입력되더라도 서브 워드의 조합으로 설명이 가능해진다.

* 1. 한국어의 특성

언어에는 해당 문화권에 따라 다양한 언어적 특성들이 존재할 수 있는데, 한국어에도 그러한 특징들이 있다. 본 논문에서는 한국어의 다양한 특성들 중 다음의 두가지 특징들을 다루려고 한다. 첫째, 한국어는 각 음절이 초성, 중성 그리고 종성으로 이루어진 자모 단위로 표기된다. 둘째, 한국어는 어미변화 등의 문법적인 존대법체계를 가지고 있다 [13].

* + 1. 자모 단위 구성

한글은 자음과 모음을 나타내는 글자가 따로 존재하는 자소 문자이며, 자음과 모음을 합쳐서 음절 단위로 사용하는 음절 문자적 성격을 지니고 있다 [14]. 한글은 19개의 자음 23개의 모음으로 구성되어 있다. 음절과 자소에 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 분석

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 해당되며 자립 형태소는 그 자체로 자립성이 있는 말로, 주로 체언의 어근, 감탄사, 관형사, 부사 등이 해당된다 [17].

* + - 1. 상대 높임법

한국어는 청자에 대한 화자의 태도에 따라 여러 문체가 존재하며 문장의 종결 유형별로 문체를 구분한다. 문장의 목적에 따라 평서법, 의문법, 명령법, 청유법, 감탄법으로 나뉘며 청자 또는 높인 정도에 따라 ‘하십시오’, ‘하오’, ‘하게’, ‘해라’, ‘해요’, ‘해’ 체로 나뉘며 그 중에서 ‘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. 관련 연구

기존의 한국어 단어의 구성에 따른 번역 연구들로는 자소와 음소 정보에 기반한 transliteration 연구가 있고 [18], 최근 tokenization 전략에 따른 성능의 변화를 확인하는 연구에서 한국어의 특수성을 전처리 과정에 반영한 사례가 있다 [19]. 본 논문은 자모 처리 후 BPE 과정을 거친다는 점에서 위 논문과는 다른 접근이다.

한국어의 존댓말 변환을 위한 연구로는 OpenNMT 를 사용한 사례가 있지만, 의미 파악에 한계가 있는 것으로 판단된다 [20]. 네이버 파파고에서는 높임말 On/Off 기능을 추가했고, 카카오 번역에서는 3가지 문체 (기본문체/높임말/예사말) 중 하나를 선택할 수 있게 했는데, 이들은 문체에 맞게 번역을 하던지 혹은 새로운 프로세싱을 추가로 진행하는 것으로 판단된다. 이는 본 논문에서 제안하는 간단한 문체 변경 프로세스와는 다른 과정으로 이해된다.

그 외에도 영한 번역 품질 향상을 위해 언어적 차이를 고려한 프리에디팅 혹은 포스트에디팅이 제안된 바 있다 [21][22].

1. 제안 모델

본 논문은 한국어의 특성을 반영한 한영 번역 성능의 개선을 위해 자모 단위의 서브워드 사용과, 높임말/반말의 변환 사용을 제안한다. 번역 모델은 Transformer 기본 구조를 그대로 사용했고, 모델 학습을 위해 AIHub 한국어-영어 번역(병렬)을 포함한 3,525,933개 문장을 사용했다 [8][23].

학습을 위해 데이터를 가공하는 전처리 과정은 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 적용하게 된다. 자모 단위 변환은 Tokenize 이후에 진행되며 높임말, 반말의 변환은 Split 전에 진행된다.

* 1. 자모 단위 변환

자모 단위에 BPE를 적용하기 위해 Tokenize와 BPE 단계 사이에 자모 단위 변환을 진행한다. 자모 단위 변환은 오로지 한국어에만 작용하며 초성, 중성, 종성의 구조로 이루어진다. 그리고 ‘나’와 같이 종성이 없는 경우, 종성은 ‘\_’로 대체한다. 예를 들어, 자모 단위 변환을 적용하면 ‘반가워’가 ‘ㅂㅏㄴㄱㅏ\_ㅇㅝ\_’로 변환된다.

이러한 자모 단위 변환을 통한 BPE 생성은 기존에 비해 Vocabulary 크기가 감소하고 [7]와 더 효율적인 서브워드 리스트를 생성하게 된다.

* + 1. Vocabulary 크기의 감소

자모 단위 변환 후 BPE를 적용하면 Vocabulary 크기가 줄어드는 것을 확인하였다. 기존의 경우 서브 워드 중 빈도수가 가장 높은 12,000개로 Vocabulary를 구성하게 되는데, 이때 총 4,340개의 단어는 표현할 수 없게 된다. 그에 반해 자모 단위 변환을 적용하면 13,068개의 고유한 단어가 나오고, 총 1,068개의 단어를 표현할 수 없게 된다. 즉, 자모 단위 변환이 동일한 Vocabulary 크기에 대해 잃어버리는 고유한 단어가 적다는 것이고, 이는 번역 모델 학습에 더 효과적인 단어를 구성하는 것을 의미한다.

* + 1. 효율적인 BPE 서브워드 생성

기존의 BPE는 음절을 최소 단위로 서브 워드 리스트를 생성하게 된다. 예를 들어 ‘영흰’이라는 단어의 경우 ‘영’과 ‘흰’을 분리하여 BPE를 진행하고 때에 따라 ‘영’과 ‘흰’ 또는 ‘영흰’으로 서브워드가 생성된다. 즉, 기존 방법으로는 ‘영희’와 ‘ㄴ’을 서브워드로 만들 수 없다. 하지만 자모 단위 BPE는 ‘ㅇ’, ‘ㅕ’, ‘ㅇ’, ‘ㅎ’, ‘ㅢ’, ‘ㄴ’으로 나뉘기 때문에 데이터에 따라 ‘ㄴ’이 떨어질 수 있다. 이는 번역 결과에 유의미한 차이를 가져올 것으로 기대된다.

입력 문장: 우리는 **영흴** 좋아한다.

BPE(기존): 우리는 **영@@ 흴** 좋아@@ 한다 .

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

입력 문장: **교횐** 지금 축제의 분위기다.

BPE(서브워드): **교@@ 횐** 지금 축@@ 제의 분위기@@ 다 .

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

위 예시를 보면 기존 BPE와 자모 단위 BPE의 차이를 알 수 있다. 기존 BPE는 ‘영흴’과 ‘교횐’의 명사와 조사을 제대로 구분하지 못하지만, 자모 단위 BPE는 명사에 붙은 조사를 분리한다.

* 1. 높임말/반말의 변환

한국어에서 형태소는 가장 작은 의미 단위이며 [23], 같은 문장을 어절 단위로 분해하는 것보다 더 세밀하게 분해할 수 있다. 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/ 아침/마다/ 출근/하/**-ㅂ니다**.

반말: **나**/는/ 아침/마다/ 출근/하/**-ㄴ다**.

이러한 형태소 분석에 기반한 단위의 어휘 치환으로 원본 문장의 의미를 훼손하지 않으면서 문장을 높임말/반말로 변환할 수 있다.

표 1. 격식체와 비격식체  
Table 1. Honorifics and Nonhonorifics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Honorific | | | | Nonhonorific | |
| ‘하십시오’ | ‘하오’ | ‘하게’ | ‘해라’ | ‘해요’ | ‘해’ |
| (ㅂ니)다 | -오 | -네 | -(ㄴ다) | -어요 | -어 |
|  | -구려 | -구먼 | -구나 | -어요 | -어 |
| (ㅂ니까) | -오 | -ㄴ가 | -냐 | -어요 | -어 |
| -시오 | -오 | -게 | -어라 | -어요 | -어 |
| -시다 |  | -세 | -자 | -어요 | -어 |

표 2. 대명사에서 상대높임법  
Table 2. Relative Honorifics in Pronoun

|  |  |  |  |
| --- | --- | --- | --- |
| Level of politeness | Ending | | Pronoun |
| Honorific | Nonhonorific |
| Highest | ‘하십시오’ | ‘해요’ | 어르신, 당신, 어른 |
| Hight | ‘하오’ | 당신, 그대 |
| Low | ‘하게’ | ‘해’ | 자네 |
| Lowest | ‘해라’ | 너 |

본 논문에서는 높임말, 반말 변환 시 표 1의 격식체와 비격식체의 종결어미간 변환과 표2의 상대 높임법을 고려하였다. 격식체의 경우 높임 표현과 낮춤 표현에는 두 어체가 있지만, 최근 사용이 줄고 있는 ‘하오’ 체와 ‘하게’ 체 대신, ‘하십시오’ 체와 ‘해라’ 체로 변환했다.

* + 1. 형태소 분석 과정

높임말 혹은 반말로의 변환 이전에 입력 문장에서 치환에 필요한 정보들을 탐색하는 과정으로 다음과 같은 과정을 거친다. 입력 문장을 형태소 단위로 분해하고, 표 2를 참고하여 높임말 또는 반말의 특성을 나타내는 어휘와 형태소를 확인한다. 즉, 아래와 같이 어말어미와 인칭대명사를 확인하게 된다.

입력 문장: 저는 아침마다 회사에 출근합니다.

형태소 분석(어휘/형태소) :

**저/인칭대명사(존대형)** + 는/보조사 + 아침/일반명사 + 마다/보조사 + 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. 표 1과 표 2를 통해 반말의 특성을 가지는 경우 아래와 같이 높임말에 해당하는 어휘를 확인한다.

반말 문장 : 나/는 오늘/도/ 회사/에 가/-ㄴ다.

치환 대상 어휘(반말)의 확인

- 나/인칭대명사 (화자에 대한 반말 특성 지님)

- ㄴ다/어말 어미(‘해라’ 체)

1. 치환 대상 어휘의 주변 어휘, 형태소를 확인하여 치환할 존댓말 어휘를 찾는다. 위의 문장에서 종결어미가 ‘ㄴ다’로 ‘해라’체이다. 이는 현재를 의미하는 종결어미로, 가장 적절한 ‘하십시오’체의 ‘ㅂ니다’로 치환된다. 또한 ‘나’의 경우에는 ‘화자’를 나타내는 주어 중 ‘반말’의 특성을 가지고 있는 어휘이다. 이를 아래와 같이 치환 가능한 ‘높임말’의 특성을 지닌 어휘들로 치환하여 높임말로 변환한다.

각각에 대응하는 높임말에 해당하는 어휘 확인

- ㅂ니다, 습니다, 입니다/하십시오(‘하십시오’ 체)

- 저/인칭대명사 (화자에 대한 높임말 특성 지님)

높임말 문장(치환 결과물) : 저/는 오늘/도/ 회사/에 가/-ㅂ니다.

* + 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 존재하는데, 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  
Table 3. Exceptions for transformation into nonhonorific expression

|  |  |  |
| --- | --- | --- |
| **No** | **Cases** | **Logic** |
| 1 | Endings with ‘-아/어’ | ‘어’ should follow verb stem not ending with ‘ㅏ, ㅗ’ |
| 2 | Use of verbs | Endings may or may not vary depending on regular or irregular use of verbs. |
| 3 | Endings with ‘-니다’ | It transforms into ‘ㄴ다’ or ‘다’ depending on verbs. |
| 4 | Present tense verb conversion | Ending for present tense verb should end with ‘는다’. |
| 5 | Endings with ‘ㅂ시다’ | Ending with ‘아/어’ (해체) instead of ‘-자’ (해라체) |
| 6 | Endings with descriptive article | Descriptive article and ending ‘요’ should be replaced by ‘-야’ |

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.

아래의 예시에서 ‘세요’ 는 ‘시어요’의 줄임말이다. ‘노’는 원형 ‘놀다’에서 ‘ㄹ’이 탈락한 ‘ㄹ’불규칙 활용으로, 해체로 변경할 경우, 원형이 복귀되어야 하는 표3의 예외규칙 2번에 해당한다. 따라서 선어말 어미와 종결어미가 결합된 ‘시어요’는 제거되고 ‘놀’로 변경된다. 종결어미 ‘어요’는 ‘해요’체로 ‘해’체로 변경 되어야 하는데, 이 경우 종결어미 ‘-아’ 혹은 ‘-어’가 올 수 있다. 표3의 예외 규칙 1번에 의해 ‘-아’가 선택이 되며 최종적으로 ‘놀아’로 변경된다. 이런 예외 사항을 제외한 변환 과정은 ‘높임말’로의 변환 과정과 동일하다.

높임말: 그럼/ 재미/있게/ 노/세/요.

예외규칙 2번 적용: 그럼/ 재미/있게/ 놀.

예외규칙 1번 적용: 그럼/ 재미/있게 놀/아.

종결어미 외에도 감탄사인 ‘예’, ‘아니요’도 알맞은 낮춤 표현으로 변경되도록 하였다.

1. 실험
   1. 데이터

실험에 사용한 데이터는 다음과 같다. AIHub에서 제공하는 데이터와 웹에서 크롤링한 데이터를 사용했다 (hancorpus.github.io). 크롤링한 데이터는 재번역을 통해 유효하지 않은 문장들은 제거하는 클렌징 작업을 거쳤다 [8]. Validation과 Test1의 경우 각각 3,000개씩 AIHub 데이터에서 따로 추출했으며, Test2의 경우 크롤링한 데이터에서 3,000개를 추출했다. Validation과 Test를 제외한 나머지 데이터 3,525,933개는 학습데이터로 사용되었다.

본 논문에서 제시한 데이터의 변형을 실험하기위해, 표 4에서 처럼, 어체 변환과 자모 변환의 적용 유형에 따라 서로 다른 전처리 단계를 진행하여 학습 데이터의 변형을 만들고 각각의 변형된 데이터에 대해 모델을 학습한다. 높임말과 반말로 변환한 경우 전체 문장들 중에 각각 69.8%와 28.1% 문장들이 변환되었다.

표 4. 적용 유형 별 데이터: 자모 변환 (Jamo), 높임말 변환 (Honor), 반말 변환 (NHonor)  
Table 4. Data with transformations into Jamo, Honorific(Honor) and/or Nonhonorific (NHonor)

|  |  |  |  |
| --- | --- | --- | --- |
| transformation  data | Jamo | Honor | NHonor |
| Original data | X | X | X |
| +Honor | X | O | X |
| +NHonor | X | X | O |
| +Jamo | O | X | X |
| +Jamo +Honor | O | O | X |
| +Jamo +NHonor | O | X | O |

* 1. 번역 결과
     1. 정량적 비교

번역 성능을 평가하기 위해 각 번역 모델에 대한 Valid 와 Test 데이터에 대한 BLEU 점수를 구하였고 표 5에 정리했다.

이 표에서 자모 단위 변환과 높임말/반말로의 변환을 적용한 번역 모델의 BLEU 점수가 기존 모델과 유사하거나 소폭 향상된 것을 확인할 수 있다. 기존 결과에 비해 한영 번역의 경우 Valid는 0.26, Test1은 0.12, Test2는 0.81가 높아졌고, 영한 번역의 경우 Valid는 0.37, Test1은 0.61, Test2는 0.73가 높아졌다. 이를 통해 자모 단위 변환을 통한 효율적인 서브워드 생성이 더 높은 성능으로 이어질 수 있으며, 높임말/반말의 통일 또한 성능을 향상할 수 있음을 확인했다.

표 5. 각 모델별 BLEU 점수 (테스트데이터 별 최고점수는 볼드체로 표시)  
Table 5. BLEU scores of models with different transformations (The best scores for each test data are in red bold.)

|  |  |  |  |
| --- | --- | --- | --- |
| Test data  Model | Vaild | Test1 | Test2 |
| Kr2En | 39.12 | 33.83 | 25.79 |
| +Honor | 39.15 | **33.95** | 25.93 |
| +NHonor | 39.08 | 33.82 | 26.27 |
| +Jamo | 39.35 | 33.64 | **26.60** |
| +Jamo+Honor | **39.38** | 33.37 | 26.45 |
| +Jamo+NHonor | 39.18 | 33.67 | 26.04 |
| En2Kr | 20.93 | 13.90 | 11.00 |
| +Honor | 21.14 | **14.51** | 11.26 |
| +NHonor | 20.98 | 13.76 | **11.73** |
| +Jamo | 21.09 | 13.83 | 11.17 |
| +Jamo+Honor | **21.30** | 13.77 | 11.22 |
| +Jamo+NHonor | 21.19 | 13.85 | 11.08 |

* + 1. 정성적 번역 비교

BLEU 점수를 통한 정량적인 성능 뿐만 아니라 정성적인 성능을 확인하기 위해 여러 번역 예시를 파파고, 카카오, 구글 번역과 함께 보이고자 한다.

* + - 1. 자모 단위 번역

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.

(파파고, 카카오, 구글 번역은 2022.04.27(일) 결과)

**입력 : 교횐 지금 축제의 분위기다.**

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.

자모 단위 : The church is in a festive mood now.

**입력 : 우리는 영흴 좋아한다.**

파파고 : We like Younghee.

카카오 : We like Young.

구글 : We like Young.

자모 단위 : We like Younghee.

반면, 일상생활에서 비교적 사용 빈도가 적은 ‘추팔 던진다(추파를 던진다)’ 같은 표현의 경우, 제대로 번역하지 못하는 모습을 보인다 (아래 예시 참고). 이는 자모 단위 BPE를 적용했음에도, 학습 데이터 상에 등장하는 빈도가 낮을 경우 번역 시 한계가 있음을 보여준다.

(파파고, 카카오, 구글 번역은 2022.08.25(목) 결과)

**입력 : 그가 계속 추팔 던진다.**

파파고 : He keeps flirting.

카카오 : He keeps flirting.

구글 : he keeps throwing

자모 단위 : He is throwing a ball.

* + - 1. 높임말 및 반말 변환 후 번역

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

(파파고, 카카오, 구글 번역은 2022.05.01(일) 결과)

**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: 저는 당신을 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?

번역 성능의 향상 외에도, 본 논문에서 제안한 높임말 반말 변환을 통해 추가적인 복잡한 과정 없이 번역 결과에 대한 어체 변환을 할 수 있다는 장점이 있다.

그렇지만, 반말로의 변환 과정에 대하여, ‘~하세요’의 경우 명령형으로 생각하여 ‘~해’로 변환되도록 했기 때문에, 명령형에 해당하지 않는 ‘~하세요’는 잘 못 변환되며, 이렇게 변환된 데이터를 바탕으로 학습되었기 때문에 아래와 같은 번역상의 오류가 발생한다.

(파파고, 카카오, 구글 번역은 2022.08.25(목) 결과)

**입력 : Hello.**

구글 번역기: 여보세요.

카카오 번역:.안녕하세요.

파파고: 안녕.

반말 통일: 안녕해.

이를 개선하기 위해서는 형태소 분석기의 발전과 더불어 ‘명령법’, ‘의문법’ 등을 구분할 수 있는 모듈의 추가가 필요하다.

* + - 1. 영한/한영 번역 결과 예시

아래에 추가적인 번역 결과들을 통해 번역의 질을 보다 객관적으로 비교 확인 할 수 있다.

**입력: Thank you very much. I tried to make it more complete and organized. I think this one is much more convenient to use for me.**

* 구글: 매우 감사합니다. 더 완벽하고 체계적으로 만들려고 노력했습니다. 저한테는 이게 훨씬 편리한 것 같아요.
* 파파고: 정말 감사합니다. 나는 그것을 더 완전하고 조직적으로 만들려고 노력했다. 나는 이것이 나에게 훨씬 더 편리하다고 생각한다.
* 카카오: 정말 감사합니다. 나는 그것을 좀 더 완벽하고 체계적으로 만들기 위해 노력했다.나는 이것이 나를 위해 사용하기에 훨씬 더 편리하다고 생각한다.
* 높임말 통일: 정말 감사합니다. 좀 더 완벽하고 조직적으로 만들기 위해 노력했습니다. 이 제품이 저한테는 훨씬 편리한 것 같아요.
* 반말 통일: 정말 감사한다. 나는 좀 더 완벽하고 체계적으로 만들려고 노력했다. 이 제품이 나한테는 훨씬 편리한 것 같아.

**입력: I think it makes more sense for me to pay the extra-baggage charge and bring them back on the plane with me. If I send them, we'll have to wait at least a week to get them. I want to get them into the lab so they can start analyzing them as soon as possible.**

* 구글: 추가 수하물 요금을 지불하고 비행기에 다시 가지고 오는 것이 더 합리적이라고 생각합니다. 보내면 적어도 일주일은 기다려야 합니다. 나는 그들이 가능한 한 빨리 분석을 시작할 수 있도록 실험실로 데려가고 싶습니다.
* 파파고: 내 생각엔 초과 수화물 요금을 지불하고 기내에 갖고 타는 게 더 나을 것 같아요. 만약 내가 그것들을 보낸다면, 우리는 그것들을 받기 위해 적어도 일주일은 기다려야 할 것이다. 최대한 빨리 분석을 시작할 수 있도록 실험실로 데려가고 싶어요.
* 카카오: 추가 수하물 요금을 지불하고 그들을 다시 비행기에 태우는 것이 더 이치에 맞는 것 같다. 내가 그들을 보내면, 우리는 그들을 얻기 위해 적어도 일주일은 기다려야 할 것이다.나는 그들이 가능한 한 빨리 그것들을 분석하기 시작할 수 있도록 그들을 연구실로 데려가고 싶다.
* 높임말 통일: 초과 수화물 요금을 지불하고 기내에 가지고 가는 게 더 말이 되는 것 같아요. 보내드리면, 받으려면 적어도 일주일은 기다려야 할 것 같아요. 연구실에 입고하여 빠른 시일 내에 분석을 시작할 수 있도록 하고 싶습니다.
* 반말 통일: 내 생각엔 초과 수화물 요금을 지불하고 기내에 가지고 타는 게 더 나을 것 같아. 보내면, 적어도 일주일 이상 기다려야 받을 수 있어. 빨리 분석할 수 있도록 실험실에 맡기고 싶어.

**입력: When you look at certain places on the new note from an angel you will see a map of Korea and the yin-yang symbol called the 'Taeguk'. Also there will be numbers printed on the note that are so small they will only be able to be read with a microscope.**

* 구글: 천사의 새 노트의 특정 장소를 보면 한국의 지도와 "태극"이라고 하는 음양 기호를 볼 수 있습니다. 또한 메모에는 현미경으로만 읽을 수 있을 정도로 작은 숫자가 인쇄되어 있습니다.
* 파파고: 여러분이 천사에서 새 지폐의 특정 장소를 볼 때, 여러분은 한국의 지도와 '태극'이라고 불리는 음양의 상징을 보게 될 것입니다. 또한 지폐에는 현미경으로만 읽을 수 있을 정도로 작은 숫자가 인쇄될 것이다.
* 카카오: 천사가 보낸 새 쪽지의 특정 장소를 보면 한국 지도와 '태극'이라고 불리는 음양 상징이 보일 것이다.또한 너무 작아서 현미경으로만 읽을 수 있는 숫자가 노트에 인쇄될 것이다.
* 높임말 통일: 새로운 지폐에 새겨진 특정 장소를 천사로부터 바라보면 한국의 지도와 ‘ 태극기 ’ 라는 음양 문양이 보입니다. 또한 노트에 너무 작아서 현미경으로만 읽을 수 있는 숫자가 적혀있습니다.
* 반말 통일: 천사의 새 지폐에 적힌 특정 장소를 보면 한국의 지도가 나오고, ‘ 태극기 ’ 라는 음양 문양이 나온다. 또한 지폐에 작은 숫자가 인쇄되어 현미경으로만 읽을 수 있을 것이다.

**입력: 영흰 평소에 교횔 자주 간다.**

* 구글: Young-hui usually goes to school often.
* 파파고: Young-hee usually goes to church often.
* 카카오: Young-huin usually goes to school often.
* 자모 단위: Younghee goes to church often.

**입력: 우리 오빤 평소에 용돈을 자주 준다. 그리고 누난 용돈을 안 준다.**

* 구글: My brother usually gives pocket money often. And noona gives no pocket money.
* 파파고: My brother often gives me pocket money. And my sister doesn't give me allowance.
* 카카오: My brother usually gives me pocket money. And he doesn't give me pocket money.
* 자모 단위: My brother often gives me pocket money. And my sister doesn't give me any allowance.

**입력: 나는 성가댈 지원했다. 그래서 목소릴 가다듬었다. 이후 앉아서 내 차롈 기다렸다. 하지만 나의 기횐 주어지지 않았다.**

* 구글: I supported the choir. So I straightened my voice. After that, I sat down and waited for my car. But my blessing was not given.
* 파파고: I applied for a choir. So I cleared my voice. After that, I sat down and waited for my turn. But my opportunity was not given.
* 카카오: I applied for annoyance. So I cleared my throat. I sat down and waited for my car. But I wasn’t given my gig.
* 자모 단위: I applied for the choir. So I got my voice straightened. I sat down and waited for my turn. However, my chance was not given.

1. 결론

본 논문은 한국어의 특성을 반영하지 못하여 발생하는 번역 성능의 하락을 해결하기 위해 자모 단위 변환과 높임말/반말의 변환을 제안했다. Transformer 모델 기반 학습을 통한 번역 성능을 확인하였을 때 정량적으로는 BLEU 점수가 유사하거나 소폭 향상됐고, 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 보여, 향후 추가적인 연구가 필요할 것으로 판단된다.

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, and I. Polosukhin (2017). “Attention Is All You Need,” *arXiv preprint arXiv:1706.03762*.
5. https://tech.kakaoenterprise.com/22
6. R. Sennrich, B. Haddow, and A. Birch (2016). “Neural Machine Translation of Rare Words with Subword Units,”*54th Annual Meeting of the Association for Computational Linguistics.*
7. C. Lee, D. Lee, Y. Hur, K. Yang, and H. Lim (2018). “Comparing Byte Pair Encoding Methods for Korean,” The 30th Annual Conference on Human and Language Technology, p. 293–294.
8. C. Kang, Y. Ro, J. Kim and H. Choi (2018). “Symbolizing Numbers to Improve Neural Machine Translation” JDCS, Vol. 19, No. 6, pp.1161-1167.
9. B. Kim, H. Kim, S.-W. Lee, G. Lee, D. Kwak, D. H. Jeon, S. Park, S. Kim, S. Kim, D. Seo, H. Lee, M. Jeong, S. Lee, M. Kim, S. H. Ko, S. Kim, T. Park, J. Kim, S. Kang, N.-H. Ryu, K. M. Yoo, M. Chang, S. Suh, S. In, J. Park, K. Kim, H. Kim, J. Jeong, Y. G. Yeo, D. Ham, D. Park, M. Y. Lee, J. Kang, I. Kang, J.-W. Ha, W. Park, 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.
10. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova (2019). “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” NAACL.
11. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever (2018). "Improving Language Understanding with Unsupervised Learning," *Technical report, OpenAI*.
12. S. Eo, C. Park, H. Moon, and 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,*Vol. 12, No. 3, p.1-7.
13. H. Kim (2015). “A new look at the United States from Korea, and Korea from the United States,” Sae-Ga-Jeong, Vol.62, p.45 – 19.
14. https://www.korean.go.kr/hangeul/principle/001.html
15. G. Han, (2002). “A Study on the Honorific speech of Modern Korean Language”, Yeongnak.
16. H. Choi. (2016).  “Korean morphology”. Yeongnak.
17. S. Kang, and Y. Kim (1992). “A Computational Analysis Model of Irregular Verbs in Korean Morphological Analyzer,” *Korean Computer Science*, Vol. 19, No. 2, p. 151-164.
18. J. Oh, and K.-S. Choi, (2005). “An English-to-Korean Transliteration Model based on Grapheme and Phoneme,” KIISE, Vol. 32, no. 4, pp.312-326.
19. K. Park, J. Lee, S. Jang, and D. Jung (2020). “An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks,” In Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 133–142, Suzhou, China.
20. J.-N. Jeong, S.-Y. Kim, S.-T. Kim, J.-J. Lee, and Y. Jung, (2021). “Implementation of Korean Honorific Converter Using OpenNMT,” Korea Society of Computer Information.
21. S. Lee and S. Kim, (2018). “Pre-editing Rules to Enhance Output Quality of Machine Translation: English-Korean and Korean-English,” Korean Association for Translation Studies.
22. M.-Y. Ahn (2020). “MT Problems and its MTPE Suggestion with Regard to Structural Differences between English and Korean,” Mirae English Language and Literature Association.
23. M. Kim, J. Nam, H. Jung, and H. Choi (2021). “Kor-Eng NMT using Symbolization of Proper Nouns” KIISE, Vol. 48, No. 10, pp.1084-1089.

|  |
| --- |
| 김 정 희 |
|  |
| 2017~ 현재: 한동대학교 전산전자공학부 재학 |
| 관심분야: 머신러닝, 딥러닝, 인공지능 |
| 허 재 무    2017~ 현재: 한동대학교 전산전자공학부 재학  관심분야: 머신러닝, 딥러닝, 인공지능 |
| 김 주 환 |
|  |
| 2017~ 현재: 한동대학교 전산전자공학부 재학 |
| 관심분야: 머신러닝, 딥러닝, 인공지능 |
|  |
| 최 희 열 |
|  |
| 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 |