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 정밀한 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, 신경 기계 학습 기반의 번역 모델과 방대한 양의 데이터를 사용해 뛰어난 한영 번역 성능을 보여준다. 하지만 ‘존댓말, 반말 간의 구분’, ‘자모단위의 구성’이라는 한국어의 특성을 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 모델 뿐만 아니라 어체 변환을 위한 별도의 모델이 필요하다(이거 확인해봐야됨). 이로 인해 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 특성을 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. 이는 일관성 있는 번역을 가능하게 하며, 추가적인 번역 모델의 학습없이 적은 비용으로 효율적인 높임말, 반말의 변환을 가능하게 한다. 또한 기존의 문장을 [ 안녕하세요. ㅇㅏㄴㄴㅕㅇㅎㅏ\_ㅅㅔ\_ㅇㅛ\_. ]와 같이 자모단위로 분리를 한 후 BPE를 진행한다. 이로 인해 더 정밀한 BPE의 생성이 가능하며, vocabulary의 크기도 상당히 줄어드는 것을 확인하였다.

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.

1. 배경지식
   1. Transformer

입력 문장의 문맥을 출력 문장에 반영하는 것으로 자연어 번역의 성능을 향상 시킬 수 있다. 문장의 문맥은 이를 구성하는 단어들 간의 관계를 통해 파악할 수 있다. 때문에 초기 RNN(Recurrent Neural Networks)을 활용한 자연어 번역 모델이 등장한 후로 LSTM(Long Short Term Memory)등을 추가한 모델도 등장하였다. 하지만 이런 시도에도 불구하고 RNN은 문장의 단어들이 순서대로 입력되기 때문에 문장 내 거리가 먼 단어 간의 관계를 모델링 하는 것이 힘들며, 문장이 길어질수록 학습하는 데 걸리는 시간이 길어진다는 한계가 있다. 이를 개선한 것이 Transformer 모델이며, 이는 단어 간 의 관계를 모델링하는 Attention Mechanism만을 사용한다. 또한 이전의 모델들보다 더 좋은 번역 성능을 보여주며, 현재 번역 모델의 대부분을 차지한다.

본 논문의 Transformer모델의 특성은 다음과 같다. 첫째, 문장 전체가 한번에 입력되므로 병렬 계산이 가능하다. 즉, RNN을 학습 시키는 데 필요한 시간보다 더 적은 시간을 활용해서 모델을 학습 시키는 것이 가능하다. 또한 Attention Mechanism만을 사용하여 구성했기 때문에 문장 내 단어 간의 거리에 따른 관계를 모델링하는 것이 효과적이다. 해외의 BERT, GPT 그리고 국내의 카카오 번역과 네이버의 Hyper Clova 등이 Transformer를 사용 중이다.

* 1. BPE

NMT(Neural Machine Translation)는 Vocabulary size에 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 완화하기위해 문장의 단어들을 더 작은 Subword 단위로 분할하여 사용하는 Subword Tokenization이 나오게 되었다. 본 논문에서는 BPE(Byte Pair Encoding)라는 대표적인 Subword 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 분할하여 사전 생성

4. 가장 많이 등장하는 문자 쌍을 찾아 병합하고, 병합한 쌍을 사전에 추가하는 행위를 hyper-parameter로 설정한 횟수만큼 반복

5. 사전을 사용해서 vocabulary 생성

3번과 4번을 실행한 결과는 다음과 같다.

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

}

또한 hyper-parameter를 10으로 설정을 하고 input으로 ‘한동대학교’를 넣은 결과는 다음과 같다.

한동대학교

한 동 대학 교 </w>

한 동 대학교 </w>

한 동 대학교</w>

한동 대학교</w>

위의 예시처럼 ‘한동대학교’는 ‘한동’과 ‘대학교’의 두 subword로 분할된다. 이처럼 BPE를 사용하면 사전에 등장하지 않는 단어가 input으로 들어오더라도 subword의 조합으로 설명이 가능해진다.

* 1. 한국어의 특성

한국어와 영어는 각각의 언어적 특성들을 가지고 있다. 이 특성들은 서로 공통된 부분도 있으나, 다른 문화권의 언어이기에 가지는 차이점들이 있다. 그중 한국어에서 나타나는 특징들은 다음과 같다. 첫째, 한국어는 영어와는 달리 각 음절이 초성, 중성 그리고 종성으로 이루어진 자모 단위로 구성되어 있다. 둘째, 한국어는 문법적인 존대법 체계를 가지고 있지만 영어는 문법적인 수준의 격식, 존대 표현을 가지고 있지 않다.

* + 1. 자모 단위 구성

단어: 한글 → 음절: 한,글 → 음소: ㅎ,ㅏ,ㄴ,ㄱ,ㅡ,ㄹ

위의 예시처럼 한국어는 자음과 모음을 나타내는 글자가 따로 존재하는 음소 문자이며, 자음과 모음을 합쳐서 음절 단위로 사용하는 음절 문자적 성격을 지니고 있다. 이는 일반적으로 한국어는 음절 단위로 사용되며, 음절을 자음과 모음, 즉 음소로 나누어질 수 있음을 의미한다. 한국어는 ~개의 자음 ~개의 모음으로 구성 되어 있다.

* + 1. 형태소

문장: 재무가 이야기 책을 읽었다.

어절: 재무가/이야기/책을/읽었다.

형태소: 재무/가/이야기/책/을/읽/었/다.

한국어는 의미를 가진 최소 단위인 형태소로 이루어져 있다. 자립성의 유무에 따라 자립 형태소와 의존 형태소로, 실질 의미 유무에 따라 실질 형태소와 형식 형태소로 구분할 수 있다.

자립 형태소: 재무가 이야기 책을 읽었다.

의존 형태소: 재무가 이야기 책을 읽 었 다.

위의 그림은 의존 형태소와 자립형태소의 예시들이다. 의존 형태소는 그 자체로 자립성이 없고 다른 말에 의존해야 하는 형태소로, 주로 조사나, 어미, 접사, 용언의 어근 등이 해당하며 자립형태소는 그 자체로 자립성이 있는 말로, 주로 체언의 어근, 감탄서, 관형사, 부사 등이 해당된다.

* + 1. 종결어미

한국어는 청자에 대한 화자의 태도에 따라 여러 문체가 존재하며 문장의 종결유형 별로 이를 구분한 문체법이 존재한다. 문장의 목적에 따라 평서법, 의문법, 명령법, 청유법, 감탄법으로 나뉘며 청자 또는 높임의 정도에  따라 종결어미에서 하십시오, 하오, 하게, 해라, 해요, 해체로 나뉘며 그 중에서 해요체와 해체의 경우 비격식체, 나머지는 격식체로 분류된다.

1. 제안 모델

본 논문은 한국어의 특수성을 반영한 한영 번역 성능의 향상을 위해 자모 단위 번역과 높임말, 반말의 변환을 제안한다. 이를 구현하기 위해 배경지식에서 설명한 Transformer 모델을 사용했으며, AI-Hub에서 제공하는 한국어-영어 말뭉치 데이터 160만개와 ~~~한 방법으로 생성한 HGU\_clean 데이터 ~~~만개를 사용했다. 학습을 위해 데이터를 가공하는 전처리 과정을 거치게 되며, 이는 크게 4가지(Split, Tokenize, Symbolize, BPE)이다. 먼저 Train, Valid, Test data로 나눠주는 Split을 진행한다. 이후 ‘ . ’, ’ ! ’, ‘ ? ‘와 같은 문장 부호를 기준으로 문장을 띄워주는 Tokenize을 진행한다. 그리고 고유명사나 우리가 적용하고자 하는 symbol을 처리해주는 Symbolize를 진행한 후, 마지막으로 BPE를 생성 및 적용하게 된다. 자모 변환의 경우 Symbolize 이후에 진행되며 높임말, 반말 변환의 경우 Split 전에 진행된다.

3.1 자모 단위 변환

본 논문은 자모 단위에 BPE를 적용하는 것이 목적이기 때문에 Symbolize 와 BPE단계 사이에 자모 단위 변환을 진행하였다. 자모 단위 변환은 오로지 한국어에만 작용하며 초성, 중성, 종성의 구조로 이루어진다. 또한 ‘나’와 같이 종성이 없는 경우는 ‘\_’로 대체하게 된다. 자모 단위 변환을 적용할 경우 ‘안녕하세요. Kim입니다.’라는 문장은 다음과 같이 ‘ㅇㅏㄴㄴㅕㅇㅎㅏ\_ㅅㅔ\_ㅇㅛ\_. Kimㅇㅣㅂㄴㅣ\_ㄷㅏ\_.’로 변환된다. 이러한 변환을 통한 BPE생성은 2가지의 이점을 가지고 있다. 바로 vocabulary size의 감소와 정밀한 BPE적용이다.

3.1.1 vocabulary size의 감소

본 논문의 자모 단위 변환 후 BPE의 생성 및 적용을 하게 되면 vocabulary size가 줄어드는 것을 확인하였다. 기존의 경우 subword 단위로 BPE를 생성하게 되고, 이들 중 빈도수가 가장 높은 12000개의 subword의 집합을 vocabulary로 사용하게 된다. 이때 총 16340개의 unique한 단어가 나오게 되고, 이 중 4340개의 단어는 표현할 수(숫자로 나타낼 수) 없게 된다. 그에 반해 자모 단위로 변환하여 적용한 경우 13068개의 unique한 단어가 나오게 되는 것을 확인하였다. 즉 동일한 vocabulary size에 대해 자모 단위가 더 많은 단어를 표현 할 수 있다는 의미이다.

3.2.2 정밀한 BPE의 생성

기존의 BPE는 subword 단위로 BPE를 생성하게 된다. 즉 ‘안녕’이라는 단어의 경우 ‘안’, ’녕’과 같이 분리하여 BPE를 진행한다. 그에 반에 자모 단위 BPE는 ‘ㅇ’, ‘ㅏ’, ‘ㄴ’, ‘ㄴ’, ‘ㅕ’, ‘ㅇ’과 같이 더 정밀하게 나뉘어 BPE를 진행한다. 이는 ‘영희’에 조사 ‘ㄴ’이 붙은 경우 ‘ㄴ’을 따로 분리하지 못한 채 ‘영흰’ 또는 ‘영’, ‘흰’으로 나뉘는 상황에서 ‘ㄴ’을 분리하여 ‘영희’, ‘ㄴ’으로 나뉘어 분석할 수 있음을 의미한다. 아래 는 정밀하게 BPE가 적용되어 더 높은 성능을 보이는 예시들이다.

Subword BPE

BPE : 우리 오@@ 빤 용@@ 돈을 많이 준다 .

Translated : We give a lot of poc@@ ket money .

Output : We give a lot of pocket money.

자모단위 BPE

BPE : ㅇㅜ\_ㄹㅣ\_ ㅇㅗ\_@@ ㅃ@@ ㅏㄴ ㅇㅛㅇ@@ ㄷㅗㄴㅇㅡㄹ ㅁㅏㄶㅇㅣ\_ ㅈㅜㄴㄷㅏ\_ .

Translated : My brother gives me a lot of allowance .

Output : My brother gives me a lot of allowance.

이 예시는 오빠에 붙은 조사 ‘ㄴ’을 완벽하게 분리하지는 않지만 subword BPE와는 다른 분리를 통해 더 정확하게 번역하는 것을 확인 할 수 있다.

Subword BPE

Bpe : 우리는 영@@ 흴 좋아@@ 한다 .

Translated : We like Young-@@ S@@ om .

Output : We like Young-Som.

자모단위 BPE

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

Translated : We like Youn@@ gh@@ ee .

Output : We like Younghee.

Subword BPE

Bpe : 교@@ 횐 지금 축@@ 제의 분위기@@ 다 .

Translated : Ky@@ o@@ to , the atmosphere of the festival is now .

Output : Kyoto, the atmosphere of the festival is now.

자모단위 BPE

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

Translated : The church is in the mood of the festival now .

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

이 예시들을 통해 본 논문을 통해 처음 예상했던 것 처럼 영희에 붙은 조사 ‘ㄹ’과 교회에 붙은 조사 ‘ㄴ’을 제대로 분리한 것을 확인 할 수 있다.

|  |  |  |
| --- | --- | --- |
|  | 격식체 | 비격식체 |
| 아주 높임 | 하십시오체 | 해요 |
| 예사 높임 | 하오체 |
| 예사 낮춤 | 하게체 | 해 |
| 아주 낮춤 | 해라체 |

본 논문에서 높임말 반말 변환 시 상대높임법을 고려했으며, 종결어미 간 변환은 격식체와 비격식체 끼리 아주 높임에서 아주 낮춤으로 변환되도록 하였다. 격식체의 경우 높임 표현과 낮춤 표현에는 두 어체가 존재하나, 최근 사용이 줄고 있는 하오체와 하게체 대신 하십시오체와 해라체로 변환되도록 했다.[[1]](#footnote-2)

반말로 변환하는 경우 높임말로 변환하는 경우와 달리, 반말로 변환할 때에는 종결어미의 경우 용언의 활용, 품사 등을 함께 고려해서 변경해야 한다. 즉, 종결어미의 종류에 따라 1:1로 변경되지 못하는 경우가 존재하게 된다. 예를 들어, ‘주세요’라는 문장은 반말로 변경 시 어간의 종성 ‘ㅜ’가 탈락하지 않지만, ‘푸세요’라는 문장은 원형이 ‘푸다’인 동사가 활용이 되어 종성 ‘ㅜ’이 탈락하는 불규칙 활용이다.[[2]](#footnote-3) 이처럼 높임말에서 반말로 변경할 때 문법적 요소를 고려해야 하며, 본 논문에서는 이를 6가지 예외적인 경우로 분류하고 처리하도록 하였다.

|  |  |  |
| --- | --- | --- |
| **예외** | **예외 경우** | **이유** |
| 1 | ‘-아/어’ 종결어미 | ‘어’는 끝 음절의 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 |

위의 규칙에 따라 변환 시 해당 형태소를 확인 하면서 예외적인 경우는 높임 종결어미를 반말 종결어미로 변경하고, 그 외에는 1:1로 높임 종결어미가 반말 종결어미로 변환되도록 하였다. 특히, 예외 규칙은 독립적으로만 사용되는 것이 아닌, 문법적 규칙에 따라 결합되어 적용되기도 한다. 종결어미 외에도 감탄사인 ‘예’, ‘아니요’도 알맞는 낮춤 표현으로 변경되도록 하였다.

1. 한국어 형태론 p 123 [↑](#footnote-ref-2)
2. 한국어 형태론 P189 [↑](#footnote-ref-3)