드림학기 연구 논문
지도교수 홍우평

다중 토큰화 전략을 적용한 한국어-영어 기계 번역

2020년 12월

건국대학교
문과대학 영어영문학과(휴먼ICT연계전공)
박 도 준

목차

표목차 ii
그림목차 iii
국문초록 iv

제1장 서론 1
제1절 연구 목표 1
제2절 연구 배경 2
제3절 논문의 구성 4

제2장 연구 방안 5
제1절 실험 데이터 5
제2절 트랜스포머 신경망 6
제3절 토큰화 적용 방법의 다양화 8

제3장 연구 결과 분석 10
제1절 실험 환경 10
제2절 정확도 및 혼잡도 분석 10
제3절 BLEU 점수 분석 14
제4절 번역 문장 분석 16

제4장 결론 및 한계점 20

참고문헌 22
ABSTRACT 23

표 목차

<표 1> 한-영 뉴스 병렬 코퍼스 예시 5
<표 2> 한국어와 영어 문장 토큰화 방법의 예시 9
<표 3> epoch 변화에 따른 각 모델의 BLEU 점수표 14
<표 4> 원문과 기계 번역 문장 예시 1 16
<표 5> 원문과 기계 번역 문장 예시 2 17
<표 6> 원문과 기계 번역 문장 예시 3 19

그림 목차

<그림 1> 트랜스포머 모델 구조 7
<그림 2> 한국어 자모 토큰화를 적용한 모델의 정확도와 혼잡도 11
<그림 3> 한국어 형태소 토큰화를 적용한 모델의 정확도와 혼잡도 12
<그림 4> 한국어 BPE 토큰화를 적용한 모델의 정확도와 혼잡도 13

국문 초록

다중 토큰화 전략을 적용한 한국어-영어 기계 번역

본 연구는 토큰화 방법이 기계 번역 모델의 학습 결과에 어떠한 영향을 미치는지 알아보기 위해 진행되었다. 실험을 위해 OpenNMT-py 라이브러리를 이용하여 동일한 하이퍼 파라미터를 설정한 트랜스포머 신경망을 구축하였고, 출발어인 한국어와 도착어인 영어에 대해 각각 자모 토큰화, 형태소 토큰화, BPE 토큰화를 적용하여 9개의 모델을 5만 epoch을 반복하며 비교 실험을 진행하였다. 모델의 학습, 검증, 평가 단계에서 AI Hub로 제공받은 80만 쌍의 한영 뉴스 코퍼스를 사용하였다.

학습된 모델의 BLEU 점수를 측정하여 모델의 성능 평가를 진행한 결과 한국어에 BPE 토큰화를 적용하고 영어에 형태소 토큰화를 적용한 모델이 35.73을 기록하며 가장 우수한 성능을 보였고, 한국어와 영어에 모두 BPE 토큰화를 적용한 모델이 23.54을 보이며 다음으로 높은 성능을 보였다. 그 외의 모델은 BLEU 점수가 1이 채 되지 않는 낮은 점수를 보였고, 부진한 모델 학습의 경과는 검증 데이터의 정확도와 혼잡도를 통해서도 확인이 가능했다.

각 학습 모델의 번역문과 원문을 비교한 결과 한국어 BPE 토큰화, 영어 형태소 토큰화 모델이 가장 원문에 충실한 번역을 보여주었으며, 한국어와 영어에 BPE 토큰화를 적용한 모델은 학습 데이터에 포함되지 않은 명사나 고유 명사가 포함된 문장의 번역에서 우수한 성능을 보여주었음을 확인해 볼 수 있었다.

제1장 서론

제1절 연구 목표

본 연구는 다중 토큰화(tokenization) 전략을 적용하여 기계 번역 모델의 성능 차이를 분석하고자 한다. 토큰화란 특정 기준에 따라 입력 문장을 적절한 단위인 토큰(token)으로 분절하는 작업을 의미하며, 토큰화 과정에서 형성된 토큰은 단어 벡터로 변환되면서 컴퓨터가 이해할 수 있는 수치화 된 의미를 형성한다. 따라서 기계 번역 과제를 수행함에 앞서 어떠한 기준으로 입력 문장에 대한 토큰을 구성할 지에 대한 고민은 신경망 구조의 설계 만큼이나 중요한 문제라 말할 수 있다.

토큰화는 어떤 언어를 대상으로 하는지에 따라 다르게 고려되어야 한다. 이는 언어마다 사용되는 문자의 종류와 표기 방법이 상이하기 때문이다. 본 연구에서 다루고자 하는 한국어와 영어는 각기 다른 문자 체계를 갖는다. 한국어는 24자의 기본 자모로 구성된 한글을 사용하고, 영어는 26자의 라틴 문자를 사용하며 대문자와 소문자를 구분한다. 또한 한글은 하나의 음절을 구성하는 자모들이 모여 독립된 문자를 형성하는 모아쓰기 규칙을 따르지만, 영어는 자모가 가로 순으로 나열되어 별도의 음절의 구분을 표기하지 않는다.

본 연구는 한국어와 영어가 가진 고유한 문자적, 언어적 특성을 반영하기 위한 세 가지 토큰화 방법을 적용하였다. 첫째는 낱개의 자모를 개별적 토큰으로 분리하는 '자모 토큰화'이며, 둘째는 언어학적 의미를 형성하는 최소 단위인 형태소로 분리하는 '형태소 토큰화', 셋째는 자주 함께 출현하는 문자 쌍을 새로운 토큰으로 학습하여 분리하는 'BPE(Byte Pair Encoding) 토큰화'이다. 위 세 가지 토큰화 방법을 한국어와 영어 데이터에 적용하여 별도의 모델 학습을 진행하고, 각 모델의 성능 차이를 비교하고자 한다.

모델 학습에는 기계 번역 과제에 획기적인 성능 개선을 가져온 트랜스포머

(Transformer) 신경망을 이용하며, 번역한 텍스트를 정량적 수치로 평가하기 위해 기계 번역 텍스트 평가 지표인 BLEU(Bilingual Evaluation Understudy) 점수를 사용하도록 한다.

제2절 연구 배경

오늘날의 인공지능은 딥러닝(Deep Learning)을 기반을 비약적인 성장을 이뤄내고 있다. 생물학의 신경망에서 아이디어를 착안한 통계학적 학습 알고리즘인 인공신경망(Artificial Neural Network)은 신체 자극 신호를 전달하는 신경 세포인 뉴런(neuron)을 입력 값에 대한 최적의 파라미터(parameter)를 찾아내는 퍼셉트론(perceptron)으로 구현해 냄으로써 시작되었다. 이후 기울기 소실 문제, 과적합 문제 등 여러 난관에 부딪히며 관련 연구는 침체되는 듯 보였으나, 컴퓨팅 하드웨어 성능의 발전과 다양한 학습 알고리즘 및 신경망 설계 기술의 발전을 토대로 오늘날의 딥러닝은 컴퓨터 비전, 음성 인식, 자연어처리 등 다양한 분야에 적용되며 전래 없던 성과를 보여주고 있다.

본 연구에서 다루고자 하는 기계 번역은 자연어처리(Natural Language Processing)에 포함되는 하위 연구 분야이다. 컴퓨터를 이용한 번역은 1940년대부터 시도되었는데 규칙 기반, 통계 기반의 방식을 거쳐 현재의 딥러닝 기반의 기계 번역으로 발전해오고 있다. 구글의 기계 번역 서비스는 '구글 번역'이라는 이름 하에 서비스가 제공되고 있으며, 2016년 딥러닝 기반의 기계 번역 시스템을 도입한 이래로 번역 품질이 크게 향상된 바 있다. 국내 기업 중에서는 네이버의 '파파고', 카카오의 '카카오 i 번역', 삼성전자의 'S번역기' 등이 딥러닝 기반의 번역 서비스를 제공하고 있다.

2017년까지 기계 번역의 주된 흐름은 순차적 정보처리였다. 현재까지도 여러 분야에서 널리 사용되고 있는 순환신경망(Recurrent Neural Network)은 시계열 정보를 포함하고 있기에 과거의 결과를 기반으로 현재의 상태를 예측하는 데 사

용되었고, 시간적 특성을 가진 언어 데이터에 접목되며 이후 LSTM(Long Short-Term Memory), GRU(Gated Recurrent Units)와 같은 신경망으로 발전되어 사용되었다. 그러나 2017년 구글 소속 연구원들이 중심이 되어 발표한 논문인 'Attention Is All You Need'(Vaswani el al, 2017)는 기존의 자연어처리의 판도를 혁신적으로 바꾼 전환점이 되었다. 이 논문에서 제시된 트랜스포머 신경망은 기존의 시계열 접근 방식의 데이터 처리의 대안으로 모든 문장을 벡터의 집합인 행렬 단위로 처리하는 방법을 제시하였고, 이에 따라 직렬 연산으로 오랜 학습 시간이 소요되는 문제를 병렬 처리라는 대안으로 빠르고 능률적으로 처리할 수 있게 되었다. 현재까지 자연어이해(Natural Language Understanding)와 자연어생성(Natural Language Generation) 분야에서 선두를 달리고 있는 BERT(Bidirectional Encoder Representations from Transformers)와 GPT(Generative Pre-trained Transformer) 모두 트랜스포머 신경망에 기반을 둔 모델들이다.

자연어처리 과제의 수행 결과에 상당한 영향을 미치는 요인으로 토큰화 방법을 꼽을 수 있다. 토큰화(tokenization)란 주어진 코퍼스(corpus)를 의미를 부여할 수 있는 단위인 토큰(token)으로 나누는 작업을 일컬으며, 토큰은 구분 기준에 따라 여러가지 다른 형태로 나눠질 수 있다. 토큰화 방법이 중요한 이유는 토큰에 따라 단어의 의미를 결정하는 단어 벡터(word vector)가 형성되기 때문이다. 입력된 문장에 대해 변환된 일련의 토큰들은 딥러닝 모델로 전달되기 전 수치화된 N-차원의 벡터로 변환되고, 모델 학습이 진행됨에 따라 단어 벡터 값이 갱신되며 최적화된 의미를 찾아 나간다. 따라서 토큰화 적용 방법은 자연어처리 성능에 상당한 영향을 미치는 요인으로 간주된다.

제3절 논문의 구성

본 논문의 구성은 다음과 같다.

먼저 2장에서는 구체적인 연구 방안에 대해 설명한다. 2장 1절에서는 연구에 사용된 실험 데이터에 대해 소개하고, 모델 학습과 검증 및 평가를 위해 데이터를 어떻게 분할하였는지, 그리고 데이터의 예시와 함께 구성 구조 및 기본적 데이터 특징에 대해 기술한다. 이어 2장 2절에서는 본 연구를 위해 설계한 트랜스포머 신경망의 특징과 내부 구조에 대해 설명하고 실험을 위해 동일하게 설정된 하이퍼 파라미터(hyper parameter) 대해 설명한다. 이후 2장 3절에서는 토큰화 방법과 데이터가 작성된 언어의 관련성에 대해 언급하며 본 연구에서 다루는 한국어와 영어의 문자적, 언어적 특성을 반영하기 위한 다중 토큰화 전략을 설명하고 토큰화가 각 적용 방법마다 어떻게 처리가 되는지를 예시를 통해 보여주고자 한다. 그리고 모델의 디코더에서 생성된 문장에 어떠한 후처리 작업이 필요로 하는 지를 설명하고자 한다.

3장에서는 실험 설계에 따라 진행된 연구에 대한 결과를 분석한다. 3장 1절에서는 연구가 수행된 실험 환경에 대해 기술한다. 다음으로 3장 2절에서는 모델 학습 과정에서 학습 데이터와 검증 데이터로 측정된 정확도와 혼잡도를 기반으로 모델의 학습 진행 경과를 설명하고, 이를 그래프로 시각화 하여 총 5만 번의 epoch(반복 학습 횟수)에서 보이는 측정 값을 지표로 삼아 모델의 학습 과정을 평가한다. 이어지는 3장 3절에서는 1만 epoch 마다 저장된 모델과 모델 학습에 사용되지 않은 평가 데이터를 이용하여 번역을 수행하고 번역 결과에 대한 BLEU 점수를 측정하여 각 모델의 성능을 분석한다. 이후 3장 4절에서는 모델이 번역한 문장들을 제시하며 각 학습 모델이 출력한 번역 결과에 대해 어떠한 특징이 관찰되는 지 분석하고 번역 품질에 대한 평가를 진행한다.

끝으로 4장에서는 실험에 대한 전체적인 결과를 종합하여 결론을 도출하고 향후 연구에 대한 방향을 제시한다. 또한 실험을 진행하며 겪은 현실적 제약 조건 및 본 연구가 가지는 한계점을 설명하고 이에 대한 개선 방안을 제안한다.

제2장 연구 방안

제1절 실험 데이터

본 연구에서는 모델 학습을 위한 데이터로 AI Hub에서 제공하는 한-영 병렬 코퍼스를 사용하였다. AI Hub는 한국정보화진흥원이 운영하는 AI 통합 플랫폼으로 2017년부터 AI 기술 및 서비스 개발에 필요한 다양한 AI 학습용 데이터를 제공하고 있다. 현재 AI Hub는 총 160만 문장의 한-영 병렬 코퍼스를 제공하고 있으며, 본 연구에서는 80만 문장 쌍으로 구성된 한-영 뉴스 코퍼스를 사용하였다. 뉴스 코퍼스를 선별적으로 사용한 이유는 뉴스 기사에서 사용된 문체의 일관성과 정제된 어휘와 문장이 번역 모델 학습에 적합한 조건을 충족하고 있다고 판단하였기 때문이다. 본 연구에 사용된 병렬 코퍼스의 예시는 아래와 같다.

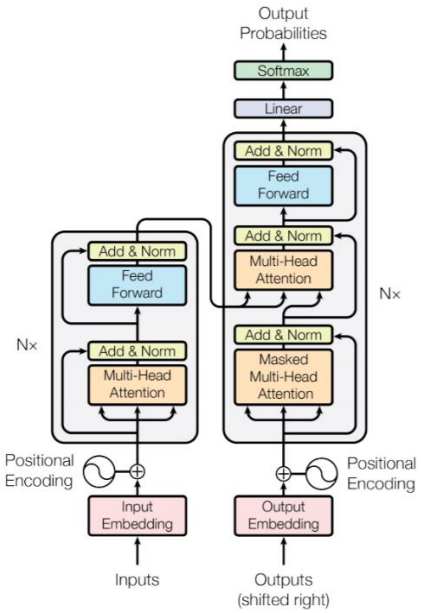
한국어	영어
스키너가 말한 보상은 대부분 눈으로 볼 수 있는 현물이다.	Skinner's reward is mostly eye-watering.
심지어 어떤 문제가 발생할 건지도 어느 정도 예측이 가능하다.	Even some problems can be predicted.
오직 하나님만이 그 이유를 제대로 알 수 있을 겁니다.	Only God will exactly know why.
중국의 논쟁을 보며 간과해선 안 될 게 기업들의 고충이다.	Businesses should not overlook China's dispute.
박자가 느린 노래는 오랜 시간이 지나 뜨는 경우가 있다.	Slow-beating songs often float over time.
보험 처리가 안 되는 비급여 시술은 엄두도 못 낸다.	I can't even consider uninsured treatments.
예수까지 합치면 모두 열세 명이 함께 식사를 하는 것이다.	Including Jesus, thirteen people eat together.

<표 1> 한-영 뉴스 병렬 코퍼스 예시

전체 80만 문장으로 구성된 한-영 뉴스 코퍼스는 98:1:1의 비율로 무작위 선별하여 학습 데이터(78만 4천 문장), 검증 데이터(1천 문장), 평가 데이터(1천 문장)로 구분하였고, 분할된 데이터를 사용하여 모델 학습, 모델 검증, 그리고 모델 평가를 진행하였다. 본 뉴스 데이터의 평균 문장 길이(문자 기준)는 한국어 69자, 영어 173자이고, 최단 문장은 한국어 15자, 영어 20자, 최장 문장은 한국어 220자, 영어 706자이다.

제2절 트랜스포머 신경망

트랜스포머 신경망(Transformer network)은 크게 문장을 입력 받는 인코더(encoder)와 출력 문장을 생성하는 디코더(decoder)로 구성된다. 트랜스포머 신경망의 주요 특징은 시계열 정보를 입력 받아 연산하는 순환신경망 구조에 의존하지 않고 주요 단어에 집중하는 어텐션(Attention) 기법으로 설계되었다는 점이다. 본 연구에서 인코더는 출발어에 해당하는 한국어 문장을 입력 받아 임베딩 과정을 거쳐 문장의 의미를 문맥 벡터(context vector)에 담아서 디코더로 전달한다. 이후 디코더는 전달받은 문맥 벡터와 도착어인 영어 문장을 입력 받아서 한국어 입력 문장에 해당하는 영어 문장을 생성하는 학습 훈련을 수행한다. 본 연구에 사용된 트랜스포머 신경망은 'Attention is all you need' 논문에서 제시한 것과 같이 인코더와 디코더를 여섯 층 씩 쌓아서 모델을 구성하였다.



<그림 1> 트랜스포머 모델 구조

모델의 학습 데이터로 사용된 78만 4천 문장의 한-영 뉴스 코퍼스는 임베딩 과정을 거쳐 각각 인코더와 디코더의 입력 값으로 모델에 전달된다. 하지만 트랜스포머 모델의 특성 상 순환신경망 구조를 배제하여 시계열 정보의 손실이 발생하게 되는데, 이 점을 보완하기 위해 단어의 위치 정보를 전달해주기 위한 방법으로 포지셔널 인코딩(Positional Encoding) 방식이 적용된다. 이는 각 단어 임베딩에 위치 정보를 반영하여 모델의 입력 데이터로 사용하는 방법이다.

인코더의 하위층(sublayer)은 멀티 헤드 어텐션(Multi-Head Attention)과 순방향 신경망(Feed Forward Neural Network)으로 구성되고, 디코더의 하위층은 마스크드 멀티 헤드 셀프 어텐션(Masked Multi-Head Self-Attention), 멀티 헤드 어텐션(Multi-Head Attention), 그리고 전방 전달 신경망(Feed Forward Neural Network)으로 구성된다. 인코더와 디코더의 하위층에서 입력 데이터에 대한 연산이 완료된 후 잔차 연결(Residual Connection)과 층 정규화(Layer Normalization)를 통해 연산 과정에서 손실되었을 수 있는 초기 정보를 다시 반영해주고 데이터의 값을 균일하게 보정하는 작업을 수행한다.

모델 설계에 적용된 하이퍼 파라미터(hyper parameter)는 다음과 같이 설정하

였다. 인코더와 디코더의 출력 차원을 512, 전방 전달 신경망의 층 수를 2048, 멀티 헤드 어텐션의 헤드(head)를 8, 드롭 아웃(drop out)을 0.1, 그리고 배치 사이즈(batch size)를 4096으로 통일하여 동일한 조건 하에 모델 학습이 이뤄지도록 하였다.

제3절 토큰화 적용 방법의 다양화

토큰화(tokenization)는 입력 문장에 대해 의미를 부여할 수 있는 최소 단위인 토큰으로 분절하는 작업으로, 자연어처리를 수행하기 전에 거치는 필수적인 절차이다. 어떠한 기준으로 토큰을 형성할 것인 지에 따라 다양한 토큰화의 방법을 적용할 수 있으며, 다뤄지는 언어의 문자적, 언어적 특성 역시 토큰화 방법을 선택하기 이전에 함께 고려되어 하는 사항이다.

기본적으로 언어학에서 구분하는 의미의 최소 단위인 형태소를 기준 단위로 삼아 토큰화 하는 방법이 일반적이다. 이와 유사하게 공백을 기준으로 토큰을 만들 수도 있지만, 한국어와 같이 체언에 조사가 붙는 언어에서는 조사가 분리되지 않은 채 토큰이 만들어지게 되며, 이는 동일한 단어도 별개의 단어로 인식하게 하여 자연어처리 학습에 부정적 영향을 미치는 원인이 된다. 그리고 한국어와 같이 자음과 모음이 결합하여 음절마다 새로운 문자가 조합되는 경우에는 자음과 모음을 분리하여 토큰을 분절하는 자모 단위의 토큰화도 적용 가능하다. 최근에는 BPE(Byte Pair Encoding) 알고리즘 방식을 적용하여 입력 문장의 문자 수준에서 가장 많이 출현하는 문자 쌍을 선정하고 이들을 병합하여 새로운 토큰으로 만드는 BPE 토큰화 방법도 많이 사용되고 있다. 이 방법은 단어 사전에 없는 토큰이 등장하였을 때 발생하는 OOV(Out Of Vocabulary) 문제를 최소화할 수 있다는 이점을 가진다. 본 연구에서는 기계 번역 모델의 출발어와 도착어로 사용되는 한국어와 영어에 적용할 토큰화 방안을 다음과 같이 제시한다.

토큰화 방법	언어	토큰화 문장 예시
자모 토큰화	한국어	ㅇ/ㅏ/ㄴ/ㄴ/ㅕ/ㅇ/ㅎ/ㅏ/ㅅ/ㅔ/ㅇ/ㅛ/.
	영어	n / i / c / e / ␣ / t / o / ␣ /m / e / e / t / ␣ / y / o / u / .
형태소 토큰화	한국어	안녕 / ␣ / 하세요 / .
	영어	nice / ␣ / to / ␣ / meet / ␣ / you / .
BPE 토큰화	한국어	안녕 / 하 / 세@@ / 요 / .
	영어	ni@@ / ce / to / me@@ / et / you / .

<표 2> 한국어와 영어 문장 토큰화 방법의 예시

우선 자모 토큰화와 형태소 토큰화 작업에서 토큰 분리 이전에 한국어와 영어의 띄어쓰기 규칙을 반영할 수 있도록 공백을 특수문자 ''␣"(U+00E6)로 치환하는 작업을 거쳤다. 한국어 자모 토큰화는 hgtk 라이브러리를 사용하여 모아쓰기 규칙에 따라 조합된 자음과 모음을 분리하여 분리된 자음과 모음 문자를 임베딩 토큰으로 사용하였으며, 영어 자모 토큰화의 경우 나열된 문자를 그대로 개별적 토큰으로 사용하였기에 별도의 라이브러리를 사용하지 않았다. 한국어 형태소 토큰화에는 Konlpy 패키지의 Okt 형태소 분석기를 사용하였고, 영어 토큰화에는 spacy 라이브러리의 영어 형태소 분석기를 사용하였다. BPE 토큰화의 경우 트랜스포머 모델 구축에 사용한 OpenNMT-py 라이브러리의 내장 모듈을 이용하여 빈도 수가 높은 동시 출현 문자를 학습하고, 이를 기반으로 입력 문장에 대한 BPE 토큰을 출력하도록 하였다. BPE 토큰화는 공백 없이 다음 토큰이 이어지는 토큰의 경우 "@@"(U+0040)을 붙이는 방법으로 토큰을 생성하여 띄어쓰기를 구분할 수 있도록 하였다. 이후 모델 평가 과정에서 띄어쓰기 규칙을 위해 출력 문장에 삽입된 특수문자를 후처리 작업을 거치며 모두 제거하도록 하였다.

토큰화를 거친 후 형성된 토큰의 개수는 다음과 같다. 자모 토큰화는 한국어 81개, 영어 50개, 형태소 토큰화는 한국어 692,838개, 영어 161,817개였으며, BPE 토큰화의 경우 한국어 37,015개, 영어 32,485개의 토큰을 형성하였다.

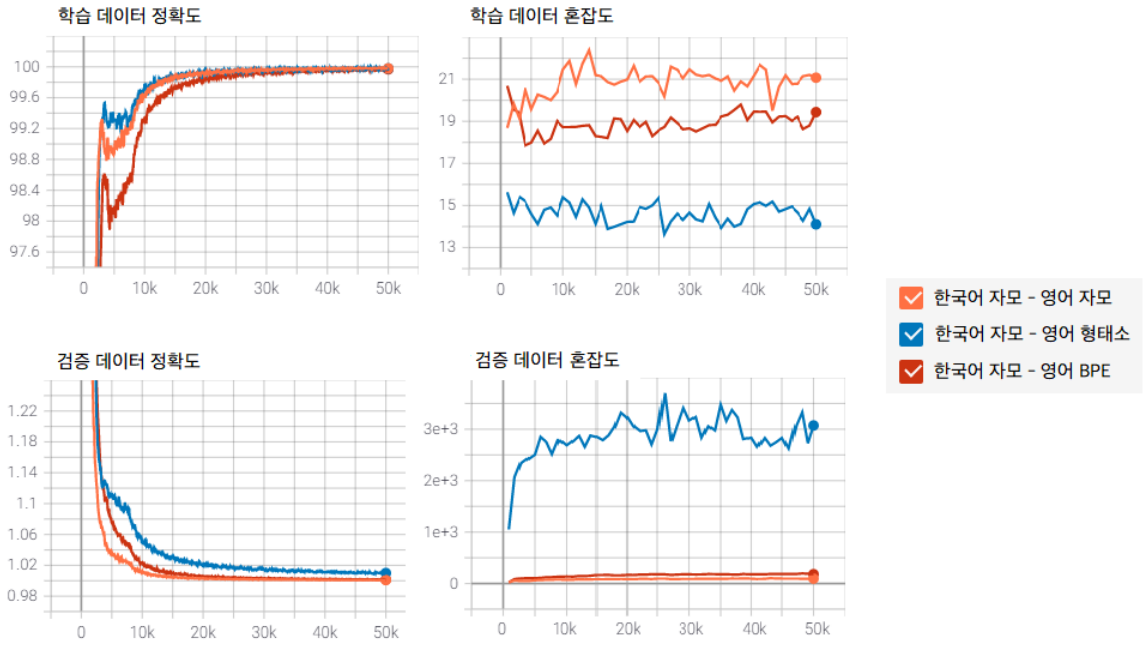
제3장 연구 결과 및 분석

제1절 실험 환경

본 연구의 한-영 기계번역 모델 학습은 정보통신산업진흥원(National IT Industry Promotion Agency)의 지원을 받아 리눅스(Linux) 기반의 운영체제인 우분투(Ubuntu) 원격 서버에서 진행되었다. 위 학습 환경에는 110GB의 램(RAM)과 RTX6000 GPU가 사용되었으며, 파이토치(PyTorch) 기반의 오픈 소스 신경망 기계 번역 시스템인 OpenNMT-py 라이브러리를 사용하여 트랜스포머 신경망의 기계 번역 모델을 설계 및 구축하였다. 모델 학습에는 Adam(Adaptive Moment estimation) Optimizer 최적화 함수를 사용하였다.

제2절 정확도 및 혼잡도 분석

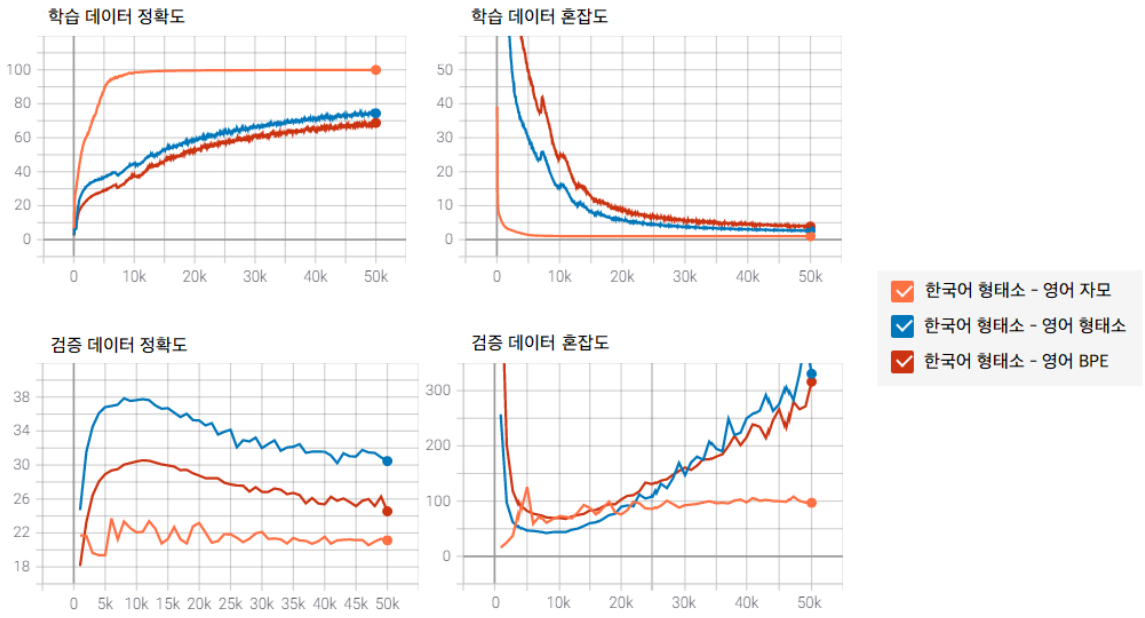
모델 학습이 진행되는 과정 중에 학습의 정도를 파악할 수 있도록 정확도(accuracy)와 혼잡도(perplexity)를 측정하였고, 이를 텐서보드(Tensorboard) 라이브러리를 이용하여 그래프로 시각화 하였다. 정확도는 입력 데이터 중에서 올바르게 예측한 데이터가 차지하는 비율을 측정하여 계산되며 100에 가까울수록 좋은 학습이 진행됨을 의미한다. 혼잡도는 예측 데이터 없이 모델 내부에서 성능을 측정하여 계산하는 내부평가 방식으로, 모델은 혼잡도를 최소화하는 방향으로 학습을 진행한다. 따라서 혼잡도가 0에 가까울수록 좋은 성능을 기대할 수 있다.



<그림 2> 한국어 자모 토큰화를 적용한 모델의 정확도와 혼잡도

위 그래프는 한국어에 자모 토큰화를 적용하고 영어에 자모, 형태소, BPE 토큰화를 적용한 세 개의 학습 모델에 대해 학습 데이터와 평가 데이터를 기반으로 측정된 정확도와 혼잡도를 보여준다. 위 그래프의 X축은 반복 학습 횟수인 epoch를 나타내고, 이에 대한 정확도와 혼잡도의 변화를 각각 Y축에서 표현하고 있다. 먼저 학습 데이터의 정확도 그래프를 보면 1만 epoch 이후 3개의 모델 모두 정확도가 99%를 넘어가고 있으며, 5만 epoch에서는 100에 근접함을 볼 수 있다. 하지만 학습에 포함되지 않은 검증 데이터로 정확도를 측정한 결과 5만 epoch 기준 1에 근접하는 수치로 측정되며 학습 데이터의 정확도와 상반된 결과를 보여준다. 이는 학습 데이터로 학습된 모델의 성능이 모델에 포함되지 않은 데이터에 대해서는 올바른 예측을 하지 못함을 의미한다.

학습 데이터로 측정된 혼잡도는 13~23 내에 분포하고 있는 반면, 평가 데이터에서는 영어 자모 모델이 94.34, 영어 BPE 모델이 181.4, 그리고 영어 형태소 모델은 눈에 띄게 높은 3072를 보이고 있다. 이 역시 학습 데이터를 기반으로 측정된 혼잡도와 반대되는 양상으로, 위 세 가지 모델의 학습 진행이 순조롭지 못함을 나타낸다.

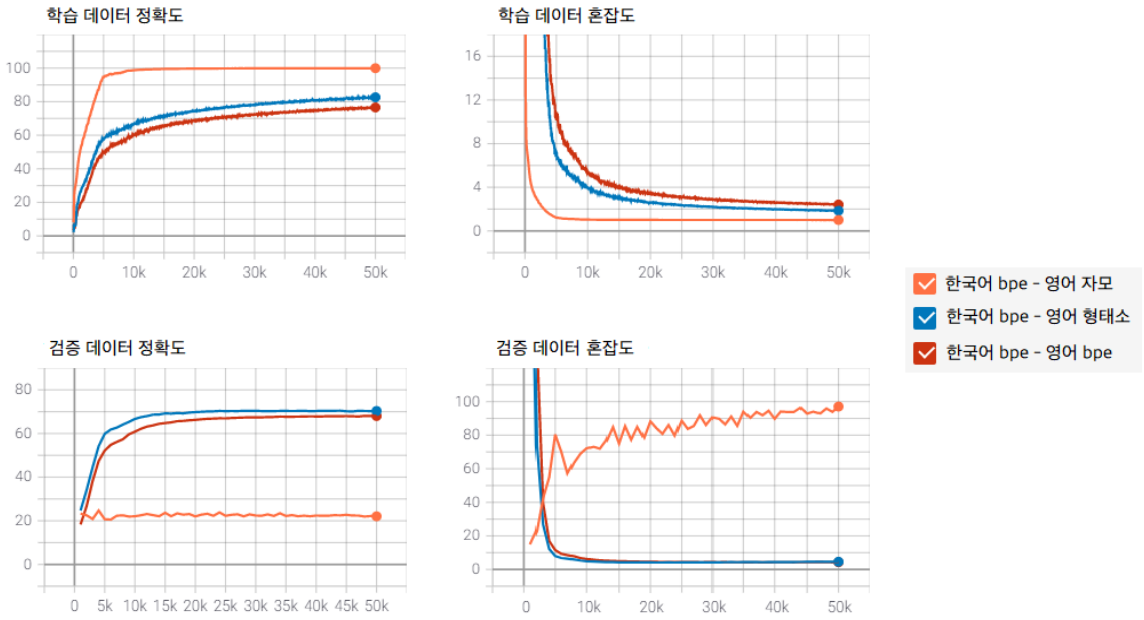


<그림 3> 한국어 형태소 토큰화를 적용한 모델의 정확도와 혼잡도

다음은 한국어에 형태소 토큰화를 적용하고 영어에 자모, 형태소, BPE 토큰화를 적용한 후 학습 데이터와 평가 데이터에 대해 모델의 정확도와 혼잡도를 측정한 그래프이다. 먼저 학습 데이터의 정확도는 영어 자모 토큰 모델이 1만 epoch 지점에서 98.02을 기록한 후 100에 수렴하는 모습을 보이고 있으며, 형태소, BPE 토큰화 모델의 경우 꾸준히 상승하는 모습을 보이며 5만 epoch에서 각각 74.37, 68.74를 보이고 있다. 반면 검증 데이터로 측정된 정확도는 형태소, BPE 토큰화 모델의 경우 1만 epoch까지 상승 곡선을 그려가다가 이후 정확도가 떨어지는 양상을 확인할 수 있는데, 이는 형태소, BPE 토큰화 모델의 경우 1만 epoch까지 적절한 최적화 학습이 진행되었지만 이후 학습 데이터에 대한 과적합이 일어남으로 오히려 정확도가 떨어지고 있음을 보여준다. 자모 토큰화 모델의 검증 데이터 정확도는 모든 epoch에서 19~23 내에 분포하며 학습이 더 이상 진행되지 않고 정체되어 있는 양상을 보인다.

학습 데이터에서 측정된 혼잡도는 자모, 형태소, BPE 토큰화 모델 순으로 낮게 분포하며 epoch을 반복할 때마다 지속적으로 낮아지고 있다. 하지만 검증 데이터를 통해 측정된 혼잡도는 1만 epoch를 기점으로 영어 형태소, BPE 토큰 모델

의 혼잡도가 지속적으로 증가하고 있으며 5만 epoch 지점에서 330.6, 316.3이라는 높은 수치를 보였고, 영어 자모 토큰 모델 역시 90을 상회하는 분포를 보이고 있다. 이는 마찬가지로 올바른 모델 학습이 진행되지 않았음을 의미한다.



<그림 4> 한국어 BPE 토큰화를 적용한 모델의 정확도와 혼잡도

마지막으로 살펴볼 그래프는 한국어에 BPE 토큰화를 적용하고 영어에 자모, 형태소, BPE 토큰화를 적용한 모델에 대해 학습 데이터와 검증 데이터로 측정한 정확도, 혼잡도 그래프이다. 먼저 영어 자모 토큰 모델의 경우 학습 데이터의 정확도가 1만 epoch 이후 100에 근접하고 있지만, 검증 데이터에서 측정된 정확도는 20 정도에 머무는 것으로 확인된다. 영어 자모 모델의 혼잡도는 학습 데이터에서는 1만 epoch 이후 1정도를 유지하며 낮은 수치를 보이고 있으나, 평가 데이터로 측정된 혼잡도는 5천 epoch까지 급격하게 상승하며 80에 도달한 이후, 등락을 반복하며 지속적인 상승 추세를 보이고 있으며 5만 epoch 기준 97.06이라는 높은 혼잡도 수치를 보여준다.

이와 대조적으로 영어에 형태소, BPE 토큰화를 적용된 두 모델은 다른 모델들과 확연히 다른 양상을 보인다. 학습 데이터로부터 측정된 정확도는 5만 epoch

기준 다른 모델들 보다는 다소 낮은 82.61, 76.56을 보였는데, 검증 데이터에서 측정된 정확도는 70.32와 68을 보이며 다른 학습 모델들이 보이지 않은 높은 정확도를 기록하며 올바른 모델 학습이 진행되고 있음을 보여준다. 혼잡도의 경우, 학습 데이터에서 형태소 토큰화 모델이 1.853, BPE 토큰 모델이 2.413을 보여주었고, 평가 데이터에서도 이와 유사한 4.595, 4.319를 보였다. 이는 한국어에 BPE 토큰화를 적용하고, 영어에 형태소 토큰화와 BPE 토큰화를 적용하였을 때 가장 바람직한 기계 번역 모델이 학습됨을 보여주는 대목이다.

제3절 BLEU 점수 분석

		epoch	영어		
			자모	형태소	BPE
한국어	자모	10,000	0	**0.12**	0.03
		20,000	0	0.11	0.03
		30,000	0.03	0.09	0
		40,000	**0.04**	0.08	**0.05**
		50,000	0.03	0.07	0
	형태소	10,000	0	**0.37**	0.28
		20,000	0	0.18	0.19
		30,000	0	0.1	0.3
		40,000	0	0.1	0.13
		50,000	**0.07**	0.08	**0.31**
	BPE	10,000	0.08	30.28	17.73
		20,000	**0.09**	34.5	21.84
		30,000	0.08	35.36	22.78
		40,000	0.08	35.7	23.22
		50,000	0.08	**35.73**	**23.54**

<표 3> epoch 변화에 따른 각 모델의 BLEU 점수표

위 표는 각 모델이 5만 번의 epoch를 반복하는 학습 과정 중에 1만 epoch 지점마다 학습 모델을 저장한 후 이를 토대로 모델에 대한 BLEU 점수를 측정한

결과이다. BLEU 점수의 측정 방식은 기계 번역 모델이 번역한 문장과 사람이 번역한 문장으로 측정된 유니그램(unigram) 정밀도에 기반하며 일치하는 문자열이 많을수록 높은 점수를 보이게 된다.

위 표에서 보여주는 BLEU 점수는 대부분의 모델에서 1 이하의 낮은 점수를 보여주며 모델 학습이 제대로 진행되지 않았음을 보여준다. 반면, 한국어에 BPE 토큰화를 적용하고 영어에 형태소, BPE 토큰화를 적용한 모델의 경우 BLEU 점수가 상당히 높게 측정되었음을 확인할 수 있다. 이는 2절에서 정확도와 혼잡도를 기반으로 모델 학습 경과를 평가한 분석 결과와 일치하는 부분이다.

한국어와 영어에 모두 BPE 토큰화를 적용한 모델에서는 17~24 사이의 BLEU 점수를 보여주고 있다. 1만 epoch에서 17.73, 2만 epoch에서 21.84를 보이고 이후 미세한 상승 추세를 보이며 5만 epoch에서 23.54를 기록하였다.

본 실험에서 가장 좋은 BLEU 점수를 보여준 모델은 한국어에 BPE 토큰화를 적용하고 영어에 형태소 토큰화를 적용한 모델이다. 1만 epoch에서 이미 30.28을 보이는데, 이는 한국어와 영어에 모두 BPE 토큰화를 적용한 모델의 5만 epoch 지점의 BLEU 점수를 뛰어넘는 높은 점수이다. 2만 epoch에서 34.50을 보인 후 미세한 상승 추세를 이어가다 5만 epoch에서 35.73을 기록하며 본 실험에서 가장 높은 BLEU 점수를 기록하였다. 이 점수는 트랜스포머 모델을 처음 공개한 'Attention is All You Need' 논문에서 제시한 영어-독일어 번역이 28.4, 영어-프랑스어 번역이 41.0 BLEU 점수를 기록한 것을 감안할 때 언어적 유사도가 현저히 낮은 한국어-영어 번역 모델이 35.73을 보였다는 점은 모델 학습에 적용된 한국어 BPE 토큰화, 영어 형태소 토큰화가 가장 효과적인 토큰화 전략이였음을 나타내는 결과이다.

제4절 번역 문장 분석

번역 문장 분석에 사용된 문장은 BLEU 점수를 측정 때와 동일하게 모델 학습에 포함되지 않은 평가 데이터를 사용하였고, 이 중 몇 개의 문장을 추려서 번역 결과와 함께 제시하였다. 문장 번역 예측에 사용된 모델은 BLEU 점수 측정하였을 때 각 모델 중에서 가장 높은 점수를 보인 epoch 지점의 모델을 선정하여 동일 문장에 대한 번역 결과를 비교하였다.

한국어	**학생이 체육과 예술을 중시하게 하기 위한 프로그램이다.**
영어	**it is a program to make students value physical education and art.**
자모-자모	the reason of of the came a reason for this work.
자모-형태소	On the contrary, SK holds the world are needed.
자모-BPE	The **smill** for **obeging pressibly** the continued by sale.
형태소-자모	choi is said to be a dream for the **trension**.
형태소-형태소	The Ministry of Land, Infrastructure and Transport said that it will conduct a survey on the number of rental businesses in the first half of this year, and will conduct a survey on the number of rental businesses in the second half of this year.
형태소-BPE	The Seoul Metropolitan Government said on the 26th that it will hold a large-scale exhibition of "Mirae Asset Daewoo Securities Perfect Evaluation" at the National Assembly in Yeouido, Seoul, to commemorate the 100th anniversary of the March 1 Movement and the establishment of the corporation.
BPE-자모	**the sports** control make itself and public or say.
BPE-형태소	**It is a program designed to make students value physical education and art.**
BPE-BPE	**It is a program designed to make students value physical education and art.**

<표 4> 원문과 기계 번역 문장 예시 1

위에 제시된 원문과 번역된 문장을 살펴보면 한국어 BPE - 영어 형태소 토큰화 모델과 한국어 BPE - 영어 BPE 토큰화 모델의 번역 결과가 동일하게 출력되었고, 문법적, 의미적으로 올바른 번역 결과를 보여주고 있음이 확인된다. 원문

영어 문장과 다른 점은 “designed”가 포함되어 번역된 부분인데, 한국어 원문을 살펴볼 때 이는 제시된 영어 문장보다 오히려 원문의 의미를 잘 반영한 번역 결과라고 평가할 수 있다.

한국어 BPE - 영어 자모 토큰화 모델에서는 원문의 체육을 의미하는 “the sports”가 포함된 결과를 보였고, 그 외의 모델에서는 원문의 의미와 유사한 단어를 찾아보기 어렵다. 한국어에 형태소를 적용한 세 가지 모델의 번역 결과를 보면 형태적, 통사적으로 문제가 없는 문장 형태를 보여주었지만, 원문과는 전혀 다른 의미의 문장을 출력하였다는 점에서 적절한 번역 결과로 보기 어렵다. 이외의 모델에서는 “smill”, “obeging”, “pressibly”, “trension”과 같이 영어사전에 검색되지 않은 단어들이 다수 포함되어 있었으며, 문장 구조 역시 통사적 규칙에 어긋나는 다수의 문장을 확인할 수 있다.

한국어	**서울시는 다음달 3 일 오전 8 시부터 오후 2 시까지 서울둘레길 2 코스(양원역~광나루역)에서 시민 1,000 명이 참여하는 보물찾기 대회를 연다고 28 일 밝혔다.**
영어	**seoul city announced on 28th that it will hold a treasure hunt contest with 1,000 citizens at seoul 2nd course (from yangwon station to gwangnaru station) from 9am to 2pm next month.**
자모-자모	the two book the 180 more at than 108 more **agun**.
자모-형태소	The price of LCD TV panels continues to decline.
자모-BPE	The place, an **accumpared** an **USeould** have table.
형태소-자모	choi is said to be a dream for the trension.
형태소-형태소	The Ministry of Land, Infrastructure and Transport said that it will conduct a survey on the number of rental businesses in the first half of this year, and will conduct a survey on the number of rental businesses in the second half of this year .
형태소-BPE	The Seoul Metropolitan Government said on the 26th that it will hold a large-scale exhibition of "Mirae Asset Daewoo Securities Perfect Evaluation" at the National Assembly in Yeouido, Seoul, to commemorate the 100th anniversary of the March 1 Movement and the establishment of the corporation.
BPE-자모	the bible are growing to the movie dust in price.
BPE-형태소	**The Seoul Metropolitan Government said on the 28th that it will hold a treasure - hunting competition involving 1,000 citizens at the Seoul Dulegil (Yangwon Station - NCR Station) from 8 a.m. to 2 p.m. on the 3rd of next month.**
BPE-BPE	**The Seoul Metropolitan Government announced on the 28th that it will hold a treasure search competition involving 1,000 citizens at the Seoul Dulle-gil Trail 2-course (Yangwon Station to Gwangnaru Station) from 8 a.m. to 2 p.m. on the 3rd of next month.**

<표 5> 원문과 기계 번역 문장 예시 2

다음으로 제시된 원문과 번역문 중에서 먼저 한국어 BPE – 영어 형태소 모델과 한국어 BPE - 영어 BPE 토큰화 모델의 출력 결과를 살펴보면 “서울시”가 “The Seoul Metropolitan Government”로 번역되었음을 확인할 수 있다. 이는 서울특별시청의 공식 영어 표기 명칭에 해당하는 표현으로 적절한 번역 결과로 평가할 수 있다. 영어 원문에 “treasure hunt contest”라고 표기되어 있는 “보물찾기 대회”는 BPE-형태소 모델에서 “treasure-hunting competition”이라 번역되었고, BPE-BPE 모델에서는 “treasure search competition”이라 번역되었다. BPE-BPE 모델이 BPE-형태소 모델에 비해 다소 직역에 가까운 번역 결과를 보인 것으로 보여진다. “서울둘레길 2코스”는 BPE-형태소 모델에서 “Seoul Dulegil”이라고 번역되며 “2코스”가 번역 결과에서 탈락되었지만, BPE-BPE 모델은 Seoul Dulle-gil Trail 2-course이라 번역하며 올바른 번역 결과는 아니지만 “2 코스”의 의미에 상응하는 단어를 포함한 번역 결과를 보여주었다. “광나루역” 명칭은 BPE-BPE 모델의 경우 “Gwangnaru Station”라고 올바르게 로마자 표기법에 따라 출력하였지만, BPE-형태소 모델은 “NCR Station”이라고 오역하며 고유 명사에 있어 다소 아쉬운 출력 결과를 보였다.

한국어에 형태소 토큰화를 적용하고 영어에 자모, 형태소, BPE 토큰화를 적용한 모델에서는 이전의 번역 문장과 동일한 문장을 출력하였다. 다른 문장을 입력한 경우에도 동일한 결과로 출력되었는데, 이는 위 모델들이 특정 문장을 기준으로 과적합이 되어 입력 문장과 무관한 출력을 내도록 학습된 것으로 판단된다. 그 외 다른 모델의 출력 문장에서는 “agun”, “accumpared”, “USeould” 등과 같은 영어사전에서 찾을 수 없는 단어들이 다수 포함되어 이전 문장의 번역 결과에서 관찰된 특징이 반복되어 확인된다.

한국어	대전시 대덕구 대화동 대전산업단지의 근무 환경이 개선된다.
영어	the working environment of daejeon industrial park in daedeok-gu, daejeon will be improved.
자모-자모	this is the one work the dieact will be ened.
자모-형태소	It is expected that the cooperation between Gijo and medium - level departments will be strengthened.
자모-BPE	One manager will be a stranger in person.
형태소-자모	choi is said to be a dream for the trension.
형태소-형태소	The Ministry of Land, Infrastructure and Transport said that it will conduct a survey on the number of rental businesses in the first half of this year, and will conduct a survey on the number of rental businesses in the second half of this year.
형태소-BPE	The Seoul Metropolitan Government said on the 26th that it will hold a large-scale exhibition of "Mirae Asset Daewoo Securities Perfect Evaluation" at the National Assembly in Yeouido, Seoul, to commemorate the 100th anniversary of the March 1 Movement and the establishment of the corporation.
BPE-자모	the human expected shots like a dramage to bloom.
BPE-형태소	The working environment of Daejeon Industrial Complex in Daehwa-dong, Daedeok-gu , Daejeon will be improved.
BPE-BPE	The working environment of Daejeon Industrial Complex in Daehwa-dong, Daedeok-gu, Daejeon is improved.

<표 6> 원문과 기계 번역 문장 예시 3

위에 제시된 원문과 번역문에서도 마찬가지로 올바른 번역으로 출력된 경우는 한국어 BPE - 영어 형태소 토큰화 모델과 한국어 BPE - 영어 BPE 토큰화 모델로 추릴 수 있다. 고유명사에 해당하는 “대전시 대덕구 대화동”을 두 모델 모두 “Daehwa-dong, Daedeok-gu, Daejeon”으로 번역하며 적절한 번역 결과를 보였다. 한국어 문장에서 “개선된다”라는 표현은 bpe-형태소 모델의 번역 결과에서 “will be improved”라고 번역하며 미래 시제로 명시되었지만, bpe-bpe 모델은 “is improved.”이라고 번역하며 단순 현재시제로 번역되었다. 이는 한국어 문장에서 명확하게 드러나지 않은 미래 시제 표현이 번역하는 과정에서 제대로 옮겨지지 못한 것으로 판단된다.

위 번역 결과에서도 한국어 형태소 토큰화를 적용한 모델들에 대한 동일 문장 출력 문제가 반복되고 있음이 확인되며, 자모-자모 모델에서 “근무”에 해당하는 “work”가 등장한 것 외에는 원문에 기반한 단어를 찾아볼 수 없었다. 그 외의 번역 문장에서 “dieact”, “ened”, “dramage”와 같은 영어사전에 검색되지 않는 단어가 다수 관찰되었는데, 이는 도착어인 영어에 자모 토큰을 적용한 경우에 종종 관찰되는 특징으로 판단된다.

제4장 결론 및 한계점

본 연구에서는 한국어와 영어에 자모, 형태소, BPE 토큰화 방법을 적용하여 총 9개의 모델 학습을 진행하였다. 모델의 BLEU 점수 측정 결과, 한국어 BPE – 영어 형태소 토큰화 모델과 한국어 BPE – 영어 BPE 토큰화 모델을 제외한 나머지 7개의 모델에서 1점이 채 되지 않는 BLEU 점수가 측정되었고, 검증 데이터의 정확도와 혼잡도를 살펴보아도 적절한 모델 학습이 진행되었다고 보기 어려웠다. 번역된 문장을 원문과 직접 대조해보았을 때 간혹 비슷한 의미의 단어가 출현함은 확인되었지만, 우연이 아닌 의도된 단어 선정이였는지 확신할 수 없으며, 어떠한 입력 문장에 대해서 동일한 문장을 출력하는 문제와 영어사전에서 검색되지 않는 단어들이 등장하는 문제들을 반복적으로 발견할 수 있었다. 또한 한국어와 영어에 자모 토큰화를 적용한 모델의 경우 epoch과 무관하게 낮은 BLEU 점수를 보여주었는데, 이는 자음과 모음 단위에 의미를 형성함이 현실적으로 어려웠기 때문일 것으로 판단된다.

한국어 BPE – 영어 형태소 토큰화 모델과 한국어 BPE – 영어 BPE 토큰화 모델은 학습 데이터와 비례하는 검증 데이터의 정확도와 혼잡도가 측정되었고, BLEU 점수도 각각 35.73, 23.54를 보이며 적절한 모델 학습이 진행되었음을 확인할 수 있었다. 실제 번역 문장 결과를 통해 살펴본 결과, 한국어 BPE – 영어 형태소 토큰화 모델에서 문법적 특징을 더 잘 반영한 번역 결과를 보이고 원문에 충실한 번역 결과를 보여주었다고 판단되었고, 한국어 BPE – 영어 BPE 토큰화 모델의 경우 학습 데이터에 포함되지 않은 명사 내지 고유명사 표기에 더 우수한 성능을 보여주었다고 평가된다. 두 모델의 각 장점을 고려해 보았을 때, 두 토큰화 방법을 적절히 조합한다면 더 좋은 성능의 모델 학습이 가능할 것이라 기대해 볼 수 있다.

본 연구에서는 한국어, 영어 문장의 토큰화 방법에 주안점이 맞춰져 있어 토큰의 개수에 변화를 주지 않았다는 한계점이 있다. 형태소 토큰의 개수를 조절해가며 실험을 진행한다면 현 학습 모델에서 데이터 과적합으로 인해 보이는 동일 문장 출력 문제 역시 개선할 수 있을 것으로 기대해 볼 수 있다. 뿐만 아니라, 본

연구의 모든 모델의 학습 반복은 5만 epoch으로 설정하여 실험이 진행되었는데, BPE-형태소와 BPE-BPE 모델 모두 epoch을 거듭할수록 미세한 성능 개선이 진행되고 있었던 점을 감안하면 epoch을 더 높게 설정하여 실험을 진행하는 모델 학습에 대한 후속 연구가 필요하다고 보여진다.

참고문헌

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation, https://arxiv.org/abs/1701.02810

Kyubyong Park, Joohong Lee, Seongbo Jang, Dawoon Jung. 2020. An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks, AACL-IJCNLP 2020

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715– 1725, Berlin, Germany. Association for Computational Linguistics.

Sangwhan Moon and Naoaki Okazaki. 2020. Jamo pair encoding: Subcharacter representation-based extreme Korean vocabulary compression for efficient subword tokenization. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 3490–3497, Marseille, France. European Language Resources Association.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015

ABSTRACT

Korean-English Machine Translation with Multiple Tokenization Strategy

This study was conducted to examine the effect of tokenization method on machine translation model training. For the experiment, Transformer Neural Network was built using OpenNMT-py library to set the same hyperparameter, and 9 models were repeated by 50,000 epochs for the source language, Korean, and target language, English, by applying the consonant-vowel tokenization, morpheme tokenization, and BPE tokenization respectively. The data used in the training, validation and evaluation stages of the models are 800,000 sentence pairs of Korean-English news corpus provided by AI Hub.

As a result of measuring the BLEU score of the trained models, the model with Korean BPE tokenization applied to English morpheme tokenization recorded 35.73, and the model with Korean BPE tokenization applied to English BPE tokenization showed the next highest performance with 23.54. These two models have been shown to be well-trained even through the high accuracy and low perplexity of training and validation data. Other models had a low BLUE score of less than 1, which was also confirmed by the accuracy and perplexity of the validation data.

Comparing the translations and the original text of each trained model, the model with Korean BPE tokenization applied to English morpheme tokenization showed the most faithful translation of the original text and the model with Korean BPE tokenization applied to English BPE tokenization showed excellent performance in the translation of sentences that contained nouns or unique nouns not included in the training data.