Transformer 를 사용한 영한 기계 번역

천진우*, 구자환*, 김응모*
*성균관대학교 소프트웨어대학
czw4653@skku.edu, jhkoo@skku.edu, ukim@skku.edu

English-Korean Machine Translation using Transformer

Jin-woo Chun*, Jahwan Koo*, Ung-Mo Kim*
* College of Software, Sungkyunkwan University
czw4653@skku.edu, jhkoo@skku.edu, ukim@skku.edu

요 약

최근 자연어 처리 기술은 지속적으로 발전하고 있으며, 많은 분야에서 활용되고 있다. 그 중 번역 기술은 가장 널리 사용되고 있는 자연어 처리 기술 중 하나이다. 본 논문에서는 기존의 seq2seq 모델의 단점을 극복하기 위해 개발된 Transformer 를 통해 영어-한국어 번역기를 만드는 것의 가능성을 제시한다.

1. 서론

최근 NLP(Natural Language Processing, 자연어처리) 기술의 발달로 인해 문서 요약, 자동완성, 챗봇, 번역기 등 관련된 다양한 서비스가 연구, 개발되고 있다. 특히 번역기는 이미 많은 사람들에게 일상적이고 보편적인 도구가 되었고, 여행과 교육 등 다양한 분야에서 활용되고 있다.[1][2] 그러나 아직까지 곳곳에서 잘못 번역되어지는 사례가 발견되고 있어 성능 개선의 필요성이 느껴지고 있다.

기존의 RNN(Recurrent Neural Network)을 활용한 모델의 문제점을 해결하기 위해 구글에서는 attention 을 사용하는 Transformer 라는 모델을 만들었다.[3] Transformer 는 문장이 길어질수록 성능이 떨어지게 되는 기존 모델의 문제점을 해결할 뿐만 아니라 학습 시간이 매우 빠르다는 장점 또한 가지고 있다. Google 은 이 모델을 활용하여 이탈리아어, 불어, 영어 등의 언어들을 서로 번역할 수 있는 모델을 만들었다.[4]

하지만 이 모델은 한국어는 아직까지 지원해주지 않는다. 그래서 본 논문에서는 한국정보화진흥원의 한국어, 영어 데이터들을 이용하여 모델을 학습시키고, Transformer 모델을 통한 개선된 영한 번역기에 대한 가능성을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구로서 Transformer 모델의 구조와 특징에 대해 기술하고, 3 장에서는 영어-한국어 번역기를 만들기 위한 방법과 실험 방법에 대해 기술하며, 4 장에서는 3 장의 결과를 기술한다. 마지막 5 장에서는 결론을 기술한다.

2. 관련 연구

본 장에서는 Transformer 모델의 구조와 각각의 기능, 모델의 사용 이유에 대해 설명한다.

2.1 Trnasformer 특징

RNN 기반 sequence to sequence 모델은 인코더와 디코더로 구성된다. 인코더는 sequence 의 모든 의미를 단일 벡터로 축소하고 이를 디코더로 전달하며, 디코더는 정보를 처리하고 예측한다. 디코더는 타임 스탭마다 다른 정보를 요구하지만 하나의 벡터로부터 시퀀스의 종속성 정보를 얻기 때문에 정보가 손실되는 문제가 발생한다.

LSTM(Long-Short Term Memory)은 정보를 변형하여 중요도에 따라 유지할 정보와 잊을 정보를 선택할 수 있도록 한다. 하지만 이 역시 문장이 길어지게 되면 모델이 먼 위치의 정보를 잃게 되는 경우가 많다. 이 외에도 그들은 단어 단위의 순차적 진행을 하기 때문에 병렬화가 힘들다는 단점도 있다.[5]

이런 문제를 해결하기 위해 Attention Mechanism 을 사용하는 Transformer 라는 새로운 모델이 제안되었다. Transformer 는 Google 연구팀에서 테스트한 기계 번역 시험에서 가장 높은 BLEU[6] 점수를 획득했으며, 다른 모델에 비해 학습하는데 드는 계산과 시간이 훨씬 적었다.[7]

2.2 Transformer 구조

2.2.1 Encoder and Decoder

Transformer 는 인코더와 디코더로 이루어져 있고, 각각은 N 개의 layer 로 이루어져 있다. 인코더의 layer 는 Multi-Head Attention 과 Point Wise Feed Forward networks 두개의 sublayer 들로 구성되며 각각은 residual connection 을 적용하고 normalize 한다. 디코더 layer 의 기본 구성은 인코더와 유사하지만 Multi-Head Attention 에 마스크를 추가하여 순차적 진행을 가능하게 하였고, 두 sublayer 사이에 Multi-Head Attention sublayer 하나를 추가하여 인코더의 아웃풋 정보를 이용할 수 있도록 하였다.

2.2.2 Attention

Multi-Head Attention 은 dimension 을 나누어 Scaled Dot-Product Attention 을 적용할 수 있도록 한다. 입력 벡터를 linear 하게 나누고 attention 을 시킨 후, 만들어진 벡터들을 concat 하고 dimension 을 다시 원래대로 만든다. 이는 문장의 종속성을 다양한 측면에서 파악할 수 있도록 해준다.

Scaled Dot-Product Attention 은 들어온 Query 와 Key 를 내적하여 단어들 사이의 유사도를 측정하고, scaling 과 softmax 를 적용하여 weight 를 만든다. 그 후, Value 를 weight 와 곱하여 중요도가 높은 값을 더 크게 키워준다.

2.2.3 Positional Encoding

모델이 시퀀스 순서를 활용하기 위해 토큰의 상대적 또는 절대적 위치에 대한 정보를 제공해 준다. Transformer 에서는 이를 위해 sine 함수와 Cosine 함수를 사용한다.[8]

2.2.4 Feed Forward Neural Network

Feed Forward Neural Network 는 두 개의 fully connected layer 와 활성화함수 ReLU 로 이루어져 있다. 이는 학습 프로세스를 강화하고 시퀀스의 새로운 종속성을 자체적으로 학습할 수 있도록 한다.

2.2.5 Residual connection

Residual connection 은 sublayer 이전의 입력값을 통해 Back-propagation 를 하여 weight 를 쉽고 효율적으로 업데이트 할 수 있도록 해준다.

3. 영한 번역기 구축

3.1 구축환경

영한 번역기 모델은 구글 Colab 에서 구현하였고, tensorflow 플랫폼과 python 언어를 사용하여 모델을 만들고 학습, 평가했다.

3.2 Dataset

한국어, 영어 번역 데이터셋을 만들기 위해 한국정보화진흥원의 한국어-영어 구어체, 대화체, 문어체(뉴스) 데이터들을 사용했다. 데이터를 구성하는 문체의 종류와 데이터의 양에 따라 번역기의 정확도 차이를 파악하고 최적의 데이터셋을 사용하기 위해 본 논문에서는 다양한 데이터셋을 구축하고 실험을 진행했다.

3.2.1 문체에 따른 데이터셋

구어체 문장 10000 개를 사용한 데이터셋 1, 대화체 문장 10000 개를 사용한 데이터셋 2, 문어체 문장 10000 개를 사용한 데이터셋 3, 문어체 4000 문장, 대화체 2500 문장, 구어체 3500 문장을 함께 사용한 데이터셋 4 를 만들어 모델 학습을 진행하고, 학습된 모델을 바탕으로 번역 정확도를 측정, 비교하였다.

3.2.2 크기가 다른 데이터셋

구어체, 대화체, 문어체를 7:5:8 의 비율로 혼합하여 한국어, 영어 문장 10000 개를 사용한 데이터셋 4, 문장 30000 개를 사용한 데이터셋 5, 문장 100000 개를 사용한 데이터셋 6, 문장 300000 개를 사용한 데이터셋 7 를 만들어 모델을 학습하고 정확도를 측정했다.

3.3 모델 구조

기본적으로 Transformer 는 모델의 입력, 출력 사이즈 d_model 을 512, layer 수를 6 으로 한다. 본 논문에서는 이들을 바꿔가며 번역기를 만들고 그 정확도를 살펴봤다.

4. 모델 평가

4.1 모델 평가 방식

본 논문에서는 모델의 번역 정확도를 평가하기 위해 BLEU 점수를 활용했다. BLEU 점수는 0 에서 1 사

이의 값으로 나타내어지며 0 에 가까울수록 낮은 정확도, 1 에 가까울수록 높은 정확도를 가진 번역기를 뜻한다.

본 논문에서는 데이터셋이 서로 다른 각 모델들의 번역 정확도를 비교하기 위해 모델들의 평균 BLEU 점수를 계산했다. 평균 BLEU 점수는 데이터셋의 모든 문장들에 대해 모델의 BLEU 점수를 측정하고 더한 뒤, 전체 문장 수를 나누어 주어 계산했다.

<표 1>, <표 3>, <표 4>에서는 테스트 데이터셋으로 구어체, 문어체, 대화체를 7:8:5 의 비율로 혼합한 1000 개의 문장을 사용했고, <표 2>에서는 각각의 문체를 100 개씩 사용하여 문체별로 따로 점수를 계산했다.

4.2 문체 구성에 따른 모델 정확도 평가

<표 1> 학습 데이터 종류에 따른 BLEU 점수 1

	평균 BLEU 점수
데이터셋 1	0.674
데이터셋 2	0.648
데이터셋 3	0.661
데이터셋 4	0.672

<표 2> 학습 데이터 종류에 따른 BLEU 점수 2

	구어체	문어체	대화체
데이터셋 1	0.668	0.680	0.672
데이터셋 2	0.635	0.642	0.667
데이터셋 3	0.644	0.691	0.634
데이터셋 4	0.662	0.689	0.675

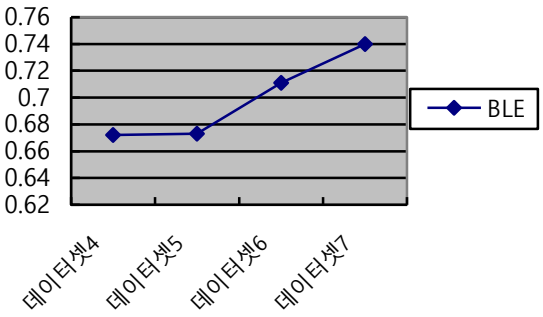
<표 1>을 보면 구어체로 학습한 모델과 혼합된 데이터셋을 통해 학습한 모델이 비교적 BLEU 점수가 높다는 것을 확인할 수 있다. 그러나 <표 2>에서 2 행과 3 행의 결과를 살펴보면 단일 문장 종류로만 모델을 학습할 경우 동일한 종류의 문장에 대한 정확도는 높아질 수 있지만 다른 문장에 대해서는 비교적 정확도가 떨어진다는 것을 알 수 있다.(문어체로 학습한 모델의 경우 문어체에 대한 BLEU 점수는 약 0.69 정도 나오지만 다른 둘은 약 0.63, 0.64 정도밖에 되지 않는다) 따라서 모든 문장에 대해 전반적인 번역 정확도를 높이기 위해서는 혼합된 데이터셋을 통해 모델을 학습하는 것이 바람직해 보인다.

4.3 데이터 양에 따른 모델 정확도 평가

<표 3> 데이터 크기에 따른 BLEU 점수

	평균 BLEU 점수
데이터셋 4	0.672
데이터셋 5	0.673
데이터셋 6	0.711
데이터셋 7	0.740

<차트 1> 데이터 크기에 따른 BLEU 점수



측정 결과에 따르면 10000 개의 데이터셋을 사용한 모델과 30000 개를 사용한 모델은 큰 차이가 나지 않았지만 100000 개, 300000 개를 사용한 모델의 BLEU 점수는 이전 모델들에 비해 높게 나타났다. 따라서 더 많은 양의 데이터를 수집해서 학습에 사용할수록 더 높은 정확도를 가진 번역기를 만들 수 있을 것이라 기대된다.

4.4 모델 구조에 따른 모델 정확도 평가

<표 4> layer 수와 dimension 에 따른 BLEU 점수

	BLEU
데이터셋 4, Layer : 2, d_mode l: 64	0.677
데이터셋 4, Layer : 4, d_mode l: 128	0.672
데이터셋 4, Layer : 6, d_model : 512	0.655
데이터셋 4, Layer : 8, d_mode l: 1024	0.505

같은 데이터셋으로 layer 수와 d_model 을 바꿔가며 모델을 학습시키고 번역 정확도를 측정한 결과 layer 수와 d_model 이 클수록 낮은 BLEU 값을 보였다.

4.5 학습시간

<표 5> 데이터 크기에 따른 학습시간

	학습시간
문장 10000 개	약 1 시간
문장 30000 개	약 3 시간
문장 100000 개	약 10 시간
문장 300000 개	약 30 시간

<표 6> 모델 구조에 따른 학습시간

	학습시간(10000 문장 당)
Layer : 4, d_model : 128	약 1 시간
Layer : 6, d_model : 512	약 4 시간

학습 시간은 데이터 크기가 커짐에 따라 그에 비례하여 증가하였으며, layer 수와 모델 사이즈의 증가 또한 학습시간을 늘리는 요인이 된다.

4.6 문장 번역 결과

4.6.1 잘 번역된 문장

영어 문장: I'm going to go to the zoo.
번역 결과 : 나는 동물원에 갈 예정입니다.
BLEU : 0.903

영어 문장: I propose two agendas.
번역 결과 : 두 가지 제안을 제안합니다.
BLEU : 0.903

영어 문장 : I sent the money to the company.
번역 결과: 내가 그 회사에 돈을 송금했습니다.
BLEU : 0.851

4.6.2 번역이 잘 안된 문장

영어 문장: I was very sad and tired.
번역 결과: 너무 피곤해서 너무 피곤해요.
BLEU : 0.594

영어 문장: Well, it tastes completely different according to the size.
번역 결과: 그러면 사이즈는 다른 사이즈입니다, 한 사이즈입니다.
BLEU : 0.557

5. 결론

자연어 처리 기술의 발달로 인해 구글, 네이버, 카카오 등 많은 기업에서 번역기를 개발하였고, 계속해서 그 성능이 향상되고 있다. 그는 많은 사람들에게 일상적으로 사용하는 도구가 되었으며, 앞으로 더 많은 분야에서 활용될 것으로 기대된다. 하지만 아직까지 번역 과정에서 다소 문제점들이 발견되고 있으며, 이를 해결할 필요성이 있다.

그를 위해 본 논문에서는 기존의 방식 대신 Transformer 모델을 사용하여 영한 번역기를 제작하게 되었고, 데이터와 모델 구조에 따른 번역 정확도를 측정, 비교했다.

데이터셋의 구조와 데이터의 양을 바꾸어 가며 모델을 학습시키고 BLEU 점수를 측정한 결과, 번역은 되었지만 그 정확도가 매우 뛰어나다고는 볼 수 없었다. 그러나 문장의 종류를 적절히 혼합하고, 데이터의 양을 증가시킴에 따라 정확도 또한 향상되는 것을 볼 수 있었다. 따라서 다양한 종류의 문장들을 많이 확보하여 학습에 이용할 수 있다면 양질의 번역기를 만들 수 있을 것이라고 기대된다.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) founded by the Ministry of Education(NRF-2018R1D1A1B07049464).

참고문헌

[1] 이윤재, “영어자동번역기 활용이 고등학생 영어 글쓰기에 미치는 영향”, 한국교원대학교 교육대학원, 2020
[2] 지혜원, “초등영어 쓰기 수업에서 효과적인 번역기 활용 방안”, 서울대학교 교육전문대학원, 2020
[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, “Attention is All you Need”. NIPS 2017.
[4] “trainformer.ipynb”, GitHub, last modified sep 10, 2020, accessed sep 12, 2020, https://github.com/tensorflow/docs/blob/master/site/en/tutorials/text/transformer.ipynb
[5] “How Transformer Work”, towards data science, last modified mar 11, 2019, accessed sep 12, 2020, https://towardsdatascience.com/transformers-141e32e69591
[6] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”, ACL, Philadelphia, 2002
[7] “Neural Machine Translation with Transformers”, Medium, 2020 년 4 월 17 일 수정, 2020 년 9 월 27 일 접속, https://medium.com/@galhever/neural-machine-translation-with-transformers-69d4bf918299
[8] “[AI TECH 컬럼] AI 분야 ‘어텐션 트랜스포머’의 위치 엔코딩에 대한 고찰”, 인공지능 신문, 2020 년 8 월 24 일 수정, 2020 년 9 월 13 일 접속, http://www.aitimes.kr/news/articleView.html?idxno=17442