BERT layer를 합성한 Transformer 모델에 적용한 Cardinality Residual connection 방법

최규현○, 이요한, 김영길
과학기술연합대학원대학교○, 한국전자통신연구원
choko93@ust.ac.kr○, carep@etri.re.kr, kimyk@etri.re.kr

The Cardinality Residual Connection Method Applied to Transformer Model combining with BERT Layer

Gyu-Hyeon Choi○, Yo-Han Lee, Young-Kil Kim
University of Science and Technology○, Electronics and Telecommunications Research Institute

요 약

본 논문에서는 BERT가 합성된 새로운 Transformer 구조를 제안한 선행연구를 보완하기 위해 cardinality residual connection을 적용한 새로운 구조의 모델을 제안한다. Transformer의 인코더와 디코더의 셀프어텐션에 BERT를 각각 합성한 모델의 잔차연결을 수정하여 학습 속도와 번역 성능을 개선하고자 한다. 그리고 가중치를 다르게 부여하는 실험으로 어텐션을 선택하는 효과적인 방법을 제시하고 원문의 언어에 맞는 BERT를 사용하는 이유를 설명한다. IWSLT14 독일어-영어 말뭉치와 AI hub에서 제공하는 영어-한국어 말뭉치를 이용한 실험에서는 제안하는 방법의 모델이 기존 모델에 비해 더 나은 학습 속도와 번역 성능을 보였다.

주제어: BERT, Transformer, 기계번역, 잔차연결, Cardinality residual connection

1. 서론

최근 기존의 순환 신경망, 합성곱 신경망 기반의 기계 번역 모델의 성능을 능가한 트랜스포머(Transformer)[1] 번역 모델이 등장하여 번역 분야에 새로운 패러다임을 제시했다. 트랜스포머는 순환 신경망을 제거하고 셀프어텐션 매커니즘(self attention mechanism)[2]만으로 모델을 구성했다. 인코더(encoder)는 입력 문장의 단어들을 양방향의 문맥을 고려하여 각각 연속적인 표현으로 생성하고 디코더(decoder)는 이 표현들을 통해 번역 단어를 출력한다.

BERT[3]는 트랜스포머의 인코더 구조를 기반으로 학습된 언어 모델(language model)이다. 문장 내에서 임의의 단어를 마스킹하고 이를 예측하는 masked 언어 모델(masked language model)을 학습하며, 주어진 두 문장이 연결된 문장인지 예측하는 다음 문장 예측(next sentence prediction)을 함께 학습한다. 이를 통해 양방향의 문맥을 고려할 수 있고 문장 수준의 이해를 바탕으로 하는 언어 표현을 생성할 수 있다. BERT는 대용량의 단일 코퍼스로부터 모델을 사전 학습(pre-trained)하고 태스크에 따라 fine-tuning 하는 방법으로 사용된다.

BERT의 구조가 트랜스포머 인코더의 구조에 기반 한다는 점을 이용하여 트랜스포머에 BERT 네트워크를 합성한 새로운 번역 모델을 제시한 다양한 연구가 진행되었다. 입력 언어로 fine-tuning한 BERT를 트랜스포머의 새로운 인코더로 대체하여 사용하거나, 기존의 인코더를 유지하고 BERT를 추가하여 두 개의 인코더를 사용하는 방법들이 제안되었다. 또, fine-tuning을 하지 않은 BERT의 출력층을 트랜스포머의 어텐션과 합성하는 방법도 제안되었다. 하지만, 선행연구에서는 트랜스포머에 BERT가 추가되어 번역 성능은 향상되어도 파라미터가 증가하고 잔차연결(residual connection)의 연산 횟수가 증가하여 학습 속도를 낮추는 단점이 존재한다.

이를 해결하기 위해 본 논문에서는 잔차연결을 병렬구조로 구성하여 학습 속도를 향상시킨 방법을 제시한 ResNeXt[4]의 연구를 토대로 BERT가 합성된 트랜스포머의 새로운 구조를 제안한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 선행연구들을 설명하고, 3장에서는 BERT가 합성된 트랜스포머를 개선하기 위해 새롭게 적용한 방법들을 설명한 뒤, 4장에서는 실험 내용과 결과를 살펴보고 5장에서는 결론을 맺는다.

2. 관련 연구

2.1 Residual connection(ResNet)

모델의 깊이가 너무 깊어질수록 그래디언트 손실(gradient vanishing) 문제 때문에 오히려 학습이 잘 이루어지지 않아 성능이 떨어지는 경향이 있었다. 이를 해결하기 위해서 ResNet[5]이 고안되었는데, 기존 CNN(convolution neural network)[6] 네트워크층에 skip connection을 추가한 것으로 2개의 합성곱(convolution) 네트워크층을 쌓은 것에 입력을 그대로 위에 더해주는 형태로 구성한 구조를 제안했다. 이를 통해 네트워크층이 깊어짐에 따른 그래디언트 손실 문제를 해결했다.

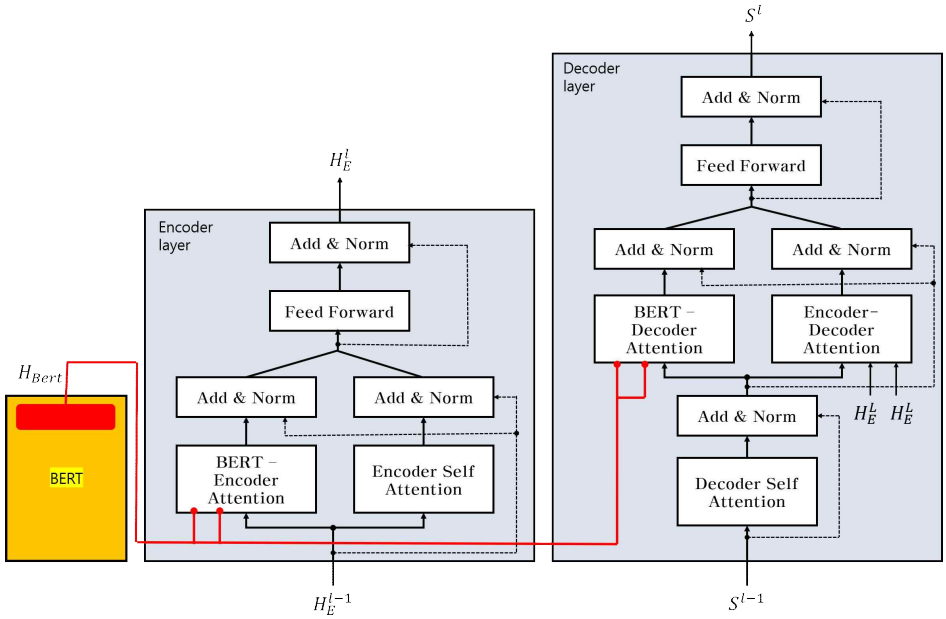

그림 1 Cardinality Residual Connection을 적용한 BERT 합성 트랜스포머 모델

2.2 Cardinality Residual Connecion

기존 ResNet의 수직적인 네트워크 구조는 깊어질수록 학습할 파라미터의 수가 늘어나 학습 속도가 느려지는 단점이 있다. 이를 해결하고자 ResNeXt[4]가 제안되었다. 동일한 블록을 병렬구조로 구성하여 블록을 반복적으로 구축했다. 이를 통해 더 적은 파라미터로 이미지를 분류할 수 있음을 확인하고 분류 정확도를 향상할 수 있다고 제안한다. 이때 동일한 블록의 개수를 Cardinality라 칭한다. 그리고 ResNeXt는 "Split-Transform-Merge"라는 개념을 사용하는데, 예를 들면 Cardinality가 32이면, 8채널씩 나눠서 하나의 그룹을 생성하고 합성곱을 하는 구조로 만드는 방법이다. 벡터 X에 대해 분리하기(split)을 사용하여 $X_1, X_2, \ldots, X_n$로 분리한다. 분리된 각각의 저차원의 벡터는 가중치 w와 곱하여 크기를 변화(transform)시킨다. 마지막에 모든 저차원의 벡터를 합치는(merge) 과정을 거친다. 분리과정을 통해 파라미터 수를 줄였기 때문에 학습 속도가 향상되었고 특정 데이터셋에 치우치는 과적용(over-adaptation)현상을 줄일 수 있다. 그리고 ILSVRC 2016 분류 분야에서 2등을 차지하였고 다양한 모델들보다 간단한 구조로 더 높은 성능을 증명했다.

2.3 BERT 출력층을 합성한 Transformer 모델

트랜스포머는 순환 및 합성곱 구조를 배제하고 어텐션 메커니즘 방식에 초점을 두어 구성한 시퀀스 변환 모델이다. 이 모델은 연산의 병렬 처리를 통해 학습시간을 줄인 것이 특징이며, 기계번역 분야에서 준수한 성능을 보였다.

BERT는 양방향성을 가진 트랜스포머를 기반으로 하고 대용량 코퍼스로 학습된 언어 모델이다. 셀프어텐션 매커니즘으로 문맥 전체를 확인하여 언어 모델을 학습한다. BERT의 구조가 트랜스포머의 인코더를 기반으로 하기 때문에 BERT의 출력층을 트랜스포머의 인코더, 디코더에 연결하여 모델의 구조를 변경한 다양한 연구가 등장했다[7,8,9]. 트랜스포머의 인코더를 BERT로 대체하여 사용한 모델들[7,8]과 트랜스포머의 인코더와 디코더의 셀프어텐션과 BERT의 출력층이 합성되어 학습된 모델도 제안되었다[9]. 제안되었던 연구들에서 원문(source language sentence)의 언어의 BERT를 사용했을 때 성능의 향상을 보였고 이를 통해 말뭉치의 양이 적은 low-resource 환경에서 활용될 수 있음을 강조했다.

3. 제안 방법

본 연구에서는 트랜스포머의 인코더와 디코더의 셀프어텐션에 BERT를 각각 추가로 합성한 선행연구의 방법[9]을 재현하고 기존 잔차연결을 병렬구조로 수정하는 cardinality residual connection 방법을 적용하여 학습 속도와 성능의 변화를 확인하고자 한다. 제안 방법으로 설계한 모델은 그림1과 같고 선행연구 모델 구조를 기반으로 새롭게 설계된 모델이다. BERT의 출력층이 인코더와 디코더의 셀프어텐션층과 각각 합성되고 각 인코더와 디코더의 추가적인 네트워크층으로 사용된다. 선행연구에서는 합성어텐션에서 계산된 결과와 셀프어텐션에서 계산된 결과 중 하나를 무작위로 하나만 선택하여 잔차연결과 정규화 과정을 거친다. 이 과정에서 하나의 어텐션에 치우쳐지는 현상이 발생하여 BERT를 효과적으로 활용하기가 어렵고 수직적인 네트워크 구조로 인해 연산량이 증가하여 학습 속도가 저하되는 단점이 있다. 이를 극복하기 위한 방법으로 cardinality residual connection을 사용한 새로운 트랜스포머 구조를 제안한다.

Cardinality는 똑같은 형태의 네트워크를 옆으로 나열한 개수를 의미한다. 이는 Residual network의 깊은 수직 구조의 네트워크를 병렬구조로 변경하여 네트워크의 깊이를 낮추고 파라미터의 수를 줄여 학습 속도를 향상시

표 1 두 개의 어텐션에 부여한 가중치에 따라 변화하는 성능 평가

기존 어텐션	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
BERT 합성 어텐션	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
BLEU	35.09	35.42	35.17	35.12	35.17	**36.12**	35.04	35.14	28.27	17.61	34.49

킨 것을 의미한다. Cardinality 개념을 사용하여 기존 방법을 보완하고자 한다. 선행연구와 다르게 합성어텐션과 셀프어텐션을 바로 합하지 않고 잔차연결을 따로 적용하여 연산을 수행한다. 정규화까지 마친 각 어텐션의 출력값은 하나의 feed forward에서 합쳐지는데, 이때 각 어텐션을 선택하는 방법은 선행연구와 다른 방법을 사용한다. 선행연구에서는 BERT 합성어텐션과 기존의 어텐션을 선택할 때 무작위로 하나만 선택하는 방법을 사용하는데, 이 방법은 문장의 임베딩 벡터에 따라 문장의 해석과 표현이 잘 반영되는 특정 어텐션을 효과적으로 사용하지 못한다는 단점이 있다. 그래서 본 연구에서는 무작위로 어텐션을 선택하는 방법이 아닌 두 개의 어텐션에 임의의 가중치를 두어 모든 어텐션을 반영하는 방법을 사용한다.

또한, 연구를 통해 우리는 트랜스포머를 학습하기 위해서 어떤 BERT를 사용해야 하는지 실험을 통해 증명하고자 한다. 제안한 방법에서 사용된 트랜스포머는 BERT를 합성한 모델이기 때문에 언어에 따라 특정 BERT를 정해야 한다. 모델을 학습하기 위해 원문의 언어에 맞는 BERT를 사용했다. 독일어-영어 번역 모델을 학습하기 위해 German BERT[1]를 사용했다. 이 모델은 Deepset[2]에서 오픈소스로 공개한 독일어로만 학습된 독일어 BERT이다. 그리고 영어-한국어 번역 모델에는 Base BERT를 사용했다. 특정 원문으로 학습된 BERT를 사용한 이유를 증명하기 위해 번역문(target language)의 언어까지 고려하여 Multilingual BERT를 사용했을 때, 번역 성능에 어떠한 영향을 주는지 확인한다.

4. 실험 및 결과

4.1 실험 데이터

아이디어를 반영한 모델을 학습하기 위해 독일어-영어 IWSLT14[10] 말뭉치를 사용했다. 그리고 AI Hub[3]에서 배포하는 번역용 영어-한국어 말뭉치를 사용했는데, 구어체와 문어체 성격을 갖고 있다. 독일어-영어 모델을 위해 160,239문장을 한 쌍으로 학습말뭉치로 사용했다. 영어-한국어 말뭉치는 구어체 20만 문장과 문어체 20만 문장을 추출하여 40만 문장을 학습말뭉치로 사용했다. 그 외 다른 데이터에 대한 자세한 내용은 표2에서 확인할 수 있다. 영어-한국어 말뭉치는 중복문장 제거와 문장 길이 필터링과 같은 전처리작업을 적용했다.

1) https://deepset.ai/german-bert
2) https://deepset.ai
3) https://www.aihub.or.kr/

표 2 학습에 사용된 말뭉치 정보

Model	독일어-영어	영어-한국어
	문장 수	
Train	160,239	400,000
Validation	7,283	36,966
Test	6,750	6,000

4.2 실험 결과

모델을 평가하기 위해 BLEU[11]를 사용했다. 평가 데이터는 단일 레퍼런스(reference)를 사용하고 4-gram BLEU를 사용하여 번역 성능을 평가했다. 제안한 방법의 실험 결과를 확인에 앞서 어텐션 선택방법의 결과와 BERT에 따른 성능 변화를 확인한다.

첫 번째로 표1은 제안 방법을 적용한 모델에 가중치를 다르게 하여 어텐션들을 선택한 방법을 적용하여 독일어-영어 번역 모델로 학습했을 때 BLEU 수치를 나타낸 것이다. 각각의 어텐션에 가중치의 총합이 1이 되는 두 개의 소수점들로 구성한 10가지 방법을 적용했을 때의 결과를 확인할 수 있다. 선행연구의 어텐션 선택방법을 사용했을 때 BLEU 수치가 36.00이 나왔지만 10가지의 방법 중 두 어텐션에 0.5씩 가중치를 부여한 모델은 BLEU 수치가 36.12가 나왔다. 이를 통해 무작위로 어텐션을 선택하는 방법보다 두 개의 어텐션을 모두 사용하는 방법이 더 효과적인 것을 알 수 있다. 각 어텐션의 정보를 모두 반영할 수 있으며 BERT의 문맥해석능력을 최대한 활용할 수 있어서 번역성능에 좋은 영향을 주는 것을 확인할 수 있다.

두 번째로 표3은 모델을 학습하는데 사용한 BERT를 원문 언어에 맞춘 이유를 증명한 실험 결과이다. 비교를 위해 각 번역 모델에 다른 2개의 BERT를 적용했다. 독일어-영어 모델에서는 독일어 전용 BERT를 적용한 것과 Multilingual BERT를 적용한 것을 비교하고 영어-한국어 모델에서는 Base BERT를 적용한 것과 Multilingual BERT를 적용한 것을 비교했다. 독일어-영어 모델에서 Multilingual BERT를 적용한 것보다 독일어 전용 BERT를 적용했을 때 더 높은 성능을 보였으며, 영어-한국어 모델에서도 유사한 결과를 볼 수 있었다. 이를 통해 원문 언어의 BERT를 사용하지 않았을 경우 BERT에 의해 구축된 사전에 원문 언어의 정보들이 많이 누락되는 현상이 발생하여 오히려 악영향을 주는 것을 확인할 수 있었다.

최종적으로 앞서 제안한 방법들에 이어 Cardinality residual connection이 모델을 학습하는데 미치는 효과를 확인하고자 한다. 표4에서 보시다시피 모든 번역에서

표 3 BERT모델에 따른 번역 성능 평가 결과

Transformer 모델	BERT모델	독일어-영어	영어-한국어
Baseline	×	34.4	10.8
BERT 합성 Transformer (Proposed)	Multilingual BERT	33.81	9.90
	German BERT	**36.12**	×
	Base BERT	×	**11.41**

표 4 잔차연결 구조 변경에 따른 모델 성능 평가 결과

Model	독일어-영어		영어-한국어	
	Epoch	BLEU	Epoch	BLEU
Baseline	47	34.4	56	10.8
Jinhua et al. (2020) [9]	96	35.67	82	10.43
Jinhua et al. (2020) [9]	147	35.92	157	11.39
Proposed	**84**	**36.12**	**82**	**11.41**

BERT를 합성한 트랜스포머의 성능이 baseline보다 뛰어나다는 것을 알 수 있다. 제안한 모델과 선행연구를 비교한 결과는 다음과 같다. 독일어-영어 모델에서는 선행연구의 BLEU 수치가 제안 모델보다 낮은 것을 볼 수 있다. 그리고 영어-한국어 모델에서는 똑같이 epoch가 82일 때 제안 모델의 BLEU 수치는 11.41로 선행연구보다 0.98이 상승한 것을 볼 수 있다. 또한, 선행연구의 epoch가 157일 때와 제안 모델의 epoch가 82일 때의 BLEU가 큰 차이가 없고 오히려 제안 모델의 epoch가 낮은 것을 확인할 수 있다. 이 결과를 통해 제안 방법을 사용하여 모델을 학습했을 때 학습 속도뿐만 아니라 번역 성능에도 효과적임을 증명할 수 있다. 번역 성능의 향상을 자세하고 쉽게 증명하기 위해 영어-한국어 모델들의 번역 결과를 표5에 정리하였다. 동등한 비교를 위해 epoch가 82인 제안 모델과 epoch가 157인 선행연구모델의 번역결과를 사용했다. 선행연구의 BLEU가 제안 모델보다 0.02 낮지만, 번역결과에서 큰 영향을 주지 않기 때문에 두 모델을 비교에 사용했다. 표5에서는 제안한 모델의 번역 결과가 선행연구보다 우수함을 보여주고 있다. 원문을 src로 표기하고, 정답번역문을 ref로 표기했다. (1)문장에서는 선행연구도 좋은 번역결과를 보여주었지만, 제안하는 모델이 문맥에 더 근접한 단어인 '권력'이란 단어를 생성해냈다. (3)문장에는 '사회복무요원'이란 단어를 제안하는 모델이 완벽하게 생성했음을 볼 수 있으며, 다른 문장들을 살펴보아도 제안하는 모델이 더 나은 번역결과를 생성했음을 확인할 수 있다. 이를 통해 제안 방법을 사용하면 번역속도뿐만 아니라 번역 성능도 향상됨을 알 수 있었다.

5. 결론

본 연구에서는 선행연구에서 제안한 방법을 보완하기 위해 cardinality residual connection을 적용한 새로운 모델을 제안했다. 트랜스포머의 인코더와 디코더의 셀프어텐션에 BERT를 각각 추가로 합성한 모델의 잔차연결을 수정하여 학습 속도의 향상과 번역 성능을 개선했다. 추가로 가중치를 다르게 부여하는 실험으로 어텐션을 선택하는 효과적인 방법을 제시하고 원문 언어에 맞는 BERT를 사용하는 이유를 설명하면서 성능이 향상되는 모델을 학습할 수 있는 방법을 증명했다. 추후에는 본 연구를 기반으로 입력되는 원문에 따라 고정이 아닌 동적으로 가중치를 부여하여 어텐션을 선택할 수 있는 방법에 대해 연구를 진행하고자 한다.

표 5 제안하는 방법을 사용했을 때 번역문 생성 결과

(1)	src	There are voices of talk that the Korean church is losing its power.
	ref	한국교회가 힘을 잃어가고 있다는 볼멘소리가 터져 나오고 있다.
	jinhua	한국교회가 힘을 잃고 있다는 목소리가 나오고 있다.
	Proposed	한국교회가 **권력**을 잃고 있다는 얘기가 나오고 있다.
(2)	src	Then, apply the toothpaste to a cotton swab or rag, and gently buff the scratched area.
	ref	그다음, 치약을 면봉이나 천에 묻혀 스크래치가 난 곳에 부드럽게 문질러줍니다.
	jinhua	이어 칫솔은 면봉이나 라크에 바르고 긁힌 부위를 부드럽게 닦아줍니다.
	Proposed	이어 **치약**을 면봉이나 랩에 바르고 긁힌 부위를 부드럽게 닦아줍니다.
(3)	src	A social service worker caught the man who was illegally taking pictures in the bathroom of a subway station after a physical fight and handed him over to police.
	ref	사회복무요원이 지하철역 화장실에서 불법촬영범을 몸싸움 끝에 붙잡아 경찰에 넘겼다.
	jinhua	사회서비스원이 몸싸움 끝에 지하철역 화장실에서 불법촬영을 하던 남성을 붙잡아 경찰에 인계했다.
	Proposed	**사회복무요원**이 몸싸움 끝에 지하철역 화장실에서 불법촬영을 하던 남성을 붙잡아 경찰에 인계했다.
(4)	src	At the Icheon Ginseng Festival held over the weekend, more than 170,000 people visited and sold ginseng sales of 1.4 billion won.
	ref	지난 주말 열린 이천 인삼 축제에는 17만 명이 넘는 사람들이 찾아 14억 원의 인삼판매 실적을 올렸습니다.
	jinhua	주말 동안 진행된 이천진삼페스티벌에서는 1만 7천여 명이 방문해 인삼판매량을 팔았다.
	Proposed	주말 동안 열린 **이천인삼페스티벌**에서는 **17만 명** 이상이 방문해 인삼판매액을 넘어섰다.
(5)	src	The police officer arrests her in the middle of the night for speeding.
	ref	경찰관은 그녀를 속도위반으로 한밤중에 체포해요.
	jinhua	경찰관이 한밤중에 과속해서 그녀에게 체포합니다.
	Proposed	경찰관은 그녀가 한밤중에 과속해서 체포해요.
(6)	src	The owner of the cafe is me, why?
	ref	그 카페의 주인은 나인데, 왜 그러시죠?
	jinhua	카페 주인이 저야, 왜 저인가요?
	Proposed	그 카페 주인이 나인데 왜 그런가요?

사 사

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)

참고문헌

[1] VASWANI, Ashish, et al., Attention is all you need, Advances in Neural Information Processing Systems, pp. 6000-6010, 2017

[2] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473, 2014

[3] J. Devlin, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2018

[4] XIE, Saining, et al., Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, p. 1492-1500, 2017

[5] HE, Kaiming, et al., Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, p. 770-778, 2016

[6] KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E., Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, p. 1097-1105, 2012

[7] CLINCHANT, Stéphane; JUNG, Kweon Woo; NIKOULINA, Vassilina, On the use of BERT for Neural Machine Translation, arXiv preprint arXiv:1909.12744, 2019

[8] YANG, Jiacheng, et al., Towards making the most of bert in neural machine translation, arXiv preprint arXiv:1908.05672, 2019

[9] ZHU, Jinhua, et al., Incorporating bert into neural machine translation, arXiv preprint arXiv:2002.06823, 2020

[10] Mauro Cettolo, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, and Marcello Federico, Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In Proceedings of IWSLT, 2014

[11] PAPINENI, Kishore, et al., BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, p. 311-318, 2002

[12] ZHANG, Jiacheng, et al. Improving the transformer translation model with document-level context, arXiv preprint arXiv:1810.03581, 2018

[13] 신재훈 외 3명., 다중 인코더 Transformer 기반 번역문 자동 사후 교정 모델의 디코더 구조 연구, 한국정보과학회, p. 634-636, 2019