MASS를 이용한 영어-한국어 신경망 기계 번역

정영준*○, 박천음*, 이창기*, 김준석**

강원대학교 컴퓨터과학과*, 현대자동차 AIR Lab**
{kongjun, parkce, leeck}@kangwon.ac.kr, junseok.kim@hyundai.com

English-Korean Neural Machine Translation using MASS

Young-Jun Jung*○ Cheon-Eum Park* Chang-Ki Lee* Jun-Seok Kim**
Kangwon National University*, Hyundai Motor Company AIR Lab**

요 약

신경망 기계 번역(Neural Machine Translation)은 주로 지도 학습(Supervised learning)을 이용한 End-to-end 방식의 연구가 이루어지고 있다. 그러나 지도 학습 방법은 데이터가 부족한 경우에는 낮은 성능을 보이기 때문에 BERT와 같은 대량의 단일 언어 데이터로 사전학습(Pre-training)을 한 후에 미세조정(Fine-tuning)을 하는 Transfer learning 방법이 자연어 처리 분야에서 주로 연구되고 있다. 최근에 발표된 MASS 모델은 언어 생성 작업을 위한 사전학습 방법을 통해 기계 번역과 문서 요약에서 높은 성능을 보였다. 본 논문에서는 영어-한국어 기계 번역 성능 향상을 위해 MASS 모델을 신경망 기계 번역에 적용하였다. 실험 결과 MASS 모델을 이용한 영어-한국어 기계 번역 모델의 성능이 기존 모델들보다 좋은 성능을 보였다.

주제어: Neural Machine Translation, 기계 번역, 사전학습, MASS

1. 서론

기계 번역은 주로 지도 학습(supervised learning)을 이용한 신경망 기계 번역(Neural Machine Translation) 모델에 대한 연구가 진행되었다[1,2]. 신경망 기계 번역 모델은 문장을 입력받아 다른 언어로 자동 번역해주는 End-to-end 방식의 모델로, 병렬 코퍼스를 사용하여 학습한다. 일반적으로 지도 학습 방법은 데이터가 부족할 경우에는 낮은 성능을 보인다.

최근에는 학습 데이터가 부족할 때 대량의 단일 언어 데이터를 사용하는 사전학습(pre-training) 방법이 자연어 처리 분야에서 많은 관심을 끌고 있다[3,4]. BERT(Bidirectional Encoder Representations from Transformers)[4]는 대량의 단일 언어 데이터로 Masked LM(Masked Language Model)과 다음 문장 예측을 통해 양방향 인코더 표현을 사전학습하여 자연어 처리 작업에서 좋은 성능을 보이고 있다.

신경망 기계 번역 작업에서도 사전학습을 사용한 XLM(Cross-lingual Language Model)[5]과 MASS(Masked Sequence to Sequence pre-training)[6]가 주목받고 있다. 그 중 MASS 모델은 인코더에서 마스크(mask) 되지 않은 토큰의 의미를 이해하고 표현할 수 있게 하여, 디코더가 인코더에서 마스크 된 토큰을 잘 예측할 수 있도록 하였다. 또한 디코더의 입력 토큰을 마스킹하는 방법을 통해 이전 토큰보다 인코더의 표현에 더 많이 의존하게 만들어 인코더와 디코더 간의 공동 학습이 더 잘되도록 하였다.

본 논문에서는 다양한 언어 생성 작업에서 좋은 성능을 보이고 있는 MASS 모델을 영어-한국어 데이터로 사전학습하고, 사전학습된 MASS 모델을 기반으로 영어-한국어 신경망 기계 번역에 적용하여 성능 향상을 보인다.

2. 관련 연구

[1]에서 제안된 모델은 어텐션 메커니즘(attention mechanism) 기반 인코더-디코더(encoder-decoder) 모델이다. 어텐션 메커니즘은 출력 단어를 예측하기 위해 집중해서 봐야 할 입력 문장의 단어에 대한 어텐션 가중치를 결정한다.

[2]에서는 순환(recurrence)과 합성곱(convolution)을 없앤 단순한 어텐션 메커니즘 기반 인코더-디코더 구조의 트랜스포머(Transformer) 모델을 제안하였다. 트랜스포머는 멀티헤드 셀프어텐션(multi-head self-attention)을 사용하며, 기계 번역에서 좋은 성능을 보이고 있다.

최근에는 대용량 코퍼스를 이용해 사전학습하고, 다양한 자연어 처리 작업에 미세조정(fine-tuning)하는 방법이 연구되고 있다.

GPT(Generative Pre-Trainging)[3]는 단방향 트랜스포머 디코더로 구성된 모델로, 이전 토큰으로 다음 토큰을 예측하는 일반적인 언어 모델링을 사용해 사전학습한다.

BERT[4]는 양방향 트랜스포머 인코더로 구성된 모델이다. 문장 내 임의의 단어를 마스킹하고 예측하는 Masked LM과 다음 문장 예측을 기반으로 모델을 사전학습한다. 사전학습된 BERT 모델은 다른 자연어 처리 작업에 미세조정하는 방법으로 적용되며, 이는 다양한 자연어 처리 작업에서 높은 성능을 보이고 있다.

XLM[5]은 BERT와 유사한 방법으로 인코더와 디코더를 사전학습하고, 이를 기계 번역에 적용하여 높은 성능을 보였다. 인코더와 디코더는 각각 사전학습되지만, 인코더-디코더 어텐션 메커니즘은 사전학습 되지 않는다.

MASS[6]는 언어 생성을 위해 인코더와 디코더를 공동으로 사전학습하는 모델이다. 어텐션 메커니즘도 같이 사전학습 되기 때문에, 기계 번역과 같은 Sequence-to-sequence 작업에서 좋은 성능을 보이고 있다.

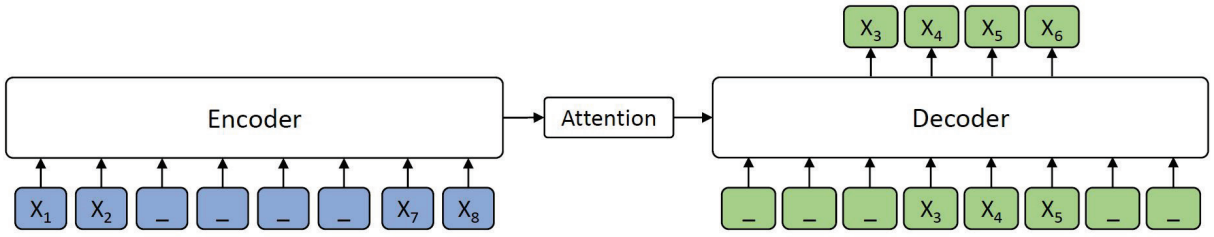

그림 1. MASS의 인코더-디코더 모델 구조. '_' 토큰은 마스크 기호 [M]을 나타낸다. [6]

3. MASS

MASS는 기존 방법(BERT나 GPT)과 다른 사전학습 방법을 사용한다. 입력 문장 x가 주어지면 위치 u에서 v까지 토큰이 마스크 되며, $0 < u < v < m$ 이다. 여기서 m은 문장 x의 토큰 개수이다. 위치 u에서 v까지 마스크 되는 토큰 수는 $k = v - u + 1$ 로 나타낼 수 있다. 마스크 된 토큰은 마스크 기호 [M]으로 대체되며, 마스크 된 문장의 길이는 바뀌지 않고 입력 문장과 동일한 길이를 가진다.

사전학습은 마스크 된 문장 x를 인코더의 입력으로 사용하며, 디코더에서 위치 u에서 v까지 마스크 된 토큰을 예측하는 Sequence-to-sequence 모델을 학습한다. 그림 1은 MASS 모델 구조의 예를 보여준다. 인코더에 $x_3x_4x_5x_6$ 토큰이 마스크 된 8개의 토큰을 가지는 문장이 입력되고, 디코더 입력으로는 위치 4-6의 토큰 $x_3x_4x_5$가 주어진다. 모델은 인코더의 입력 문장에서 마스크 된 토큰 $x_3x_4x_5x_6$만 예측한다. 디코더에서 위치 4-6을 제외한 다른 위치에 대한 입력으로는 특수 마스크 기호 [M]을 사용한다 (위치 1-3, 7-8).

MASS는 언어 생성 작업을 위해 인코더와 디코더를 공동으로 사전학습하도록 설계되었다. Sequence-to-sequence 모델을 통해 마스크 된 토큰만 예측하게 함으로써, 인코더가 마스크 되지 않은 토큰의 의미를 이해하도록 하고, 디코더는 인코더로부터 유용한 정보를 추출하도록 한다. 디코더에서는 연속적인 토큰을 예측하여 기존의 Masked LM 보다 더 좋은 언어 모델링을 할 수 있다. 또한 마스크 되지 않은 디코더의 입력 토큰을 마스크 함으로써, 디코더는 이전 토큰의 정보를 활용하는 대신 인코더에서 더 유용한 정보를 추출하도록 한다.

기존의 언어 모델링과 MASS의 차이를 마스크 된 토큰의 길이를 나타내는 하이퍼파라미터(hyperparameter) k 값이 다른 경우라고 생각할 수 있다. $k = 1$인 경우, 입력 문장의 하나의 토큰만이 마스크 되며, 이는 BERT에서 사용되는 Masked LM과 같다. $k = m$인 경우, 인코더의 모든 토큰이 마스크 되고, 디코더는 이전 토큰으로 다음 토큰을 예측하는 일반적인 언어 모델링(이 경우 GPT)이 된다.

본 논문에서는 신경망 기계 번역에서 좋은 성능을 보이고 있는 인코더-디코더 모델인 트랜스포머 모델을 사용하여 MASS 사전학습 및 미세조정 실험을 진행하였다.

4. 실험

본 논문에서는 영어-한국어 언어 쌍의 단일 언어 데이터에 대해 MASS 모델을 사전학습한다. 학습 과정에서 입력 언어와 출력 언어를 식별하기 위해 인코더와 디코더 입력 문장의 각 토큰에 언어 임베딩을 더하여 End-to-end로 학습한다.

학습에 사용한 트랜스포머 모델의 하이퍼파라미터는 다음과 같다. 인코더와 디코더의 레이어 수는 6, 임베딩과 히든 레이어 차원 수는 1024, 헤드 수는 8, 피드포워드(feed-forward)의 차원 수는 4096이다. 실험에는 XLM 코드베이스를 기반으로 구현된 MASS를 사용하였다[1].

사전학습에 사용한 단일 언어 데이터는 영어 WMT 뉴스 크롤링 데이터 500만 문장, 한국어 세종 코퍼스 약 200만 문장을 사용하였다. 영어는 Moses[2], 한국어는 Mecab-ko[3]를 이용하여 문장을 토크나이즈(tokenize) 하였다. 어휘는 BPE(Byte Pair Encoding)[7]를 사용해 영어와 한국어에 대한 60,000개의 공유 어휘(shared vocabulary)를 구성하여 사용하였다.

사전학습에서는 임의의 시작 위치 u를 가지는 연속적인 토큰을 [M]으로 교체하며, 마스크 길이 k를 문장의 총 토큰 수의 50%로 정하고 마스킹한다. BERT에서의 마스크 방법과 유사하게, 그중 80%는 마스크, 10%는 임의의 토큰으로 변경하고, 나머지 10%는 변경하지 않고 그대로 사용한다.

병렬 데이터는 IWSLT2017 TED 영어-한국어 데이터를 사용하였다. 이 데이터는 TED 번역 데이터로, 다양한 도메인을 가지고 있는 구어체 문장으로 다른 데이터보다 번역 난이도가 높다. 사전학습한 모델을 기반으로 학습 데이터 약 23만 문장 쌍을 이용하여 미세조정하였다. 테스트 데이터는 test2016과 test2017을 사용하였으며, 각각 1,143 문장, 1,429 문장으로 구성되고, BLEU를 사용하여 테스트 데이터의 성능을 평가하였다.

표 1은 MASS를 이용한 영어-한국어 번역 실험에 대한 성능을 나타낸 것이다. [8] 모델은 본 논문에서 미세조정에 사용한 데이터와 같은 IWSLT2017 TED 영어-한국어

[1] https://github.com/microsoft/MASS
[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
[3] https://bitbucket.org/eunjeon/mecab-ko-dic/src/master

데이터를 사용하여 학습한 모델의 성능이다. Baseline 모델은 사전학습을 진행하지 않고 en-ko, ko-en 각각 언어 방향에 대하여 트랜스포머 모델을 사용해 번역 작업을 학습한 결과이다. en→ko, ko→en 는 단방향, en↔ko 는 양방향, en↔ko + BT 는 양방향에 역번역(back-translation)[9]을 사용한 미세조정 학습 결과이다. Baseline과 비교하여 사전학습을 사용하였을 때 모두 1 BLEU 이상 향상되었다. 미세조정에 사용한 병렬 데이터는 사전학습에 사용한 단일 언어 데이터와 다른 도메인을 가지고 있지만, 사전학습을 통해 성능이 향상된 것으로 보인다. 양방향 학습으로 미세조정한 모델은 단방향으로 미세조정한 모델과 비교하여 test2016에서 0.19, test2017에서 0.11 BLEU 향상되었고, ko-en 태스크에서는 test2016에서 0.03, test2017에서 0.02 BLEU 하락하며, 단방향으로 미세조정한 모델과 큰 성능 차이를 보이지 않았다. 역번역을 포함하는 양방향 모델의 경우에는 사전학습에 사용한 단일 언어 데이터를 역번역에 사용하고, 병렬 데이터를 번역 작업에 사용하였다. 단방향이나 양방향으로 미세조정하였을 때 보다 양방향에 역번역을 추가로 사용한 모델이 en-ko 태스크에서 test2016에서 17.59, test2017에서 15.71 BLEU를 보였고, ko-en 태스크에서는 test2016에서 17.11, test2017에서 15.34 BLEU를 보이며 en-ko, ko-en 모두에서 성능이 향상되었다. 이와 같은 이유는 역번역을 통해 단일 언어 데이터를 활용하여 적은 병렬 데이터를 보완해 성능이 향상된 것으로 보인다. 실험 결과, 사전학습한 MASS 모델을 기반으로 미세조정한 모델이 기존 기계 번역 모델에 비해 모두 높은 성능을 보였다.

표 1. 영어-한국어 번역 실험 결과

미세조정	en-ko		ko-en	
	test2016	test2017	test2016	test2017
허광호[8]	-	-	10.09	8.98
Baseline	15.33	14.12	14.23	12.48
en→ko	16.43	15.34	-	-
ko→en	-	-	15.57	13.50
en↔ko	16.62	15.45	15.54	13.48
en↔ko+BT	**17.59**	**15.71**	**17.11**	**15.34**

5. 결론

본 논문에서는 Sequence-to-sequence 작업을 위한 사전학습 방법을 사용하는 MASS 모델을 소개하고, 이를 한국어-영어 데이터로 사전학습하여 신경망 기계 번역 작업에 적용하였다.

실험 결과, 사전학습한 MASS 모델을 기계 번역 작업을 위해 미세조정하여 기존의 모델보다 성능이 향상된 것을 보였다. 또한 평가에 사용되는 데이터와 다른 도메인으로 사전학습을 하였음에도 불구하고, MASS 모델이 기존의 모델보다 높은 성능을 낼 수 있음을 보였다. 더 좋은 품질의 대용량 데이터를 사전학습에 사용하면 지금보다 성능이 개선될 것으로 예상된다.

향후 연구로는 사전학습 및 미세조정을 사용하는 개선된 모델을 연구하고, 높은 성능을 위해 좋은 품질을 가지는 학습 데이터 정제 및 구축을 시도할 예정이다. 또한 한국어 데이터에 대해 기계 번역 작업 이외에도 문서 요약과 같은 다른 언어 생성 작업에서 MASS 모델을 적용할 예정이다.

감사의 글

이 논문은 현대 자동차 AIR Lab의 “딥러닝 기반 기계 번역 기술 연구” 과제의 지원을 받아 연구되었음

참고문헌

[1] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, ICLR, 2015.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, NIPS, 2017.

[3] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, Improving Language Understanding by Generative Pre-Training, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805, 2018.

[5] Guillaume Lample, Alexis Conneau, Cross-lingual Language Model Pretraining, arXiv:1901.07291, 2019.

[6] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu, MASS: Masked Sequence to Sequence Pre-training for Language Generation, ICML, 2019.

[7] Rico Sennrich, Barry Haddow, Alexandra Birch, Neural Machine Translation of Rare Words with Subword Units, ACL, 2016.

[8] 허광호, 고영중, 서정연, 저-자원 언어의 번역 향상을 위한 다중-언어 기계번역, 한국정보과학회 학술발표논문집, 2019.

[9] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, Marc’aurelio Ranzato, Phrase-Based & Neural Unsupervised Machine Translation, EMNLP, 2018.