단어 정렬을 이용한 한국어-영어 비자기회귀 신경망 기계 번역

정영준○, 이창기

강원대학교 빅데이터메디컬융합학과
{kongjun, leeck}@kangwon.ac.kr

Korean-English Non-Autoregressive Neural Machine Translation using Word Alignment

Young-Jun Jung○, Chang-Ki Lee
Department of Big Data Medical Convergence, Kangwon National University

요 약

기계 번역(machine translation)은 자연 언어로 된 텍스트를 다른 언어로 자동 번역 하는 기술로, 최근에는 주로 신경망 기계 번역(Neural Machine Translation) 모델에 대한 연구가 진행되었다. 신경망 기계 번역은 일반적으로 자기회귀(autoregressive) 모델을 이용하며 기계 번역에서 좋은 성능을 보이지만, 병렬화할 수 없어 디코딩 속도가 느린 문제가 있다. 비자기회귀(non-autoregressive) 모델은 단어를 독립적으로 생성하며 병렬 계산이 가능해 자기회귀 모델에 비해 디코딩 속도가 상당히 빠른 장점이 있지만, 멀티모달리티(multimodality) 문제가 발생할 수 있다. 본 논문에서는 단어 정렬(word alignment)을 이용한 비자기회귀 신경망 기계 번역 모델을 제안하고, 제안한 모델을 한국어-영어 기계 번역에 적용하여 단어 정렬 정보가 어순이 다른 언어 간의 번역 성능 개선과 멀티모달리티 문제를 완화하는 데 도움이 됨을 보인다.

주제어: 신경망 기계 번역, 비자기회귀, 단어 정렬

1. 서론

기계 번역(machine translation)은 자연 언어로 된 텍스트를 다른 언어로 자동 번역 하는 기술로, 최근에는 주로 신경망 기계 번역(Neural Machine Translation) 모델에 대한 연구가 진행되었다[1,2]. 신경망 기계 번역은 일반적으로 자기회귀(autoregressive) 특성을 갖는 디코더(decoder)가 포함된 sequence-to-sequence 모델을 이용하며, 기존의 통계적 기계 번역(Statistical Machine Translation) 모델에 비해 좋은 성능을 보여준다. 하지만, 이전에 생성된 단어를 조건으로 각 단어를 생성하는 자기회귀 디코더 때문에 병렬화할 수 없어 디코딩 속도가 느린 문제가 있다.

최근에는 자기회귀 모델의 느린 디코딩 속도를 개선하기 위한 비자기회귀(non-autoregressive) 모델이 연구되고 있다[3,4]. 비자기회귀 모델은 일반적으로 이전에 생성된 단어 대신에 소스(source) 단어 표현을 복사하여 디코더의 입력으로 사용한다. 따라서 모든 타겟(target) 단어를 동시에 독립적으로 생성할 수 있으므로 병렬 계산이 가능하여 자기회귀 모델보다 디코딩 속도가 상당히 빠른 장점이 있다. 하지만, 비자기회귀 모델은 타겟 단어 간의 종속성이 없기 때문에, 중복, 누락 또는 잘못된 단어를 생성하는 멀티모달리티(multimodality) 문제가 발생할 수 있고, 한국어-영어와 같이 어순이 크게 다른 언어 간 번역에 어려움이 있다.

본 논문에서는 비자기회귀 모델에서 어순이 다른 언어 간의 번역 성능이 낮은 문제를 해결하기 위해 단어 정렬(word alignment)을 이용한 비자기회귀 신경망 기계 번역 모델을 제안하고, 제안한 모델을 한국어-영어 기계 번역에 적용하여 번역 모델의 성능이 개선될 수 있음을 보인다.

2. 관련 연구

[1]에서 제안된 모델은 어텐션 메커니즘(attention mechanism) 기반 인코더-디코더(encoder-decoder) 모델이다. 어텐션 메커니즘은 타겟 단어를 예측하기 위해 집중해서 봐야 할 소스 문장의 단어에 대한 어텐션 가중치를 결정한다. [2]에서는 멀티헤드 셀프어텐션(multi-head self-attention)을 사용하는 어텐션 메커니즘 기반 인코더-디코더 구조의 트랜스포머(Transformer) 모델을 제안하였다.

최근 비자기회귀 신경망 기계 번역 연구는 주로 멀티모달리티 문제를 해결하기 위한 연구가 진행되었다. [3]은 트랜스포머 기반의 비자기회귀 신경망 기계 번역 모

· 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.NRF2021R1F1A106440311, 딥러닝 기반의 한국어 텍스트 스타일 변환 기술 연구).
· 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능 진화형 Wise QA 플랫폼 기술 개발).

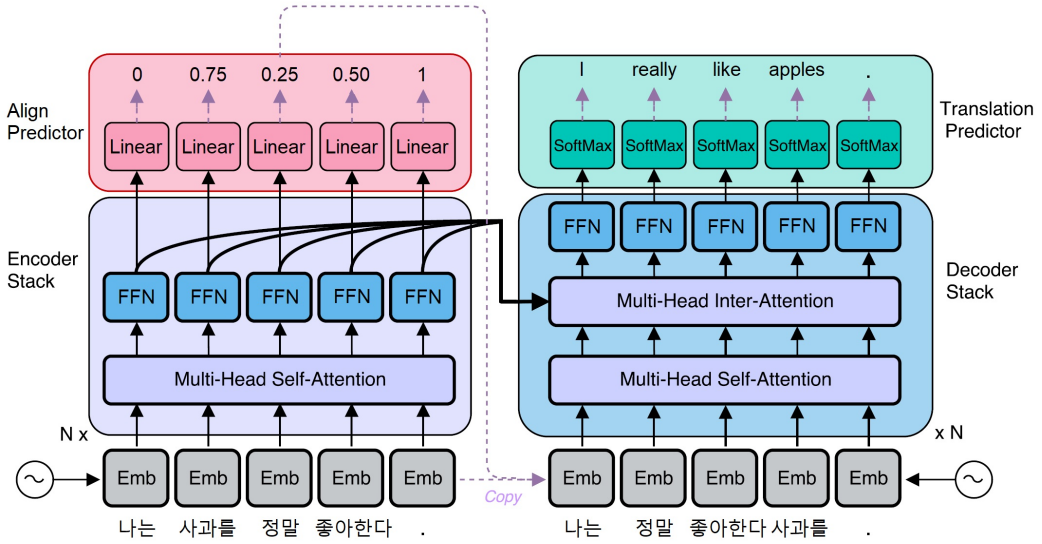

그림 1 단어 정렬을 이용한 비자기회귀 신경망 기계 번역 모델

델을 제안하였다. 인코더에 퍼틸리티(fertility) 예측기를 추가해 소스 단어를 0회 이상 복사하여 디코더의 입력으로 사용하였고, 퍼틸리티 값의 합으로 출력 길이를 결정하였다. [4]는 비자기회귀 모델의 멀티모달리티 문제를 완화하는 것을 목표로 하는 ReorderNAT 모델을 제안하였다. 재정렬 모듈을 추가로 사용하여 디코딩을 도와주는 재정렬 정보를 명시적으로 모델링하였고, 멀티모달리티 문제를 효과적으로 완화하였다.

3. 비자기회귀 신경망 기계 번역 모델

비자기회귀 신경망 기계 번역 모델은 소스 문장 $\mathrm{X} = \{x_1, \ldots, x_n\}$와 타겟 문장 $\mathrm{Y} = \{y_1, \ldots, y_m\}$가 주어지면, 조건부 독립(conditional independence) 타겟 단어 확률의 곱으로 X에서 Y로의 생성 확률을 식 1과 같이 모델링한다.

$$P(\mathrm{Y}|\mathrm{X}) = \prod_{i=1}^{m} P(y_i|\mathrm{X}) \quad (1)$$

비자기회귀 모델은 일반적으로 이전에 생성된 단어 대신에 소스 단어 표현을 복사하여 디코더의 입력으로 사용해 타겟 단어 간의 종속성을 제거한다. 따라서 디코딩 시간에 독립적인 최대가능도(maximum likelihood)로 모든 타겟 단어를 병렬적으로 생성이 가능해 디코딩 속도가 매우 빠르다. 하지만, 순차적 종속성이 없기 때문에 멀티모달리티 문제가 발생할 수 있어 성능이 떨어질 수 있는 문제가 있고, 어순이 다른 언어 간 번역에 어려움이 있다. 결과적으로 소스 문장에서 타겟 문장으로의 복잡한 번역 관계를 효과적으로 학습할 수 없어 번역 성능이 저하될 수 있다. 따라서 본 논문에서는 어순이 다른 언어 간 번역 성능 개선에 도움이 될 수 있는 단어 정렬 정보를 이용한 비자기회귀 모델을 제안한다.

그림 1은 단어 정렬을 이용한 비자기회귀 신경망 기계 번역 모델을 보여준다. [3]에서 사용한 소스 문장 단어들의 개수만을 조정하는 퍼틸리티 예측기 대신에, 어순이 다른 언어 간 번역에 효과적이기 위해 정렬 예측기를 사용하였다. 정렬 정보를 모델링하기 위해 인코더에 정렬 예측기를 추가하고, 식 2와 같이 정렬 정보를 모델링한다.

$$\mathrm{Z} = \sigma(W_a \mathrm{H} + b_a) \quad (2)$$

H는 소스 문장을 인코딩한 결과이고, Z는 디코더의 입력으로 사용하는 소스 문장 단어에 대한 상대적인 위치 정보를 나타낸다. Z는 선형(linear) 레이어(layer)를 사용하여 계산되고, 시그모이드(sigmoid) 함수를 통해 0에서 1까지의 값을 갖게 된다. 정렬 예측기는 단어 정렬 정보를 이용하여 소스 문장을 재정렬하는 방법을 학습하게 된다. [4]에서는 재정렬을 위해 재정렬 모듈을 추가로 사용하였지만, 본 논문에서 제안하는 모델은 인코더에 정렬 예측기만을 추가하여 소스 문장을 재정렬하였다.

디코더에서는 정렬 위치 $\mathrm{Z} = \{z_1, \ldots, z_n\}$를 통해 재정렬된 소스 문장 X를 입력받아 비자기회귀 디코더로 Y를 생성한다. 최종적으로 X에서 Y로의 생성 확률은 식 3과 같다.

$$P(\mathrm{Y}|\mathrm{Z},\mathrm{X}) = \prod_{i=1}^{m} P(y_i|\mathrm{Z},\mathrm{X}) \quad (3)$$

디코딩 과정에서는 타겟 문장의 길이를 결정해야 한다. 이를 위해 길이 예측기를 사용하여 학습 과정에서 타겟 문장의 길이에 대한 정보를 학습한다. 길이 예측은 식 4와 같다.

$$P(\mathrm{L}|\mathrm{X}) = \mathrm{softmax}\big(W_l(\mathrm{mean}(\mathrm{H})) + b_l\big) \quad (4)$$

표 2 비자기회귀 신경망 기계 번역 결과 예제

Source [Align Predict]	서울 강서구 공무원들이 행정 경험을 공유하고 정책 발전방안을 찾는다. [서울(0.09) 강서구(0.07) 공무원들이(0.06) 행정(0.44) 경험을(0.45) 공유하고(0.36) 정책(0.85) 발전(0.85) 방안을(0.84) 찾(0.66) 는다.(0.80)]
	언론인의 역할에 대해서 많은 혼란이 있는 시기다. [언론(0.79) 인의(0.34) 역할에(0.80) 대해서(0.52) 많은(0.31) 혼란이(0.42) 있는(0.39) 시(0.64) 기다.(0.37)]
Reference	Gangseo-gu officials in Seoul will share administrative experiences and seek ways to develop policies.
	This is a time of great confusion about the role of journalists.
Non-Autoregressive	Officials officials in Gangseo-gu, Seoul, share administrative experiences and find ways policy develop measures.
	It is a time when many confusion confusion the role of journalists.
Non-Autoregressive (+Word Alignment)	Public in Gangseo-gu, Seoul, will share administrative experiences and find for policy development measures.
	It is a time of many confusion about the role of journalists.

길이 예측기는 인코딩된 소스 문장 H를 평균 풀링(mean pooling)하고, 선형 레이어와 소프트맥스(softmax) 함수를 통해 타겟 문장의 길이 L을 예측한다.

4. 실험 및 결과

단어 정렬을 이용한 비자기회귀 신경망 기계 번역 모델은 트랜스포머 모델을 기반으로 구현하였다. 실험에 사용한 모델의 하이퍼파라미터(hyperparameter)는 다음과 같다. 인코더와 디코더 레이어 수는 6, 임베딩과 히든 레이어 차원 수는 512, 헤드(head) 수는 8, 피드포워드 네트워크(feed-forward network)의 차원 수는 512이다. 학습 데이터는 AIHub[1] 한국어-영어 번역 말뭉치 중 뉴스 분야 약 80만 문장을 사용하였으며, 그중 3200문장을 각각 절반씩 개발 데이터와 평가 데이터로 나누어 실험에 사용하였다. 토크나이즈(tokenize)는 BPE(Byte Pair Encoding)[5]를 이용하였고, 어휘 사전은 한국어와 영어에 대한 32000 크기의 공유 어휘(shared vocabulary)를 사용하였다. 학습에 사용한 단어 정렬은 한국어-영어 학습데이터에 대해 fast_align[2]을 이용하여 만든 단어 정렬 정보를 사용하였다. 기계 번역 모델의 성능 평가 지표는 BLEU(Bilingual Evaluation Understudy)를 사용하여 평가 데이터로 모델의 성능을 측정하였다.

[1] https://aihub.or.kr
[2] https://github.com/clab/fast_align

표 1 비자기회귀 신경망 기계 번역 실험 결과

Model	BLEU	Speedup
Non-Autoregressive	9.93	18.20×
(+init encoder +KD)	18.67	
Non-Autoregressive (+WA)	10.62	18.16×
(+init encoder +KD)	19.08	
Autoregressive (b=1)	31.36	1.00×

표 1은 비자기회귀 신경망 기계 번역 실험 결과를 보여준다. 모든 모델의 디코딩에 사용한 배치(batch) 크기는 1이고, 자기회귀 모델의 빔(beam) 크기는 1로 설정하였다. 비자기회귀 모델은 추가로 성능 개선을 위해 인코더를 자기회귀 모델 인코더의 가중치로 초기화하고, 지식 증류(knowledge distillation)[6]를 사용하였다. 실험 결과, 단어 정렬을 이용하였을 때 비자기회귀 모델의 성능이 향상되었다. 이는 한국어와 영어는 어순이 다른 언어이기 때문에, 학습 과정에서 정렬 정보를 학습한 정렬 예측기의 재정렬이 유의미한 영향을 끼친 것으로 보인다. 표 2는 비자기회귀 신경망 기계 번역 결과 예제를 보여준다. 노란색으로 표시한 것은 중복되어 생성된 단어들을 나타내는데, 단어 정렬을 이용한 모델이 예측한 결과에서 중복으로 생성한 단어가 적게 나타난 것으로 보인다. 이는 단어 정렬 정보가 멀티모달리티 문제를 완화하는 데 도움이 되어 번역 성능이 개선된 것으로 보인다. 비자기회귀 모델 간의 디코딩 속도는 유의미한 차이

를 보이지 않는데, 추가된 단어 정렬기가 디코딩 속도에 큰 영향을 미치지 않는 것을 보여준다. 자기회귀와 비자기회귀 모델의 디코딩 속도는 약 18배 이상 차이가 나는 결과를 보이며 비자기회귀 모델이 상당히 빠른 디코딩 속도를 갖는 것을 보여준다.

5. 결론

본 논문에서는 단어 정렬을 이용한 비자기회귀 신경망 기계 번역 모델을 제안하고, 한국어-영어 기계 번역에 적용하였다. 실험 결과, 단어 정렬 정보를 이용한 비자기회귀 모델이 어순이 다른 언어 간의 번역 성능 개선과 멀티모달리티 문제를 완화하는 데 도움이 됨을 보였다. 추후 연구에서는 비자기회귀 모델의 성능 향상을 위해 사용되는 기존의 방법론들을 적용하여 모델을 개선할 예정이다.

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] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, Richard Socher, "Non-Autoregressive Neural Machine Translation," *ICLR*, 2018.

[4] Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, "Guiding Non-Autoregressive Neural Machine Translation Decoding with Reordering Information," *AAAI*, 2021.

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

[6] Yoon Kim, Alexander M. Rush, "Sequence-Level Knowledge Distillation," *EMNLP*, 2016.