|  |
| --- |
| **신경망 기반 기계 번역을 위한 역-번역을 이용한**  **한영 병렬 코퍼스 확장**  **Expanding Korean/English Parallel Corpora using Back-translation for Neural Machine Translation**  **요 약**  최근 제안된 순환 신경망 기반 Encoder-Decoder 모델은 기계번역에서 좋은 성능을 보인다. 하지만 이는 대량의 병렬 코퍼스를 전제로 하며 병렬 코퍼스가 소량일 경우 데이터 희소성 문제가 발생하며 번역의 품질은 다소 제한적이다. 본 논문에서는 기계번역의 이러한 문제를 해결하기 위하여 단일-언어(Monolingual) 데이터를 학습과정에 사용하였다. 즉, 역-번역(Back-translation)을 이용하여 단일-언어 데이터를 가상 병렬(Pseudo Parallel) 데이터로 변환하는 방식으로 기존 병렬 코퍼스를 확장하여 번역 모델을 학습시켰다. 역-번역 방법을 이용하여 영-한 번역 실험을 수행한 결과 +4.7 BLEU 점수의 성능 향상을 보였다.  주제어: 기계번역, 역-번역, 가상 병렬 코퍼스 |

**1. 서론**

최근 신경망을 이용한 기계번역(NMT: Neural Machine Translation) 모델이 영어-불어, 영어-독일어의 대용량 코퍼스에서 우수한 성능을 보였다[1,2]. NMT 모델이 매력적인 이유는 End-to-End 방식으로 번역할 수 있기 때문이다. 즉, 기존의 통계 기반 기계번역 모델은 단어 혹은 구(Phrase) 기반 번역 후 재배치하는 방식으로 번역하는 반면, NMT 모델은 Encoder 모델로 전체 소스 문장을 고정 사이즈 벡터로 표현한 후 Decoder 모델로 전체 타깃 문장을 번역하여 출력한다. 또한, End-to-End 방식으로 번역하는 과정에서 발생하는 장기 종속성(Long-term Dependency) 문제를 Attention 매커니즘을 이용하여 효과적으로 해결할 수 있다[3,4]. 하지만 NMT 모델의 이러한 장점은 대량의 병렬 코퍼스가 부재한 언어 쌍의 경우 다소 무색해진다. 일반적으로 병렬 코퍼스가 소량일 경우 데이터 희소성(Sparsity) 문제가 발생하며 이는 번역 모델의 성능을 제한하는 주요한 요소로 작용한다[5].

최근 NMT 모델의 병렬 데이터 희소성 문제를 해결하기 위하여 단일-언어 데이터를 이용하여 NMT 모델을 학습하는 다양한 방법들이 제안되었다. Gulcehre는 언어 모델을 NMT 모델에 통합하여 Decoder의 언어 생성 능력을 향상하는 방법을 제안하였다[6]. Artetxe와 Lample은 각각 단일-언어 데이터만 이용하여 비지도 학습 방법으로 기계번역 모델을 학습하였고 준수한 성능을 보였다[7, 8]. 마지막으로 Sennrich는 역-번역을 이용하여 영어-독일어 번역에서 향상된 성능을 보였다[9]. 이중 역-번역 방법은 NMT 모델의 구조를 변경하지 않고 독립적으로 적용할 수 있다는 장점이 있다.

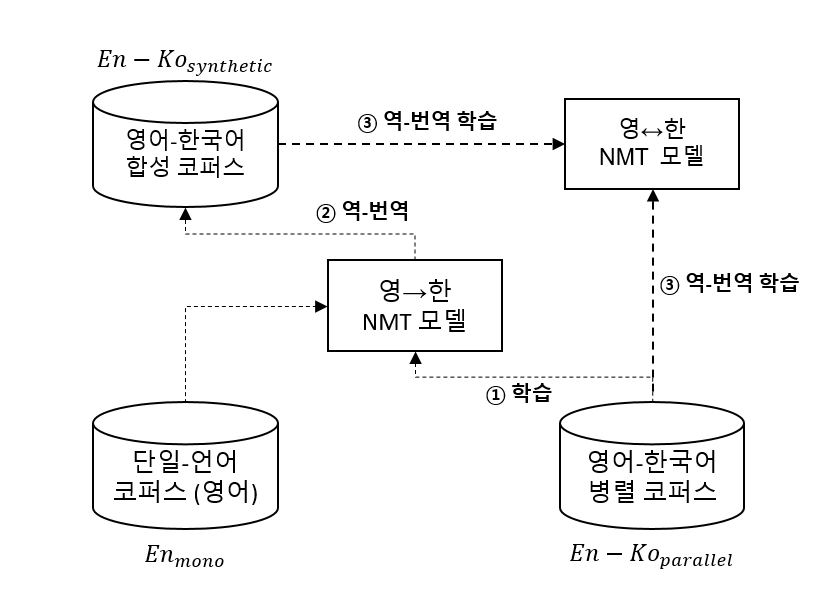
본 논문에서는 단일-언어 코퍼스를 이용하여 가상 병렬 코퍼스를 합성한 후 기존 병렬 코퍼스를 확장하는 역-번역 방법을 이용하여 한국어-영어 기계번역 모델을 학습하였다. 역-번역을 이용한 경우 기본 모델보다 영어→한국어 번역에서 +4.7 BLEU 점수의 성능 향상을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 NMT모델과 Attention 매커니즘을 설명하고 3장에서 역-번역을 이용한 기계번역 모델 학습 방법을 소개한다. 4장에서는 실험환경 설명하고 5장에서는 역-번역 모델과 기본 모델의 성능을 비교한다. 마지막으로 6장에서 결론 및 향후 연구를 제시한다.

**2. Attention기반 Encoder-Decoder 기계번역 모델**

기계번역에서 Encoder-Decoder NMT 모델은 소스 문장 을 타깃 문장 으로 변환하는 조건부 확률 값 을 신경망으로 모델링한다. Encoder는 소스 문장을 고정 사이즈 벡터 로 인코딩하고 Decoder는 인코딩 벡터 를 기반으로 를 최대로 하는 타깃 단어 시퀀스를 수식 (1)을 이용하여 생성한다.

(1)

순환 신경망(RNN: Recurrent Neural Network)을 이용한 Decoder의 타깃 단어 디코딩 과정은 수식 (2)와 같이 표시할 수 있다.

(2)

여기서 는 RNN의 은닉 층 벡터 를 출력 층 벡터로 변환하는 함수이고 는 순환함수 에 의하여 시간 축에 따라 수식 (3)을 이용하여 순환한다. 순환함수 는 함수를 사용할 수 있지만 장기 종속성 문제를 적절하게 다루기 위하여 LSTM(Long-short Term Memory)이나 GRU(Gated Recurrent Unit)을 사용한다[9,10].

그림 1. 역-번역을 이용한 기계번역 모델 학습 과정

(3)

Encoder-Decoder 모델은 소스 문장 전체를 고정 사이즈 벡터 로 표현하기 때문에 인코딩하는 소스 문장이 길어질 경우 장기 종속성 문제가 발생한다. 또한 타깃 문장의 단어와 소스 문장의 단어는 연관성을 가지는데 디코딩 단계에서 이러한 연관성을 적절하게 다루지 못한다. 이러한 문제를 해결하는 방법이 바로 Attention 매커니즘이다.

Attention기반 NMT는 디코딩 단계에서 생성하는 타깃 단어와 소스 단어의 연관성을 Context 벡터 로 수식 (4)에 의해 표현한다. 또한 를 이용하여 Attention 벡터 를 계산하고 이는 수식 (5)에 의해 타깃 단어 생성확률 계산에 사용된다.

(4)

(5)

수식 (4)에서 는 Attention weight인데 이를 계산하는 방법으로 Bahdanau Attention과 Luong Attention이 제안되었다[3, 4].

**3. 역-번역을 이용한 기계번역**

본 논문에서는 역-번역 방법을 이용하여 단일-언어 코퍼스를 NMT모델의 학습과정에 사용하였다. [그림 1]은 역-번역을 이용한 한국어(소스)→영어(타깃) 번역 모델의 학습과정을 나타낸다. 우선, 영-한 병렬 코퍼스 를 이용하여 영어→한국어 역-번역 모델을 학습한다(단계 ①). 다음, 단일-언어 코퍼스 에 대하여 단계 ①에서 학습한 영어→한국어 기계번역 모델로 역-번역을 진행하여 한국어 문장으로 번역한다. 이어서 번역된 한국어 문장을 단일-언어 코퍼스와 문장 단위로 매핑하여 병렬 코퍼스 를 구성한다. 이 코퍼스에서 한국어 문장은 기계번역 모델이 자동으로 번역한 문장이기 때문에 오류가 포함되어 있을 수 있으며 이 코퍼스를 가상 병렬 코퍼스라고 명명하였다. 이어서 기존의 병렬 코퍼스 와 가상 병렬 코퍼스 를 일정한 비율[[1]](#footnote-1)로 조합하여 영어-한국어 병렬 코퍼스를 확장한다. 마지막으로 확장한 병렬 코퍼스 을 이용하여 NMT 모델을 학습한다(단계 ③).

**4. 실험 환경**

본 논문의 실험에서는 한국어-영어 병렬 코퍼스로 OpenSubtitles2018[[2]](#footnote-2)를 사용하였다. 전체 코퍼스에서 3 어절 이하인 한국어 문장은 제외시켰고 Uniform 확률로 각각 5,000 문장을 선택하여 개발(Dev) 셋과 평가(Test) 셋을 구성하였다. [표 1]는 한국어-영어 병렬 코퍼스의 통계를 보여준다. 단일-언어 코퍼스의 경우 WMT2016 영어-독일어 병렬 코퍼스[[3]](#footnote-3) 중 영어 문장들을 추출하여 별도의 단일-언어 코퍼스를 구성하였다. [표 2]는 WMT2016 영어 코퍼스의 통계를 나타낸다. 어절의 길이가 50이상인 문장은 실험 코퍼스에서 제외시켰다.

병렬 문장의 토큰화는 다음과 같이 처리하였다. 영어 문장은 Moses[[4]](#footnote-4) 토큰화 스크립트를 사용하였고 한국어 문장은 Komoran[[5]](#footnote-5) 형태소 분석기를 사용하였다. 또한 희소 단어 (Rare Word) 문제를 효율적으로 다루기 위하여 BPE(Byte-pair Encoding)[[6]](#footnote-6) 방법을 이용하여 영어, 한국어 문장을 Sub-word 유닛 시퀀스로 변환하였다[11]. BPE Sub-word 유닛의 사전 크기는 20,434(16k merge ops)이고 영어, 한국어는 같은 유닛 사전을 공유한다.

표 1. OpenSubtitles2018 한국어-영어 코퍼스 통계

|  |  |  |  |
| --- | --- | --- | --- |
| 데이터 | 학습 | 개발 | 평가 |
| 문장 수 | 919,621 | 5,000 | 5,000 |
| 어절 수 (한) | 5,737,505 | 31,433 | 31,335 |
| 어절 수 (영) | 9,775,178 | 53,223 | 53,826 |
| 문장 당 어절 (한) | 6.24 | 6.29 | 6.27 |
| 문장 당 어절 (영) | 10.64 | 10.64 | 10.77 |

표 2. WMT 2016 영어-독일어 코퍼스 통계

|  |  |
| --- | --- |
| 데이터 | 학습 |
| 문장 수 | 4,500,966 |
| 어절 수 (영) | 113,548,249 |
| 문장 당 어절 (영) | 25.23 |

표 3. NMT 모델의 파라미터

|  |  |
| --- | --- |
| 파라미터 | 값 |
| RNN Cell | LSTM |
| RNN Layers | 2 |
| RNN Hidden Units | 256 |
| RNN Dropout Rate | 0.2 |
| Batch Size | 128 |
| Optimizer | Adam |
| Learning Rate | 0.0001 |
| Beam Width | 5 |

본 논문의 실험에서 사용한 NMT모델은 2장에서 소개한 Luong Attention 매커니즘을 적용한 LSTM 기반 Encoder-Decoder 신경망이다. [표 3]는 실험에서 사용한 NMT 모델의 자세한 파라미터를 보여준다.

**5. 실험 및 분석**

실험은 영-한 병렬 코퍼스 로 학습한 NMT모델을 기본 모델로 하고 역-번역[[7]](#footnote-7)으로 확장한 병렬 코퍼스 로 학습한 NMT 모델의 성능을 비교하였다. 모델의 성능은 BLEU점수를 사용하고 한국어의 경우 형태소 단위를 기준으로 BLEU1 ~ BLEU4를 계산하여 평가한다.

[표 4]는 기본 모델과 제안 모델의 성능을 나타낸다. 한국어→영어 번역 모델의 경우 역-번역으로 합성한 가상 병렬 코퍼스는 번역 성능 향상에 큰 기여를 하지 못하였다. 반면, 영어→한국어 번역의 경우 확장 병렬 코퍼스로 학습한 모델의 성능은 기본 모델에 비하여 BLEU4

표 4. 기본 모델과 제안 모델의 기계번역 성능 비교

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NMT Models | BLEU1 | BLEU2 | BLEU3 | BLEU4 |
| Parallel (Baseline)  (한국어→영어) | 35.95 | 24.64 | 18.46 | 14.39 |
| Parallel + Synthetic  (한국어→영어) | 35.69 | 24.47 | 18.34 | 14.34 |
| Parallel (Baseline)  (영어→한국어) | 25.92 | 17.71 | 12.88 | 9.64 |
| Parallel + Synthetic  (영어→한국어) | 36.11 | 24.78 | 18.59 | 14.50 |

기준으로 +4.7로 크게 향상되었다. 본 논문에서 영어 단일-언어 코퍼스를 이용하여 병렬 코퍼스를 확장한 점을 고려할 때 Encoder의 입력으로 오류가 포함된 합성 언어(한국어)를 사용하는 것보다 오류가 없는 단일-언어(영어)를 사용할 때 성능 향상에 크게 기여한다는 것을 유추할 수 있다. 이러한 실험결과는 Sennrich가 제시한 결과[[8]](#footnote-8)와 비교할 때 다소 상반되지만 이러한 차이가 발생하는 이유에 대해서는 향후 추가 실험을 통하여 밝혀낼 계획이다.

**6. 결론 및 향후 연구**

본 논문에서는 역-번역 방법으로 단일-언어 코퍼스를 가상 병렬 코퍼스로 확장하였고 이를 기존 병렬 데이터와 함께 NMT모델의 학습에 사용하였다. 실험 결과 가상 병렬 코퍼스를 사용할 경우 영어→한국어 NMT모델에서 유의미한 성능 향상을 보였다.

역-번역 방법의 가장 큰 장점 중 하나는 기존의 NMT모델을 수정하지 않고 바로 적용할 수 있다는 점이다. 이러한 장점 때문에 역-번역 방법을 SMT 모델에도 동일하게 적용할 수 있다. 따라서 역-번역 방법은 한국어를 포함한 기계번역 모델을 구축할 때 적극적으로 고려해볼 만한 방법이다.

본 연구에서는 역-번역 방법으로 합성한 가상 병렬 코퍼스에 어떠한 선택 알고리즘도 적용하지 않았다. 가상 병렬 코퍼스에서 모델의 성능 향상에 도움될 만한 문장을 선택해낼 수 있다면 더 큰 성능 향상을 달성할 수 있을 것으로 기대한다. Unknown단어[[9]](#footnote-9)가 포함된 합성 문장을 제외하는 것이 그중 한가지 방법이 될 것이다. 또한 기존 병렬 코퍼스와 같은 도메인에서 단일-언어 코퍼스를 수집한다면 더 효과적인 가상 병렬 코퍼스를 합성할 수 있을 것으로 예상한다. 향후, 상술한 두 가지 이슈에 대하여 자세한 연구를 진행할 계획이다.

**참고문헌**

1. M. T. Luong, I. Sutskever. Q. V. Le, Q. Vinyanls, and W. Zaremba, Addressing the rare word problem in neural machine translation, ACL 2015.
2. S. Jean, K. Cho, R. Memisevic and Y. Bengio, On using very large target vocabulary for neural machine translation, ACL 2015.
3. D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate, ICLR 2015.
4. M. T. Luong, P. Hieu and M. D. Christopher, Effective approaches to attention-based neural machine translation, arXiv:1508.04025, 2015.
5. P. Koehn and R. Knowles, Six challenges for neural machine translation, pp. 28-39, Proceedings of the First Workshop on Neural Machine Translation, ACL 2017.
6. C. Gulcehre et. al., On using monolingual corpora in neural machine translation, arXiv:1503.03535, 2015.
7. M. Artetxe, G. Labaka, E. Agirre and K. Cho, Unsupervised neural machine translation, ICLR 2018.
8. G. Lample, L. Denoyer and M. Ranzato, Unsueprvised machine translation using monolingual corpora only, ICLR 2018.
9. S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Computation, 1997.
10. J. Chung, C. Gulcehre, K. Cho and Y Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, NIPS 2014.
11. 이재환, 김보성, 허광호, 고영중, 서정연, Subword 유닛을 이용한 영어-한국어 신경망 기계번역, 한국컴퓨터종합학술대회 2018.

1. 본 논문에서는 Parallel, Synthetic 2대1의 비율로 사용하였다. [↑](#footnote-ref-1)
2. http://opus.nlpl.eu/OpenSubtitles2018.php [↑](#footnote-ref-2)
3. http://www.statmt.org/wmt16/translation-task.html [↑](#footnote-ref-3)
4. http://www.statmt.org/moses [↑](#footnote-ref-4)
5. http://konlpy.org/ko/v0.4.3 [↑](#footnote-ref-5)
6. https://github.com/rsennrich/subword-nmt [↑](#footnote-ref-6)
7. 역-번역은 영어→한국어 기본 모델을 사용하여 진행하였다. [↑](#footnote-ref-7)
8. Sennrich는 합성 언어(소스)를 단일-언어(타깃)로 번역하였다. [↑](#footnote-ref-8)
9. 사전에 등록되지 않은 단어 (Out of vocabulary word) [↑](#footnote-ref-9)