DBpia

Bidirectional LSTM-CRF 기반 띄어쓰기 모델을 활용한 한영/영한 번역 시스템

김영표°, 안동민, 최슬기, 황소정, 최희열
한동대학교
menstoo9504@gmail.com, adominic022@gmail.com, ch062404@naver.com,
sojeoung536985@gmail.com, hchoi@handong.edu

KR-EN, EN-KR NMT System with Bidirectional LSTM-CRF Korean Spacing

Youngpyo Kim°, Dongmin An, Seulgi choi, Sojung Hwang, Heeyoul Choi
Handong Global University

요 약

구글 번역기와 같은 딥러닝에 의존하는 현대 번역 시스템의 도입으로 현재의 최신 번역은 원어민 또는 전문 번역가와 유사한 번역 수준에 도달했다. 현재 모델들은 각 언어의 고유 특성 문제에 직면했음에도 불구하고 성능이 뛰어나다. 일반적인 번역 모델들은 KR-EN, EN-KR 번역에 탁월한 성능을 보이지만, 한국어의 특징인 띄어쓰기를 전혀 고려하지 않는다. 본 논문에서는 일반적인 Transformer 번역 모델과 결합한 Bidirectional LSTM 및 CRF를 사용한 모델을 제시한다. 실험은 KR-EN, EN-KR 문장으로 구성된 AI Hub 데이터로 수행한다. 제시된 모델은 번역 결과에서는 Google, Papago와 비슷한 수준의 정확도에 도달하지만, 띄어쓰기에서 더 우수한 면을 보이는 것을 확인할 수 있다.

1. 서 론

한국어는 어절 간의 띄어쓰기를 이용하기 때문에 다른 나라 언어와는 다르게 띄어쓰기에 따라 문장의 의미가 바뀌는 언어이다. 사람은 띄어쓰기가 잘못되어 있어도 의미를 유추할 수 있지만, 인공신경망 기반 기계 번역기(NMT)는 띄어쓰기가 제대로 되어있지 않은 문장을 전혀 다른 문장으로 번역한다. NMT를 학습시킬 때에는 한 언어로 이루어진 말뭉치와 일대일 매칭이 되는 번역된 말뭉치의 전처리 과정이 필요하다. 전처리 과정 중에 학습할 모델이 문맥을 파악하는 것을 도와주는 토큰화 작업이 띄어쓰기를 기준으로 문장을 나누기 때문에 올바른 띄어쓰기가 필요하다[1].

본 논문에서는 이러한 띄어쓰기의 중요성에 주목해, 첫째로 말뭉치 데이터의 띄어쓰기를 Bidirectional LSTM (Long-Short Term Memory) 과 CRF (Conditional Random Field)를 활용한 모델로 교정한다. 그 다음으로 띄어쓰기를 교정하지 않은 데이터로 학습한 Transformer 기반 번역기 모델과 띄어쓰기를 교정한 데이터로 학습한 번역기를 비교한다. 번역기의 학습 데이터로 사용된 말뭉치는 AI Hub에서 구축한 160만 개의 문장을 활용하여 실험을 진행하였다.

2. 본 문

2.1 띄어쓰기 모델

본 실험에 쓰인 띄어쓰기 모델은 그림 1과 같이 Word Embedding을 거친 후, 양방향 LSTM을 활용해 입력 문자들에 대한 문맥 정보를 학습하고, CRF를 이용해 출력 레이블(0: Non-Spacing, 1: Spacing)에 대해 학습을 하는 Bi-LSTM+CRF를 사용하였다. 이 구조는 LSTM의 장점인 긴 의존기간을 필요로 하는 학습을 잘 할 뿐만 아니라 양방향의 특성도 가지고 있기 때문에 데이터를 효율적으로 사용할 수 있다는 것이다. 또한 레이블의 인접성에 대한 정보를 바탕으로 레이블을 추측하는 CRF를 함께 사용하기 때문에 문장 수준의 태그 정보를 학습할 수 있어서 Bi-LSTM보다 더 좋은 결과를 얻을 수 있다[2].

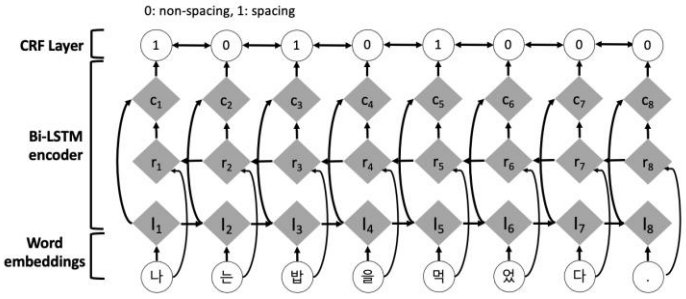

그림 1. Bi-LSTM+CRF을 사용한 띄어쓰기 모델의 구조

실험에 쓰인 데이터는 AI Hub 160만개의 데이터를 나누어 159만개를 학습으로, 3천개를 검증으로, 3천개를 평가로 사용하였다. 학습 데이터를 통해서 구축한 사전의 길이는 4280개였다. 모델의 성능을 나타내는 F1 -score는 0.94 이다.

2.2 번역기 모델

LSTM 이후 Transformer 모델[3]의 소개로, RNN 구조를 사용하지 않으면서 순수한 Attention으로 동일한 성능을 낼 수 있다. 똑같은 Encoder-Decoder 구조이지만 Transformer 모델은 기존 RNN의존

www.dbpia.co.kr

모델들과 달리 Self-attention을 통해 문장 안 단어들의 유사 점수를 비교하며, Positional Embedding을 통해 문장 안의 토큰에 값을 주어서 반복되는 연산을 줄인다. 결정적으로 순차적 데이터를 순서대로 처리할 필요가 없으므로, 학습 시 병렬 처리에 더욱 유리하다. 이러한 이점을 활용해 많은 양의 데이터를 처리할 수 있는 Transformer 모델을 사용한다. 위 띄어쓰기 모델부분에서 설명한 바와 같이, 번역기 모델에서도 데이터는 동일하게 나눴다. 어휘 사전 같은 경우에는 BPE(Byte Pair Encoding) 과정을 통해 1만개를 추출하였고, 3천개의 단어가 포함된 고유명사 사전도 이용하였다.

2.3 번역기 모델에 띄어쓰기 모델 적용 점

Transformer 모델을 사용한 번역 모델 결과의 질 향상과 위에서 언급한 문제 해결을 위해 EN-KR, KR-EN으로 번역하는 과정 중에 띄어쓰기 모델을 적용하였다. 띄어쓰기 과정은 번역의 종류와 띄어쓰기 모델을 어느 시점에 적용하느냐에 따라 3가지로 나누어진다.

먼저, 기존 번역기의 잘못된 한국어 띄어쓰기를 보완하기 위해 번역 모델을 EN-KR 데이터로 학습할 때, 후처리로 띄어쓰기 모델을 사용하였다. 다시 말해 기존의 AI Hub 데이터로 영한 번역 모델을 먼저 학습한 후, 번역된 한국어 데이터에 후처리로서 띄어쓰기 모델을 추가하였다.

둘째로, 번역 모델을 EN-KR 데이터로 학습할 때, 데이터 전처리 과정으로 띄어쓰기 모델을 적용하였다. 즉, 띄어쓰기가 잘못 되어있는 학습 데이터를 올바르게 띄워준 후, 학습을 진행하였다.

마지막으로, 한영 번역 모델을 학습할 때 전처리로서 AI Hub 한국어 데이터에 띄어쓰기 모델을 적용한 후, 번역 모델을 학습하였다. 표 1을 통해 학습데이터의 띄어쓰기 예시를 확인할 수 있다.

표 1. 띄어쓰기 모델 적용 후, 변화된 학습 데이터의 예시.

띄어쓰기 모델 적용 전	같은 지점에서 연이어 발생한 사고로 ***대체건설이*** 불가피했던 ***것으로판단된다.***
띄어쓰기 모델 적용 후	같은 지점에서 연이어 발생한 사고로 ***대체 건설이*** 불가피했던 ***것으로 판단된다.***

3. 수행 결과

3.1 띄어쓰기 모델로 후처리한 영한 번역 결과

후처리로 띄어쓰기를 적용한 EN-KR 모델의 경우, 표 1을 보면 알 수 있듯이 BLEU 13.61을 기록했다. 이 모델은 기존 번역기들의 띄어쓰기 오류를 올바르게 고치는 등, 질적인 측면에서 우수한 성과를 보였음을 다음 예시를 통해 알 수 있다. 2021년 3월 20일 검색 기준, 구글과 파파고 번역기에서 다음과 같은 띄어쓰기 오류가 있었다.

입력문장

Local taxes for the property retained by a foreigner-invested enterprise to carry out any business reported therefrom shall be reduced or exempted in accordance with the following: Provided, That if it becomes a subject of collection pursuant to Article 121-5(3) of the Restriction of Special Taxation Act, the reduced or exempted local tax shall be collected.

구글 번역

외국인 투자 기업이 사업 신고를 위해 보유한 재산에 대한 지방세는 다음 각 호에 따라 감면된다. 다만, ***제 121 조의 5 제 3 항***에 따라 징수 ***대상이되는*** 경우 조세특례 제한법에 따라 ***감면 된*** 지방세를 징수한다.

파파고 번역

외국인투자기업이 신고한 사업을 수행하기 위하여 보유하는 재산에 대한 지방세는 다음 각 호에 따라 감면한다. 다만, ***「조세특례제한법」 제121조의5제3항에 따라 징수대상이 되는 경우에는 감면된 지방세로 한다. 수집의***

기존 Transformer 번역기 번역 결과

외국인투자기업이 신고한 사업을 수행하기 위하여 실시하는 재산에 대하여는 다음 각 호에 따라 지방세를 감면한다. 다만, 「조세특례제한법」 제121조의5제3항에 따라 ***징수대상이*** 되는 경우에는 감면 또는 지방세를 징수한다.

띄어쓰기 적용한 번역 결과

외국인투자기업이 신고한 사업을 수행하기 위하여 실시하는 재산에 대하여는 다음 각 호에 따라 지방세를 감면한다. 다만, 「조세특례제한법」 제121조의5제3항에 따라 ***징수 대상이 되는 경우***에는 ***감면 또는 지방세를 징수한다.***

구글에서는 '대상이 되는', '감면된', '제121조의5제3항'을 "대상이되는', '감면 된', '제 121 조의 5 제 3 항'이라고 잘못 띄어쓰기를 하였고, 파파고에서는 '감면된 지방세로 한다. 수집의' 라는 부자연스러운 번역 결과를 내놓았다.

반면, 띄어쓰기를 후처리로 사용한 모델의 번역 결과에서는 기존의 번역기들과 다르게 올바른 띄어쓰기와 자연스러운 번역을 확인할 수 있었다. 즉, 질적 비교를 통해 본 논문의 번역 시스템이 기존 번역 시스템보다 더 나은 띄어쓰기 문법과 자연스러운 한국어 번역을 제공하는 것을 확인할 수 있었다.

3.2 띄어쓰기 모델로 전처리한 영한 번역 결과

띄어쓰기 모델을 전처리로 번역 모델에 적용한 후,

www.dbpia.co.kr

EN-KR 번역을 실행하였다. 표 2를 보면 번역 모델만 사용하였을 때 BLEU 결과는 17.26이 나왔지만 띄어쓰기 모델을 전처리로 적용하였을 때에는 BLEU 점수가 17.29가 나왔다. 번역에 대한 차이를 조금 더 확인해 보기 위해 구글과 파파고의 번역 결과를 비교했다.

입력문장

Gyeongsangbuk-do announced on the 8th that the 'local cousin youth demonstration complex construction project' submitted by Uiseong-gun was selected as a result of the application for the '2019 Regional Development Investment Convention Pilot Project' hosted by the National Balanced Development Committee and the Ministry of Land, Infrastructure, and Transport.

구글 번역

경상북도는 국가 균형 발전위원회 주최 '2019 지역 개발 투자 컨벤션 시범 ***사업'신청*** 결과 의성군이 ***제출한*** '지역 사촌 청년 시범 단지 건설 사업'이 선정됐다고 8 일 밝혔다. ***국토 교통부.***

파파고 번역

경상북도는 8일 국가균형발전위원회와 국토교통부가 주최한 '2019년 ***지역개발투자협약*** 시범사업' 신청 결과 의성군이 제출한 '지역 ***사촌형*** 청년시범단지 조성사업'이 선정됐다고 밝혔다.***d 운송***

기존 Transformer 번역기 번역 결과

경상북도는 의성군이 제출한 '지역사촌 청년 실증단지 조성사업' 이 국가균형발전위원회와 국토교통부 주관 '2019년 ***지역발전투자협약*** 시범사업' ***신청결과*** 선정됐다고 8일 밝혔다.

띄어쓰기 적용한 번역 결과

경상북도는 의성군이 제출한 '지역친척 청소년 시범단지 조성사업' 이 국가균형발전위원회와 국토교통부가 주관한 '2019년 ***지역개발 투자협약*** 시범사업' ***신청 결과*** 선정됐다고 8일 밝혔다.

파파고에서는 '지역개발투자협약', '사촌형' 등 띄어쓰기 부분에서도 문제가 있었고, 번역 또한 'd 운송'이 추가되는 모습을 보이며 낮은 번역의 질을 보였다. 구글도 '제출 한', '8 일' 등의 띄어쓰기 오류를 보였고, 문장의 마지막에 '국토 교통부.'라는 단어가 들어가서 번역의 질이 낮았다. 띄어쓰기를 전처리로 적용하지 않은 번역기의 번역 결과와 전처리를 한 번역기의 결과 차이를 보면 '신청결과'를 '신청 결과'로, '지역발전투자협약'을 '지역개발 투자협약'으로 띄어쓰는 등 띄어쓰기 부분에서 개선된 모습도 보였다. 또한 '국토교통부 주관'이 '국토교통부가 주관한'처럼 조사가 자연스럽게 들어가 질적으로 향상된 모습을 보였다.

표 2. 일반 EN-KR Transformer 번역기와 띄어쓰기 모델을 전처리, 후처리로 적용한 번역기 BLEU 점수 비교

	BLEU 점수
일반 Transformer 번역기	17.26
전처리로 띄어쓰기 모델이 적용된 번역기	17.29
후처리로 띄어쓰기 모델이 적용된 번역기	13.61

3.3 띄어쓰기 모델로 전처리한 한영 번역 결과

아래 표 3을 통해 알 수 있듯이 전처리로 띄어쓰기를 적용한 KR-EN 모델의 경우, BLEU 34를 기록했다. 이는 BLEU 34.2인 기존 KR-EN 모델과 비슷한 성능으로 기존 KR-EN 모델과 번역 결과의 문맥과 의미는 유사했으나, 문장의 구조가 상이했다.

표 3. 일반 KR-EN Transformer 번역기와 띄어쓰기 모델이 적용된 번역기 BLEU 점수 비교

	BLEU 점수
일반 Transformer 번역기	34.2
전처리로 띄어쓰기 모델이 적용된 번역기	34

4. 결 론

본 논문은 Bidirectional LSTM과 CRF 모델을 통해 기존 번역기들이 한국어-영어 번역 중 띄어쓰기를 잘못 번역하는 경우를 개선시키는 시스템을 제시하였다. 한국어 띄어쓰기를 모델에 적용하여서 기계번역기가 잘못된 문장으로 오역하는 것을 방지할 수 있다.

향후 연구에서는 기존 띄어쓰기 모델에 쓰인 Bidirectional LSTM과 CRF 모델 구조 성능 향상을 위해 Bidirectional Encoder Representations from Transformers (BERT)로 구조를 개편할 예정이다. 결론적으로 연구의 목표는 번역된 문장의 띄어쓰기 오류를 줄여 인공신경망 기반 기계번역기의 성능을 향상시키는 것에 있다.

참 고 문 헌

[1] Webster, Jonathan J., and Chunyu Kit. "Tokenization as the initial phase in NLP." COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics. 1992.

[2] Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." arXiv preprint arXiv:1508.01991 (2015).

[3] Vaswani, Ashish, et al. "Attention is all you need." arXiv preprint arXiv:1706.03762 (2017).

www.dbpia.co.kr