BERT를 활용한 한국어 개체명 인식기

황석현*, 신석환**, 최동근*, 김성현*, 김재은*
*㈜솔트룩스 AI Labs.
**㈜솔트룩스 파트너스
e-mail : shhwang@saltlux.com

Korean Named Entity Recognition using BERT

Seokhyun Hwang*, Seokhwan Shin**, Donggeun Choi*, Seonghyun Kim*, Jaieun Kim*
* AI Labs, Saltlux Inc
**Saltlux Partners

요 약

개체명이란, 문서에서 특정한 의미를 가지고 있는 단어나 어구를 뜻하는 말로 사람, 기관명, 지역명, 날짜, 시간 등이 있으며 이 개체명을 찾아서 해당하는 의미의 범주를 결정하는 것을 개체명 인식이라고 한다. 본 논문에서는 BERT(Bidirectional Encoder Representations from Transformers) 활용한 한국어 개체명 인식기를 제안한다. 제안하는 모델은 기 학습된 BERT 모델을 활용함으로써 성능을 극대화하여, 최종 F1-Score 는 90.62 를 달성하였고, Bi-LSTM-Attention-CRF 모델에 비해 매우 뛰어난 결과를 보였다.

키워드: 자연어 처리, BERT, 개체명 인식, 딥러닝

1. 서론

개체명 인식(Named Entity Recognition)이란 문서에서 개체명을 인식하고, 인식된 개체명의 종류를 분류하는 자연어처리의 한 분야이다.

개체명이란 사람 이름, 회사 이름, 지명, 영화 제목, 날짜 등 문서에서 특정한 의미를 가지고 있는 단어 또는 어구이다. 즉, 특정한 의미를 가지는 명사 혹은 숫자표현 등의 하나 이상의 단어를 말한다. 정보 검색에서 개체명은 중요한 검색 대상이 될 수 있으므로 자연어처리 분야에서 활발히 연구가 진행되고 있고, 중요하다고 볼 수 있다[1]. 본 논문에서는 15 개의 개체명 태그(정보통신단체표준 TTA.KO-10.0852)인 AF(대상물), AM(동물). CV(문명), DT(날짜), EV(사건), FD(학문), LC(지역), MT(금속), OG(기관), PS(사람), PT(식물), QT(수량표현), TI(시간), TM(용어), TR(이론)를 대상으로 분류하였다. 아래는 개체명을 표시한 문장의 예시이다.

“독일<LC>에서 태어난 아인슈타인<PS>은 1916 년 <TI>에 일반상대성 이론<TR>을 발표하였다.”

개체명은 품사로는 고유명사 또는 미등록어인 경우가 많으며, 항상 새롭게 만들어지고, 때로는 같은 단어라도 사용되는 문장에 따라 상이한 의미를 나타낸다. 또 개체명은 고유명사나 미등록어 하나가 하나의 개체명을 이룰 수도 있고, 또한 2 개 이상의 고유명사나 일반명사가 결합하여 복합 명사 혹은 명사구 형태를 보이기 때문에 그 경계를 인식하기는 쉽지 않아 사전을 기반한 방법은 정확도가 높지 못하다.

기계 학습을 활용한 방법이 등장하면서 HMM (Hidden Markov Model)[2], SVM (Support Vector Machine)을 활용한 개체명 인식기[4]가 연구되었으며, 딥러닝이 발전하면서 Bi-LSTM (Bi-directional Long Short Term Memory)와 CRF (Conditional Random Fields)를 활용한 모델[3],[5],[6]이 등장하면서 더 높은 성능을 보이기 시작하였다.

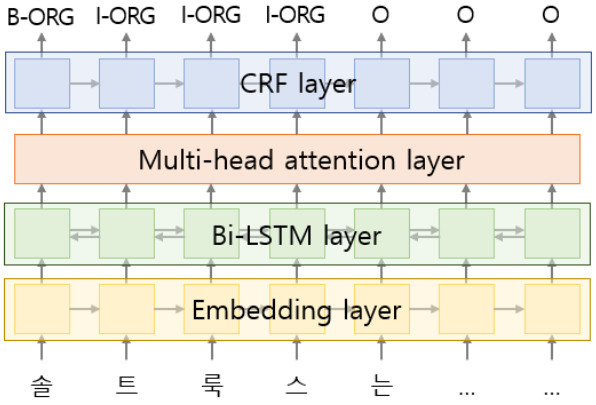
본 논문에서는 자연어처리 분야에서 좋은 성능을 보이는 BERT[7] 모델을 활용하여 개체명 인식의 성능을 높이는 실험을 진행하였다. 비교 모델로는 Bi-LSTM+Attention+CRF 를 사용했고, 기학습된 BERT 모델을 활용하여 테스트 하였으며, 성능향상을 위해 WIKI 데이터를 활용하여 연장학습을 수행하였다. 논문의 구성은 2 장에서 개체명 인식 모델 이해를 위한 배경 지식 소개와 논문의 비교 모델을 소개한다. 3 장에서는 실험 및 결과에 대해 토의하며, 마지막으로 4 장에서는 결론 및 향후 연구 방향에 대하여 제안한다.

2. 개체명 인식 모델

2.1 Bi-LSTM+Attention+CRF

Bi-LSTM 과 CRF 를 사용한 모델로 구조는 아래 그림과 같다. Embedding layer, Bi-LSTM, attention, CRF 네트워크를 거쳐서 개체명이 태깅 된다. 음절 단위의

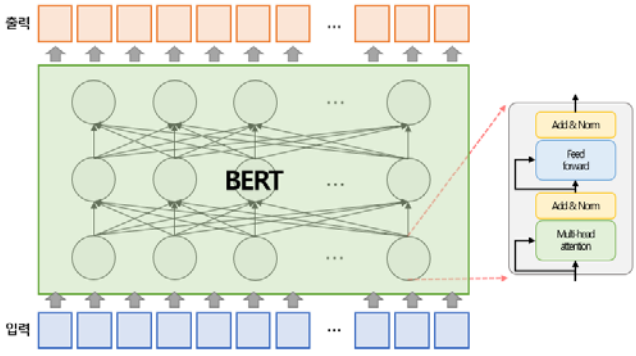
입력을 사용하였으며, 형태소와 사전 정보를 추가로 입력으로 넣어주었다. 형태소는 따로 임베딩 레이어를 만들어서 사용하였으며, 개체명 사전의 경우에는 15 개의 개체명에 대하여 사전에 있는 개체명의 경우 1, 없을 경우 0 으로 입력을 넣어주었다.



<그림 1> Bi-LSTM+Attention+CRF 모델

2.2.3 BERT 기반의 개체명 인식 모델

BERT-NER 모델은 BERT 를 활용한 개체명 인식 모델로, BERT 모델에 한 개의 layer 를 추가하여 해당 layer 에서 개체명을 태깅하도록 설계했다. 개체명 태깅에서 음절 단위의 토크나이징을 활용했다. BERT 기반의 개체명 인식 모델 구조는 아래 그림과 같다.



<그림 2> BERT 기반 개체명 인식 모델

BERT 기반의 개체명 인식 모델은 2 가지로 평가를 진행했다. BERT-NER 모델은 구글에서 공개한 BERT-multilingual[7] 모델을 활용하였다. BERT-NER+ 모델은 WIKI 데이터를 활용하여 BERT-multilingual 에 연장 학습을 진행한 모델이며 V100 32GB GPU 를 활용하여 16 Batch size 로 100K 스텝을 연장 학습 하였다.

3. 실험 및 결과

3.1 실험 환경

Bi-LSTM 기반의 개체명 인식 모델과 BERT 기반의 개체명 인식 모델의 학습과 평가를 위해 솔트룩스에서 구축한 데이터를 동일하게 사용했다. 데이터 구성은 학습 데이터는 총 95,786 문장에 257,387 개의 태그가 태깅된 데이터이고, 평가 데이터는 총 10,502 문장에 27,781 개의 태그가 태깅된 데이터로 이루어져 있다. 평가 지표는 precision, recall, F1-Score 를 활용했고, 15 개의 개체명 태그를 각각 평가하고, 전체 태그에 대한 평가도 동일하게 진행했다.

Bi-LSTM+Attention+CRF 모델의 입력은 음절 임베딩 레이어 300 차원, 형태소 정보 30 차원, 개체명 사전 정보 15 차원을 붙여서 345 차원을 입력으로 활용하였다. LSTM 층은 2 층으로 구성하고 cell 크기는 100 을 사용하였으며 multi-head attention 에서 head 의 갯수는 4 개를 사용하였으며, dropout 은 0.5 를 적용하였다. Batch size 는 128 을 활용하여 100 epochs 을 학습하였다.

BERT-NER 과 BERT-NER+ 모델의 학습은 V100 32GB GPU 에서 Batch size 를 16 으로 지정하여 3 epochs 를 학습하였다.

3.2 실험 결과

<표 1> 모델 별 성능

model	precision	recall	F1-Score
Bi-LSTM-Attention-CRF-Dic	83.36	78.87	81.05
BERT-NER	89.22	90.22	89.71

<표 1>을 보면 Bi-LSTM+Attention+CRF 모델에 개체명 사전까지 활용한 모델이 81 점의 F1-Score 를 기록하였으며 BERT 기반의 모델의 경우 기본적인 Multilingual 모델을 활용하여 3 epochs 만 학습하였음에도 89.71 의 F1-Score 를 기록했다.

<표 2> 태그 별 Bi-LSTM-Attention-CRF 와 BERT-NER

	Bi-LSTM-Attention-CRF	BERT-NER
TAG	F1-Score	F1-Score
TOTAL	81.05	89.71
AF	80.75	87.61
AM	79.34	85.67
CV	71.62	84.04
DT	92.48	96.43
EV	49.90	80.46
FD	70.09	79.91
LC	89.54	95.07
MT	71.25	86.91
OG	73.09	88.48
PS	80.03	90.59
PT	80.15	87.61
QT	91.07	94.91
TI	69.60	87.29
TM	70.44	83.33
TR	52.35	77.88

<표 2>는 각 태그별 Bi-LSTM-Attention-CRF 모델과 BERT-NER 모델의 F1-Score 이다. 모든 태그에 대해서 BERT-NER 모델이 성능이 좋았다.

<표 3> 태그 별 BERT-NER 과 BERT-NER+

	BERT-NER	BERT-NER+
TAG	F1-Score	F1-Score
TOTAL	89.71	90.62
AF	87.61	90.38
AM	85.67	84.22
CV	84.04	84.91
DT	96.43	96.99
EV	80.46	80.11
FD	79.91	81.47
LC	95.07	96.14
MT	86.91	87.15
OG	88.48	89.95
PS	90.59	91.88
PT	87.61	85.65
QT	94.91	95.86
TI	87.29	85.72
TM	83.33	82.59
TR	77.88	80.23

<표 3>은 태그 별 BERT-NER 모델과 연장학습을 진행한 BERT-NER+ 모델의 점수다. 대체적으로 근소하게 연장학습을 한 BERT-NER+ 모델이 성능이 좋았다.

4. 결론 및 향후 연구 방향

본 논문에서는 BERT 를 이용한 한국어 개체명 인식 방법을 제안하였다. BERT-multilingual 모델을 활용하되 음절이 입력되도록 수정만 하여 활용하여도 매우 높은 성능을 낼 수 있었으며, WIKI 덤프 데이터를 활용하여 연장학습을 할 경우 더 성능이 향상되었다.

Bi-LSTM+Attention+CRF 모델에 개체명 사전까지 활용한 모델에 비해서 사전을 활용하지 않음에도 불구하고 높은 성능을 내어 실제 서비스에도 충분히 활용할 수 있는 성능으로 여겨진다.

향후 연구로는 BERT 에 CRF 와 같은 네트워크를 붙여서 성능 향상을 시도할 수 있을 것이며, BERT-multiligual 모델이 아닌, 처음부터 한국어로 학습한 BERT 모델을 활용하면 성능이 더 향상될 것으로 기대된다.

Acknowledgement

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-01376, 다중 화자간 대화 음성인식 기술 개발)

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2019-0-01376, Development of the multi-speaker conversational speech recognition technology)

참고문헌

[1] 이경희, 이주호, 최명석, 김길창. (2000). 한국어 문서에서 개체명 인식에 관한 연구. 한국정보과학회 언어공학연구회 학술발표 논문집, 292-299.

[2] 황이규, 윤보현, "HMM 에 기반한 한국어 개체명 인식", 정보처리학회논문지(B), 제 10 권, 제 2 호, pp.229-236, 2003.

[3] 이창기, 황이규, 오효정, 임수종, 허정, 이충희, 김현진, 왕지현, 장명길. (2006). Conditional Random Fields 를 이용한 세부 분류 개체명 인식. 한국정보과학회 언어공학연구회 학술발표 논문집, 268-272.

[4] 이창기, 장명길. Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식. 인지과학, 21(4), 655-667, 2010.

[5] 신유현, 이상구. 양방향 LSTM-RNNs-CRF 를 이용한 한국어 개체명 인식”, 한국어정보학회, page 340-341, 2016.

[6] 유홍연, 고영중. (2017). Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장. 정보과학회논문지, 44(3), 306-313.

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