CNN-CRFs를 이용한 한국어 개체명 인식기

유연수○, 박혁로
전남대학교 전자컴퓨터공학과
powinz00@naver.com, hyukro@jnu.ac.kr

Korean Named-entity Recognition Using CNN-CRFs

Yeon-Soo You○, Hyuk-Ro Park
Department of Electronics & Computer Engineering. Chonnam National University

요 약

개체명 인식 연구에서 우수한 성능을 보이고 있는 bi-LSTM-CRFs 모델은 처리 속도가 느린 단점이 있고, CNN-CRFs 모델은 한국어 말뭉치를 사용하여 제대로 분석되지 않았다. 본 논문에서는 한국어 개체명 인식 말뭉치를 이용한 CNN-CRFs 모델의 음절 단위 한국어 개체명 인식 방법을 제안한다. 실험 결과 bi-LSTM-CRFs 모델보다 CNN-CRFs 모델의 F1 score가 0.4% 높았고, 27.5% 빠른 처리 속도를 보였다.

주제어: 개체명 인식, 심층학습, bi-LSTM-CRFs, CNN-CRFs

1. 서론

개체명(Named Entity)이란 고유한 의미를 가지는 단어나 어구를 말한다. 인명(Person), 지명(Location), 기관명(Organization), 날짜(Date), 시간(Time) 등의 범주로 구분한다. 개체명 인식(Named Entity Recognition)은 입력되는 문장에서 개체명을 추출하고 정의된 범주에 맞추어 분류하는 작업이다. 개체명 인식은 기계가 자연어를 이해하고 요청을 처리하는 검색, 추출, 대화 등의 시스템에서 중요한 핵심 기술이다.

최근의 개체명 인식 연구는 심층학습을 이용하여 좋은 성능을 보이고 있다. 기존의 연구는 언어 처리 단위에 따라서 형태소 단위[1-3], 음절 단위[4]의 방법으로 나뉘고, 모델은 통계 기반의 방법, 심층학습 기반의 방법으로 나뉜다[5].

형태소 단위 개체명 인식 방법은 형태소 분석기에 의존적으로 형태소 분석기의 오류가 전파될 수 있다는 단점이 있다. bi-LSTM-CRFs 모델은 입력 데이터의 흐름에 따라 학습을 진행하기 때문에 속도가 느리다. CNN모델은 입력 데이터의 순서나 의존성을 고려하여 학습하지 못한다.

본 논문에서는 형태소 분석기의 의존성을 배제하고 기존 bi-LSTM-CRFs 모델보다 성능과 속도가 개선된 음절 단위의 CNN-CRFs 모델을 제안한다. 실험결과 bi-LSTM-CRFs 모델보다 CNN-CRFs 모델의 F1-score가 0.4% 높았고, 27.5% 빠른 처리 속도를 보였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 소개하고, 3장에서는 기존 모델인 bi-LSTM-CRFs 모델과 제안 모델인 CNN-CRFs를 소개한다. 4장에서는 소개된 모델의 실험 결과를 비교 분석하고, 5장에서는 결론 및 향후 연구에 관해 기술한다.

2. 관련 연구

과거에는 개체명 인식을 위한 방법으로 통계 기반의 구조적 SVM(Support Vector Machine)을 이용한 방법을 이용하였으나 최근에는 심층학습을 이용한 방법이 보다 우수한 성능을 보이고 있다[5].

심층학습 방법은 데이터의 부분적인 특징을 학습하는 CNN(Convolutional Neural Network)을 이용한 방법보다 연속된 데이터의 의존성을 학습할 수 있는 LSTM 모델이 우수한 성능을 보였다[6].

입력되는 데이터의 앞, 뒤의 순서를 고려하여 학습하는 bi-LSTM(Bi-directioanl LSTM)과 출력 결과 사이의 의존성을 고려하여 학습하는 CRFs(Conditional Random Fields)를 결합한 학습 방법이 적용되었다[1-4].

한국어 개체명 인식의 성능을 향상시키기 위해 형태소 분석된 데이터와 품사 등이 활용되었다[1-3]. 입력되는 단어의 특징을 보다 잘 학습하여 성능을 향상시키는 단어 임베딩을 적용한 방법이 연구되었다[1,4].

개체명 인식 연구에서 CNN 모델은 제안이 되었지만 CRFs를 결합하지 않았거나 영어 말뭉치를 사용한 연구가 진행되었다[6-8]. 또한 LSTM의 속도 문제는 중요하게 고려하지 않았다.

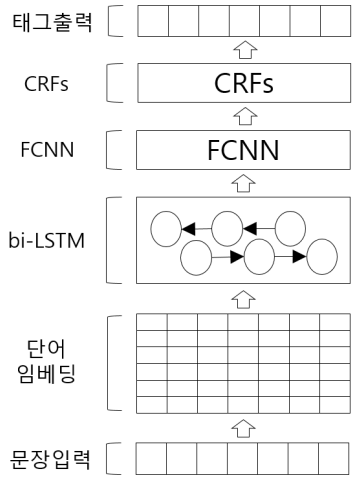
본 논문에서는 CNN-CRFs 모델을 제안하고, 각각의 모델을 한글 개체명 말뭉치 대상으로 비교하는 실험을 통하여 CNN-CRFs 모델이 기존 모델과 비슷한 성능을 낼 수 있고, 속도 면에서 우수하다는 것을 보인다.

3. 개체명 인식 모델

본 장에서는 기존의 bi-LSTM-CRFs 모델을 설명하고, 제안하는 CNN-CRFs 모델에 대하여 소개한다.

3.1 기존 bi-LSTM CRFs 모델

LSTM은 RNN의 기울기 소실(vanishing gradient) 문제를 해결하여 장기 의존성을 잘 학습할 수 있는 모델이



[그림 1] bi-LSTM-CRFs 구조도

다. bi-LSTM은 입력 데이터를 정방향과 역방향으로 학습하여 입력 데이터의 순서와 의존성을 학습한다.

[그림 1]과 같은 구조로 구현된 bi-LSTM 모델은 문장이 입력되면 임베딩 lookup table을 이용하여 각 음절을 벡터로 변환한다. 정방향, 역방향 LSTM을 이용하여 앞뒤 문맥을 고려하여 학습하고 양방향의 출력 결과를 연결(concatenate)한다. 연결된 값을 FCNN(Fully Connected Neural Networks)에 입력하여 각 음절에 개체명 태그 수가 할당되도록 변환한다. 이 결과를 CRFs 층에 입력으로 사용하여 출력 결과 사이의 의존성을 학습하여 보다 정확한 결과를 구하도록 한다.

LSTM 모델의 경우는 입력되는 데이터의 순서에 맞추어 학습을 진행한다. 이런 방식은 입력 데이터의 문맥을 고려하여 학습할 수 있지만 병렬처리가 불가능하고 속도가 느려지는 단점이 있다.

3.2 제안하는 CNN-CRFs 모델

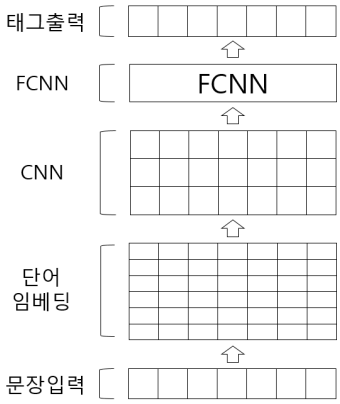
3.2.1 CNN 모델

Convolution이란 합성곱을 이용하여 데이터의 특징을 추출하는 방법이며, CNN은 여러개의 합성곱 층(convolution layer)를 이용하여 데이터의 특징을 추출하여 학습한다. 데이터의 차원에 따라 1d, 2d, 3d 합성곱 층을 이용하며, 순서가 있는 언어 데이터를 임베딩하여 사용하는 개체명 인식에서는 1d 합성곱 층을 이용한다.

CNN 모델의 경우 데이터의 일부분의 특징을 각각 추출하여 학습하기 때문에 전체 입력 데이터의 순서나 의존성을 고려하지 못한다. 순서가 있는 데이터를 처리하는 개체명 인식에서는 LSTM 모델보다 낮은 성능을 보였다. 하지만 순서를 고려하여 학습이 진행되는 LSTM 모델보다 CNN이 빠르다는 장점이 있다.

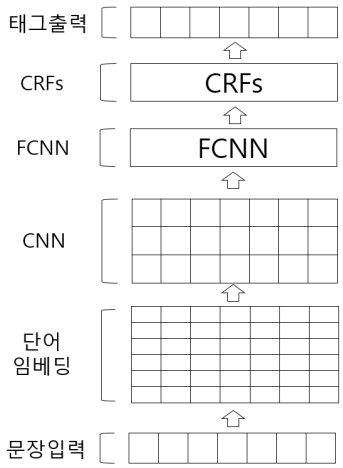
3.2.2 CNN-CRFs 모델

제안하는 모델에서는 1d 합성곱 층과 CRFs를 같이 사용하여 출력 결과 사이의 의존성을 학습하도록 하여 성능을 향상시켰다. CNN 모델은 입력되는 데이터의 순서나 의존성을 학습할 수 없지만 CRFs를 같이 사용한 CNN-CRFs 모델은 CRFs 층(layer)에서 출력되는 결과의 순서와 의존성을 학습할 수 있다. 또한 기존의 개체명 인식에서 사용되었던 bi-LSTM-CRFs 모델보다 CNN-CRFs 모델이 처리 속도가 빠르다는 장점이 있다.



[그림 2] CNN 구조도

본 논문에서는 bi-LSTM-CRFs 모델과의 비교를 위해 [그림 3]과 같이 합성곱 층을 1개만 사용하여 실험한다.



[그림 3] CNN-CRFs 구조도

4. 실험

본 실험에서는 한국어처리를 위한 기초 분석 도구 및 서비스 개발 과제를 통해 해양대학교에서 배포한 개체명 인식용 말뭉치[9]를 사용하여 실험하였다. 개체명의 범주는 인명(PER), 지명(LOC), 기관명(ORG), 기타(POH)로 나누어진다. 시간표현은 날짜(DAT), 시간(TIM), 기간 (DUR)이 있으며, 수량표현은 통화(MNY), 비율(PNT), 기타 수량표현(NOH)이 있다.

공개된 말뭉치는 형태소 분석된 정보가 포함되어 있지만, 본 실험에서는 형태소 분석에서 발생할 수 있는 오류를 최소화 하고자 말뭉치를 음절 단위로 변환하였다. 변환된 말뭉치는 BIO 태그를 사용하여 음절 단위에서 개

체명의 시작과 끝을 구분하였다.

본 실험에서는 구글(Google)에서 오픈소스로 공개한 Tensorflow 프레임워크를 이용하여 총 5가지 모델을 비교하였다. 단어 임베딩 사이즈는 128, Dropout rate는 0.2로 모두 동일하며 그 외 파라메터는 [표 2]와 같다.

모델	hidden units	filter size	filters
bi-LSTM	64	-	-
bi-LSTM-CRFs	64	-	-
CNN	-	5	64
CNN-CRFs	-	5	64
CNN-CRFs-2	-	5	256

표 2 모델 파라메터

평가방법은 macro average를 이용한 precision, recall, F1-score와 180 음절의 문장 2000개를 처리하는 평균을 소요시간을 사용하였다.

모델	prec	recall	F1	소요시간
bi-LSTM	84.4	80.9	82.4	2.07초
bi-LSTM-CRFs	85.5	81.3	**83.2**	**5.75초**
CNN	77.1	65.2	69.2	0.13초
CNN-CRFs	86.5	80.0	83.0	3.85초
CNN-CRFs-2	86.1	81.6	**83.6**	**4.17초**

표 3 모델 성능 비교

본 실험에서 사용한 장비는 [표 1]과 같다.

CPU	Intel® Xeon® CPU E5-2630 v4
RAM	32GB
GPU	Geforce GTX 1080

표 1 실험 장비 환경

실험 결과 [표 3]과 같이 bi-LSTM 모델과 CRFs를 결합한 경우 성능이 약간 향상되었지만, 속도가 매우 느려졌다. CNN 모델은 속도는 빠르지만, bi-LSTM 모델보다 매우 낮은 성능을 보였다. 64개 필터를 사용한 CNN-CRFs 모델은 bi-LSTM-CRFs 모델과 비슷한 성능을 보였고, 256개 필터를 사용한 CNN-CRFs-2 모델은 bi-LSTM-CRFs 모델보다 0.4% 높은 성능과 27.5% 빠른 처리 속도를 보였다.

5. 결론

본 논문에서는 LSTM 모델을 사용한 기존 한국어 개체명 인식 연구와는 달리 CNN 모델을 사용하였다. 입력 데이터의 순서와 상호 의존성을 학습할 수 없었던 CNN의 단점을 CRFs와 결합하여 개선하였다. 실험 결과 기존 모델인 bi-LSTM-CRFs 모델보다 CNN-CRFs 모델이 우수한 성능과 빠른 처리 속도를 보였다.

향후 연구로는 합성곱 층, 풀링 층(pooling layer)을 추가하여 한국어 개체명 인식에 최적화된 모델로의 구조개선 등이 있다.

참고문헌

[1] 유홍연, 고영중, “Bidirectional LSTM CRF 기반의 개 체명 인식을 위한 단어 표상의 확장,” 정보과학회논문지 44권 제3호, pp306-313, 2017.

[2] 신유현, 이상구, “양방향 LSTM-RNNs-CRF를 이용한 한국어 개체명 인식기,” 제28회 한글 및 한국어 정보처리 학술대회 논문집, pp.340-341, 2016.

[3] 장윤정, 민태홍, 이재성, “Stacked Bi-LSTM-CRF 앙상블 모델을 이용한 개체명 인식,” 2018년 한국컴퓨터종합학술대회 논문집, pp.2049-2051, 2018.

[4] 천민아, 김창현, 박호민, 노경목, 김재훈, “Multi-Head Attention 방법을 적용한 문자 기반의 다국어 개체명 인식,” 제30회 한글 및 한국어 정보처리 학술대회 논문집, pp.167-171, 2018.

[5] 이창기, 김준석, 김정희, 김현기, “딥 러닝을 이용한 개체명 인식,” 한국정보과학회 2014년 동계학술발표회 논문집, pp.423-425, 2014.

[6] 이창기, “Long Short-Term Memory 기반의 Recurrent Neural Network를 이용한 개체명 인식,” 한국컴퓨터 종합학술대회 논문집, pp.645-647, 2015.

[7] 최경호, 황현선, 이창기, “LSTM-CRF를 이용한 생명과학분야 개체명 인식,” 제27회 한글 및 한국어 정보처리 학술대회 논문집, pp.85-89 , 2015.

[8] 박성재, 차정원, “CNN을 이용한 대화와 같은 짧은 문장에서 개체명 인식,” 2017년 한국컴퓨터종합학술대회 논문집, pp.596-598, 2017.

[9] 한국어 개체명 말뭉치 배포 사이트 (https://github.com/kmounlp/NER)