양방향 LSTM-RNNs-CRF를 이용한 한국어 개체명 인식

신유현○, 이상구
서울대학교
shinu89@europa.snu.ac.kr, sglee@europa.snu.ac.kr

Bidirectional LSTM-RNNs-CRF for Named Entity Recognition in Korean

Youhyun Shin○, Sang-goo Lee
Seoul National University

요 약

개체명 인식은 질의 응답, 정보 검색, 기계 번역 등 다양한 분야에서 유용하게 사용되고 있는 기술이다. 개체명 인식의 경우 인식의 대상인 개체명이 대부분 새롭게 등장하거나 기존에 존재하는 단어와 중의적 의미를 갖는 고유한 단어라는 문제점이 있다. 본 논문에서는 한국어 개체명 인식에서 미등록어 및 중의성 문제를 해결하기 위한 딥 러닝 모델을 제안한다. 제안하는 모델은 형태소 및 자음/모음을 이용하여 새롭게 등장하는 단어에 대한 기존 단어와의 형태적 유사성을 고려한다. 또한 임베딩 및 양방향 LSTM-RNNs-CRF 모델을 이용하여, 각 입력 값의 문맥에 따른 의미적 유사성, 문법적 유사성을 고려한다. 제안하는 딥 러닝 모델을 사용하여, F1 점수 85.71의 결과를 얻었다.

주제어: 개체명 인식, 딥 러닝, 미등록어

1. 서론

개체명 인식은 질의 응답, 정보 검색, 기계 번역 등 다양한 분야에서 성능 향상을 위해 사용되는 기술이다. 개체명이란 하나 이상의 단어로 이루어지며, 문서에 나타나는 고유 명사에 해당하는 인명(Person name), 지명(Location name), 기관명(Organization name)을 의미하며, 숫자 혹은 시간과 같은 고유한 성질로도 확장되기도 한다. 개체명 인식은 문서에서 이러한 개체명을 추출하고 추출된 개체명의 종류를 인명, 지명, 기관명 등으로 분류하는 것을 말한다.

개체명의 경우 대부분이 새롭게 등장하는 고유한 단어이거나 기존에 존재하는 단어가 고유 명사로 쓰이는 경우이므로, 사전에 없거나 문맥에 따라 그 뜻이 다르게 사용되는 경우가 많다. 따라서 사전 기반의 방법을 사용하게 될 경우, 이러한 미등록어 및 중의성 문제로 인해 개체명 인식의 정확도가 높지 못하다. 최근에 연구되는 개체명 인식 알고리즘에는 주로 기계학습 방법이 사용되고 있다. 한국어 개체명 인식의 경우, HMM[1], CRF[2], CNN(Convolutional Neural Network)[3], LSTM Recurrent CRF[4] 등을 이용한 한국어 개체명 인식 알고리즘 연구가 존재한다.

본 논문에서는 딥 러닝 알고리즘을 이용하여, 미등록어 및 중의성을 해결하는 알고리즘을 제안한다. 문장의 형태소를 입력 값으로 받아, 해당 형태소의 임베딩 및 자음/모음 특징과 품사, 기구축 사전 정보를 이용하여 딥 러닝 모델을 학습하였다. 실험을 통해 해당 알고리즘의 성능을 측정하였다.

2. 양방향 LSTM-RNNs-CRF 모델

본 논문에서는 개체명 인식을 위한 학습 모델로 양방향(Bidirectional) LSTM-RNNs-CRF[5]를 이용하였다. 양방향 LSTM-RNNs-CRF의 경우 LSTM의 특징에 의해 입력 값 간의 긴 의존(long dependency) 정보 및 문맥 정보를 충분히 반영할 수 있다. 양방향 LSTM-RNNs-CRF 중 RNN은 자음/모음 특징 추출을 위해 사용된다. 양방향 LSTM 및 CRF는 각각 입력 시퀀스의 레이블링, 레이블 간의 의존 정보에 따른 보정에 사용된다.

양방향 LSTM-RNNs-CRF는 그림 1과 같이 형태소/품사 정보를 입력 값(ex. 첫/MM, 타자/NNG)으로 받아 형태소 임베딩, 자음/모음 특징, 품사, 기구축 사전의 4가지 벡터 표현을 생성한다. 생성된 4가지 벡터 표현은 양방향 LSTM의 입력 값으로 사용된다. 출력 값의 형태는 “O, O, B-PS, O, ...” 이며, 출력 값에 사용되는 인코딩(encoding) 기법에는 BIOES 방법을 사용하였다.

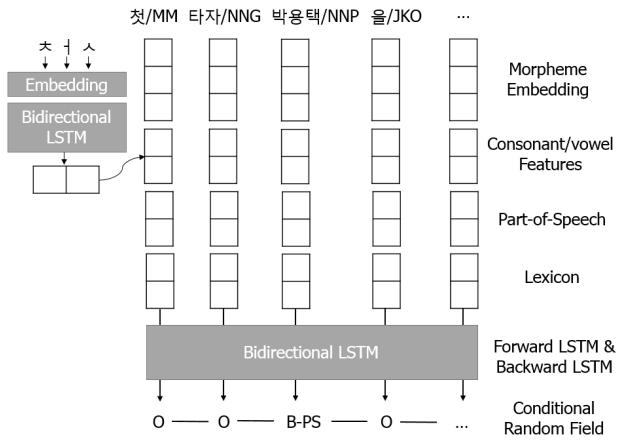

그림 1 개체명 인식 알고리즘의 전체 구조

2.1 형태소 임베딩 및 자음/모음 특징

새로운 단어가 등장하더라도 형태적 유사함을 토대로

기존 단어와의 연관성을 이용하기 위해 입력 값으로 형태소 및 형태소의 자음, 모음을 이용(ex. ㅎ, ㅕ, ㅇ, ㅌ, ㅐ, ㅅ, ㅗ)하였다. 또한 형태소 단위의 임베딩을 사용함으로써 형태적으로 유사한 단어의 의미적, 문법적 유사도를 고려하였다. 형태소 임베딩은 사전 학습(pre-train)된 워드 임베딩 값을 lookup table에서 가져온다. 자음/모음 특징은 RNN으로부터 추출된다. 첫/MM”의 경우에는, “ㅊ, ㅓ , ㅅ”을 입력 값으로 하는 양방향 LSTM layer를 추가하였다.

2.2 품사 및 기구축 사전

고유 명사와 고유 명사가 연속하여 나올 경우에는 개체명일 확률이 높다는 것과 같은 문법적 특징을 학습하기 위해 품사를 벡터로 나타내었다. 품사 벡터를 양방향 LSTM을 이용해 학습하여 각 품사 간의 의존(dependency) 관계를 학습하였다. 품사 표현 벡터는 그림 2의 왼쪽과 같이 품사 전체 개수를 길이로 하는 원 핫(one-hot) 벡터를 이용하였다.

기구축사전(lexicon) 정보를 이용하기 위해, BIOES 인코딩 방법[1]을 적용하였다. 기구축사전 벡터는 가능한 레이블의 개수를 길이로 하는 벡터로, 사전 내에서 해당 단어가 어떻게 쓰였는지에 따라 정해진 레이블에 0 또는 1의 값을 갖는다. 예를 들어, “박용택 PS”, “LA 레이커스 OG”, “LA LC” 세 가지 경우가 있다고 하자. 이 경우 “박용택”은 “박용택 PS”에서만 쓰였으므로 “S-PS”의 값만 1이다. “LA”의 경우 “B-OG, S-LC”의 값이 1이며, “레이커스”는 “E-OG”의 값만 1이다.

3. 실험

3.1 실험 환경

한국어 개체명 인식 시스템 성능 측정에는 3,555개의 학습 문장과 501개의 테스트 문장이 사용되었다. 각 문장은 그림 2와 같이 형태소 분석 및 품사 태깅이 된 상태이다. 해당 형태소에 대한 개체명 학습은 총 50번의 에폭(epoch)을 통해 이루어졌다. 성능 측정에는 CoNLL 평가 스크립트를 이용하여, [B-PS], [B-PS, I]과 같이 개체명의 청크(chunk) 단위로 성능을 측정하였다.

```
첫  MM  O
다자  NNG O
박용택  NNP B-PS
을  JKO O
```

그림 2 입력 형식 예시

3.2 실험 결과

표 1에서 볼 수 있듯이 각 특징 별 알고리즘으로는 4가지 벡터 표현 방법을 모두 사용한 경우 F1 점수 85.71로 가장 좋은 성능을 보였다. 표 2는 가장 좋은 성능을 보인 4번째 모델에 대한 개체형 별 성능을 나타내고 있다. 5가지의 개체명 중 인명(PS)에 해당하는 F-Score가 0.899로 제일 높고, 기관명(OG)에 해당하는 F-Score가 0.797로 제일 낮다. 전반적으로는 인명 > 날짜 > 시간 > 지명 > 기관명 순으로 좋은 성능을 보였다.

	특징 (Feature)	F1 점수
1	형태소 임베딩	79.69
2	형태소 임베딩+자음/모음 특징	80.66
3	형태소 임베딩+자음/모음 특징+품사	83.09
4	형태소 임베딩+자음/모음 특징+품사+사전	85.71

표 1 특징에 따른 F-Score 값

	정밀도	재현율	F1 점수
인명(PS)	91.5	88.5	89.9
지명(LC)	79.3	85.3	82.2
기관명(OG)	82.4	77.2	79.7
날짜(DT)	89.4	88.0	88.7
시간(TI)	87.2	81.0	84.0

표 2 개체형 별 정밀도, 재현율, F 1Score

4. 결론

본 논문에서는 양방향 LSTM-RNNs-CRF를 이용한 한국어 개체명 인식 알고리즘을 제안하였다. 해당 알고리즘은 “형태소/품사”의 입력 값을 받아 형태소 임베딩, 자음/모음 특징, 품사, 기구축 사전 정보의 4가지를 이용하여 개체명을 인식하였다. 그 결과 85.71의 F1 점수를 얻었다. 개체형 기준으로는 인명(PS)의 경우 F1 점수 89.9로 가장 좋은 성능을 보였다.

향후 연구로는 개체명 인식 이외의 한국어 자연어 처리에 사용되는 시퀀스 레이블링 문제에 해당 알고리즘을 적용할 예정이다.

참고문헌

[1] 황이규, 윤보현, "HMM 에 기반한 한국어 개체명 인식", 정보처리학회논문지(B), 제10권, 제2호, pp.229-236, 2003.

[2] 이창기, 황이규, 오효정, 임수종, 허정, 이충희, 김현진, 왕지현, 장명길, "Conditional Random Fields 를 이용한 세부 분류 개체명 인식", 제 18 회 한글 및 한국어 정보처리 학술대회, pp. 268-272, 2006.

[3] 이창기, 김준석, 김정희, 김현기, "딥 러닝을 이용한 개체명 인식", 한국정보과학회 제 41 회 정기총회 및 동계학술발표회, pp. 423-425, 2014.

[4] 이창기, "Long Short-Term Memory 기반의 Recurrent Neural Network 를 이용한 개체명 인식", 한국정보과학회학술발표논문집, pp. 645-647, 2015.

[5] Xuezhe Ma, Eduard Hovy. "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF", arXiv preprint arXiv:1603.01354, 2016.

1) B(Begin), I(Inside), O(outside), E(End), S(Single)의 5가지 태그를 이용하여 인코딩 하는 방법