딥러닝 기반의 개체명 인식을 위한 효과적인 사전 자질 사용 방법

김홍진○, 김학수
강원대학교 컴퓨터정보통신공학과
jin3430@gmail.com, nlpdrkim@kangwon.ac.kr

How to Use Effective Dictionary Feature for Deep Learning based Named Entity Recognition

Hong-Jin Kim○, Hark-Soo Kim
Kangwon National University Department of Computer and Communications Engineering

요 약

개체명 인식은 입력 문장에서 인명, 지명, 기관명, 날짜, 시간과 같이 고유한 의미를 갖는 단어들을 찾아 개체명을 부착하는 기술이다. 최근 개체명 인식기는 형태소 단위나 음절 단위의 입력을 사용하는 연구가 주로 진행되고 있다. 그러나 형태소 단위 개체명 인식은 미등록어를 처리하지 못하는 문제점이 존재하고 음절 단위 개체명 인식은 단어의 의미를 제대로 반영하지 못하는 문제점이 존재한다. 본 논문에서는 이 문제점을 보완하기 위해 품사 정보를 활용한 음절 단위 개체명 인식기를 제안한다. 또한 개체명 인식 성능에 큰 영향을 미치는 개체명 사전 자질을 더 효과적으로 사용할 수 있는 방법을 제안하며 이 방법을 사용했을 때 기존의 방법보다 향상된 개체명 인식 성능(F1-score 0.8576)을 보였다.

주제어: 개체명 인식, 딥러닝, Stacked BiGRU-CRFs

1. 서론

개체명 인식(Named Entity Recognition)은 입력된 문장에서 인명, 지명, 기관명, 날짜, 시간과 같이 고유한 의미를 갖는 단어들을 찾아 개체명을 부착하는 기술이다. 대부분의 개체명 인식 연구에서 입력으로 형태소 단위나 음절 단위를 사용한다[1-5]. 형태소 단위 개체명 인식기에서는 신조어와 같은 미등록어의 개체명 인식에 어려움이 있다[1,2]. 음절 단위 개체명 인식기에서는 단어가 가지는 의미가 잘 반영되지 않는 문제가 있다[3,4]. 본 논문에서는 음절 단위의 품사 태깅(Tagging) 결과를 사용하여 미등록어가 입력되었을 때 형태소 분석 오류가 전파되는 문제를 보완한다. 각 음절 품사 정보를 임베딩 자질(Embedding Feature)로 사용하여 단어의 의미 정보를 보강한 모델을 제안한다. 또한 본 논문은 다층 순환신경망(Stacked Recurrent Neural Network) 기반 모델에서 개체명 사전 자질을 보다 효과적으로 사용하기 한 새로운 방법으로 마지막 계층(Layer)에 자질을 추가하는 방법을 제안한다.

2. 관련 연구

기존 개체명 인식 연구에는 기계학습 알고리즘이 많이 사용되었으며 대표적인 모델로 Structural SVM(Support Vector Machine), CRFs(Conditional Random Fields) 등이 있다[6]. 기계학습 기반 개체명 인식 연구에서는 사람이 직접 자질을 설계하여 입력으로 주어야 했지만 최근 이러한 단점을 보완하기 위해 딥러닝(Deep-learning)을 이용한 개체명 인식 연구가 활발히 이루어지고 있다. 특히 딥러닝 기법 중 순차적 레이블링(Sequence Labeling) 문제에 좋은 성능을 보이는 양방향 순환신경망(Bidirectional Recurrent Neural Network)과 CRFs를 결합한 모델인 BiLSTM(Bidirectional Long Short-Term Memory)-CRFs 모델이 개체명 인식 연구에서 우수한 성능을 보였다[3-5,7]. LSTM에 비해 속도가 빠른 GRU(Gated Recurrent Unit)를 적용한 BiGRU-CRFs 모델을 사용하여 개체명 인식을 수행한 연구도 있다[8]. 이러한 딥러닝을 이용한 개체명 인식 연구의 성능을 향상시키기 위한 방법으로 모델의 구조를 개선시키는 방법, 입력되는 임베딩을 확장시키는 방법, 그리고 개체명 사전 등의 자질을 사용하는 방법이 있다[5]. 모델의 구조를 개선시키는 방법으로, [5]는 음절 기반 개체명 인식 모델과 형태소 기반 개체명 인식 모델의 앙상블(Ensemble)을 통해 결과를 보완하여 성능의 향상을 보였다. 순환신경망을 다층으로 구성하여 자연어처리 분야에서 좋은 성능을 보인 연구도 있었다[9]. 입력되는 임베딩을 확장시키는 방법으로는 대용량의 말뭉치를 사전 학습(Pre-training)하여 사용하거나 음절 단위 임베딩에서 단어 단위 임베딩을 유도하는 방법이 있다[3]. 또한, 개체명 사전 자질을 사용해 개체명 인식의 성능을 크게 향상시킨 연구들도 있다[3,4,8]. 이러한 연구들을 통해 입력 문장에서 단어의 개체명 정보를 주는 것은 개체명 인식을 위한 중요한 자질임을 알 수 있다. 본 논문에서 제안하는 모델은 두 개의 양방향 GRU 계층과 하나의 CRFs 계층으로 구성되며, 단어 임베딩을 확장하기 위해 음절 임베딩과 품사 임베

딩 그리고 개체명 사전 자질 벡터(Vector)를 추가한다. 또한 개체명 사전 정보를 더 잘 반영할 수 있도록 양방향 GRU의 가장 마지막 계층에 개체명 사전 자질 정보를 입력하는 새로운 개체명 사전 자질 사용 방법을 제안한다.

3. 음절 단위 개체명 인식기

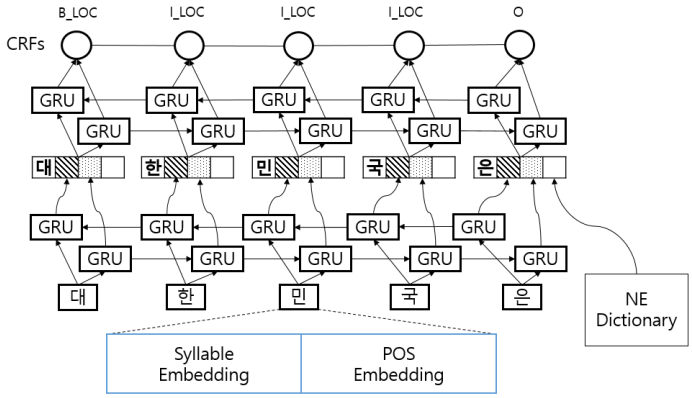

그림 1 음절 단위 개체명 인식기 전체 구조도

3.1 Stacked BiGRU-CRFs

그림 1은 음절 단위 개체명 인식기의 전체 구조도이며 두 층으로 쌓은(Stacked) 양방향 GRU 계층과 CRFs 계층으로 구성되어 있다. 첫 번째 양방향 GRU 계층은 각 음절의 임베딩 벡터와 품사 임베딩 벡터를 연결(Concatenation)한 값을 입력받아 양방향의 문맥 정보가 반영된 인코딩 벡터를 출력한다. 두 번째 양방향 GRU 계층에서는 첫 번째 계층에서 사용한 입력 벡터, 첫 번째 층의 인코딩 벡터와 개체명 사전 자질 벡터를 연결하여 입력으로 사용한다. CRFs 계층에서는 두 번째 양방향 GRU 계층의 결과 값을 입력으로 사용해서 인접한 결과 간의 전이 확률이 반영된 음절 단위 개체명 태그를 부착한다.

3.2 임베딩

본 논문에서 음절 임베딩은 무작위로 초기화(Random Initialize)하여 사용한다. 양방향 순환신경망 모델은 입력계층에서 사용하는 임베딩에 따라 성능이 좌우된다[4]. 음절 임베딩만으로 단어를 표현하기에 부족하고, 한국어의 개체명은 모두 명사에 속하기 때문에 개체명 인식에는 품사 정보가 중요하다. 즉, 음절 단위 개체명 인식에 품사 정보를 활용하면 성능의 향상을 보일 수 있다. 따라서 본 논문에서는 각 음절의 품사 정보를 주기 위하여 무작위로 초기화한 음절 품사 임베딩을 사용한다. 음절이 형태소의 시작 음절이면 “B” 태그를 부착하였고, 형태소의 중간이나 끝 음절이면 “I” 태그를 부착하였다. “대한민국은”이라는 어절에 대응되는 음절 품사 태그는 표 1과 같다.

표 1 음절 단위 품사 태그 예시

	대	한	민	국	은
품사 태그	B_NNG	I	I	I	B_JK

3.3 개체명 사전

개체명 사전은 개체명으로 분류되는 단어들을 사전으로 구축한 것이다. 특정 단어가 개체명 사전에 포함되어 있다는 정보를 주면 개체명 인식 성능을 향상 시킬 수 있다. 본 논문에서 개체명 사전 자질을 생성한 방법은 다음과 같다. 현재 입력 음절과 양 옆의 음절들을 조합하여 만들 수 있는 2 음절(Bigram) 단어부터 5 음절(Pentagram) 단어가 개체명 사전에 존재하면 “1”, 그렇지 않으면 “0”으로 표기한다. 예를 들어, “홍길동이다”라는 어절에서 현재 입력 음절이 “동”인 경우 조합되는 단어들은 표 2와 같다.

표 2 입력 음절 기준으로 조합되는 단어

	조합되는 단어
홍길동이다	“길동”, “동이”, “홍길동”, “길동이”, “동이다”, “홍길동이”, “길동이다”, “홍길동이다”

한 음절에 대응되는 개체명 사전 자질 차원 수는 8개(Bigram * 2, Trigram * 3, Quadgram * 2, Pentagram * 1)의 N 음절 단어와 5개의 개체명 태그(인명, 지명, 기관명, 날짜, 시간)의 조합으로 40차원이 된다. 표 2와 같은 예제에서 개체명 사전 자질은 표 3과 같다.

표 3 개체명 사전 자질 벡터 예시

	인명	지명	기관명	날짜	시간
길동	1	0	0	0	0
동이	1	0	0	0	0
홍길동	1	0	0	0	0
길동이	1	0	0	0	0
...	...	...	...	...	...
홍길동이다	0	0	0	0	0

3.4 효과적인 개체명 사전 자질 사용 방법

본 논문에서는 효과적인 개체명 사전 자질 벡터 반영 방법으로 개체명 사전 자질을 그림 1과 같이 두 번째 계층 입력에 반영하는 방법을 제안한다. 그림 1의 첫 번째 계층 입력으로 개체명 사전 자질 벡터를 준다면, 상위 층으로 갈수록 개체명 사전 자질이 가지고 있는 정보가 희석되어 CRFs 계층에 잘 반영되지 않는 문제가 있다. 반면 두 번째 계층 입력으로 사용하면 개체명 사전 자질이 가진 정보를 CRFs 계층에 최대한 반영할 수 있다.

4. 실험 및 결과

4.1 실험 환경

본 논문에서는 2017 국어 정보 처리 시스템 경진대회의 말뭉치를 실험에 사용한다. 전체 말뭉치는 3,660 문장이며 3,294 문장을 학습 데이터로, 366 문장을 평가 데이터로 사용하였다. 개체명 종류는 인명, 지명, 기관명, 날짜, 시간 5개이고 각 음절의 개체명 경계 구분은 BIO 태그를 이용하였다.

4.2 실험 결과

4.2.1 개체명 사전 자질 실험

본 논문에서 사용한 개체명 사전은 학습데이터에 나타난 개체명을 추출하여 구축하였다. 표 4에서 BiGRU-CRFs는 단층 양방향 순환신경망 모델이다. Stacked BiGRU-CRFs는 그림 1의 구조와 같이 BiGRU 계층을 다층으로 쌓은 모델이며, 개체명 사전 자질의 유무에 따른 실험 결과를 비교했다. Stacked BiGRU-CRFs는 개체명 사전 자질을 사용한 기존 모델의 방법처럼 첫 번째 계층에 자질을 추가한 성능이다. 표 4에서 BiGRU-CRFs가 단층일 경우 보다 다층일 경우에 F1 성능이 2.29% 향상됨을 보였다. 개체명 사전 자질을 추가하였을 때 단층 BiGRU-CRFs는 F1-score가 5.93% 향상되었으며 Stacked BiGRU-CRFs는 5.26%의 성능 향상을 보였다. 결과적으로 BiGRU를 다층으로 쌓은 모델인 Stacked BiGRU-CRFs에 개체명 사전 자질을 추가한 실험이 가장 높은 성능을 보였다.

표 4 개체명 사전 자질 실험 결과

	precision	recall	F1-score
BiGRU-CRFs	0.8184	0.7022	0.7554
BiGRU-CRFs + 개체명 사전	0.8655	0.7695	0.8147
Stacked BiGRU-CRFs	0.8328	0.7305	0.7783
Stacked BiGRU-CRFs + 개체명 사전	**0.8800**	**0.7871**	**0.8309**

4.2.2 개체명 사전 자질 사용 방법 실험

표 5는 개체명 사전 자질 사용 방법에 따른 실험 결과이다. 하위 층 입력에 반영하는 방법보다 상위 층 입력에 반영하는 방법이 2.67% 더 높은 F1-score 성능을 보였다. 특히, 정밀도(Precision)는 거의 동일 하지만 재현율(Recall) 성능이 큰 폭으로 증가했다. 이는 상위계층의 입력으로 개체명 사전 자질을 주는 것이 특정 단어가 개체명 사전에 포함되어 있다는 정보가 효과적으로 반영되어 성능의 향상을 보인 것으로 판단된다.

표 5 개체명 사전 자질 사용 방법에 따른 실험결과

	precision	recall	F1-score
기존 사전 자질 반영 방법	**0.8800**	0.7871	0.8309
제안 방법	0.8778	**0.8383**	**0.8576**

5. 결론

본 논문에서는 BiGRU-CRFs을 다층으로 구성하여 음절 단위 개체명 인식을 수행했다. 각 음절의 품사 임베딩을 사용하여 미등록어 문제를 보완하면서 단어의 의미 정보를 보강하였다. 또한 효과적인 개체명 사전 자질 사용을 위한 새로운 반영 방법을 제안하였다. 실험 결과, 제안하는 개체명 사전 자질 반영 방법을 사용했을 경우 기존 방법 보다 향상된 개체명 인식 성능을 보였다. 향후 연구에서는 [5,8]과 같이 단어 표상을 더 확장하고, 3개 이상의 BiGRU-CRFs 계층 모델에서 개체명 사전 자질 뿐만 아니라 형태소 분석, 구문 분석 등과 같은 다양한 분야에 쓰이는 자질들의 효과적인 사용방법에 대하여 연구할 것이다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2013-0-00109,WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] 이성희, 송영길, 김학수, "원거리 감독과 능동 배깅을 이용한 개체명 인식", *정보과학회논문지,* 제43권, 제2호, pp. 269-674, 2016.

[2] 최윤수, 차정원, "Word Embedding 자질을 이용한 한국어 개체명 인식 및 분류", *정보과학회논문지*, 제43권, 제6호, pp. 678-675, 2016.

[3] 나승훈, 민진우, "문자 기반 LSTM CRF를 이용한 개체명 인식", *한국정보과학회 2016년 한국컴퓨터종합학술대회 논문집*, pp. 729-731, 2016.

[4] 유홍연, 고영중, "Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장", *정보과학회논문지*, 제44권, 제3호, pp. 306-313, 2017.

[5] 박건우, 박성식, 장영진, 최기현, 김학수,

"KACTEIL-NER: 딥러닝과 앙상블 기법을 이용한 개체명 인식기", *제29회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 324-326, 2017.
[6] 이창기, 김준석, 김정희, 김현기, "딥 러닝을 이용한 개체명 인식", *한국정보과학회 제41회 동계학술발표회*, pp. 423-425, 2014.
[7] 이창기, "Long Short-Term Memory 기반의 Recurrent Neural Network를 이용한 개체명 인식", *한국정보과학회 2015년 한국컴퓨터종합학술대회*, pp. 645-647, 2015.
[8] 김선우, 최성필, "Bidirectional GRU-CRF 기반의 한국어 개체명 인식을 위한 어휘 사전 자질 적용 네트워크 토폴로지 연구", *정보과학회 컴퓨팅의 실제 논문지*, 제25권, 제2호, pp. 99-105, 2019.
[9] 배장성, 이창기, "Stacked Bidirectional LSTM-CRFs를 이용한 한국어 의미역 결정", *정보과학회논문지*, 제44권, 제1호, pp. 36-43, 2017.