한국어 어휘의미망[1]을 활용한 Conditional Random Fields 기반 한국어 개체명 인식

박서연○, 옥철영, 신준철
울산대학교 한국어처리연구실
seoyeon9695@gmail.com, okcy@ulsan.ac.kr, ducksjc@gmail.com

Conditional Random Fields based Named Entity Recognition Using Korean Lexical Semantic Network

Seo-Yeon Park○, Cheol-Young Ock, Joon-Choul Shin
University of Ulsan, Korean Language Processing Lab

요 약

개체명 인식은 주어진 문장 내에서 OOV(Out of Vocaburary)로 자주 등장하는 고유한 의미가 있는 단어들을 미리 정의된 개체의 범주로 분류하는 작업이다. 최근 개체명이 문장 내에서 OOV로 등장하는 문제를 해결하기 위해 외부 리소스를 활용하는 연구들이 많이 진행되었다. 본 논문은 의미역, 의존관계 분석에 한국어 어휘지도를 이용한 자질을 추가하여 성능 향상을 보인 연구들을 바탕으로 이를 한국어 개체명 인식에 적용하고 평가하였다. 실험 결과, 한국어 어휘지도를 활용한 자질을 추가로 학습한 모델이 기존 모델에 비해 평균 1.83% 포인트 향상하였다. 또한, CRF 단일 모델만을 사용했음에도 87.25% 포인트라는 높은 성능을 보였다.

주제어: 개체명 인식, 기계학습, 어휘의미망, CRF

1. 서 론

개체명 인식(Named Entity Recognition)은 주어진 문장 내에서 OOV(Out of Vocaburary)로 자주 등장하는 인명이나 장소, 기관명 등과 같이 고유한 의미가 있는 단어들을 미리 정의된 개체의 범주로 분류하는 작업이다. 이를 전처리로 사용하면 정보검색이나 질의응답, 기계번역, 대화 시스템 등의 자연어 처리 분야의 성능을 향상 시킬 수 있는 중요한 분야이다.

최근에는 개체명이 문장 내에서 OOV로 등장하는 문제를 해결하기 위하여 개체명 사전을 구축하는 등의 외부 리소스를 활용하는 연구들이 진행되었다[1, 8]. 본 논문에서는 의미역[2], 의존관계 분석[4]에 한국어 어휘지도(UWordMap)[4]를 이용한 자질을 추가하여 성능 향상을 보인 연구들을 바탕으로 한국어 개체명 인식에 적용하고 평가한다.

본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장은 한국어 어휘지도를 이용한 자질 추가 방법을 소개한다. 4장에서는 실험 결과를 분석하고, 마지막 5장에서 결론에 관해 기술한다.

2. 관련 연구

개체명 인식 문제는 사전기반 및 규칙기반을 사용하는 전통적인 방법과 개체명 태그를 부착한 말뭉치를 학습하여 분석하는 학습기반 방법으로 나뉜다. 규칙기반과 사전기반의 개체명 인식 방법은 정의된 규칙과 사전에 대해서는 정확한 분석이 가능하고 분석 속도가 빠르다는 장점이 있지만, 미리 식별된 특정 데이터에만 유효하여 OOV에 취약하며 규칙과 사전을 수동으로 구성을 해야 하는 문제가 있다.

학습기반 방법은 개체명 태그가 부착된 말뭉치를 학습에 사용하여 문장의 성분들을 조합하고 선택하여 효과적인 자질을 찾아내어 사용한다. 이를 바탕으로 순차적 레이블링(sequential labeling)에 효과적인 기계학습과 딥러닝 모델에 적용한 개체명 인식 연구가 진행되었다. [5]는 기계학습 모델인 Structural SVMs와 Pegasos 알고리즘을 이용한 개체명 인식 모델을 제안하였다. [6]은 또 다른 기계학습 모델인 CRF를 이용하여 특허 문서에 대한 개체명 인식 태그를 학습 및 평가한 연구를 진행하였다. 딥러닝 기법 중에는 순차적 레이블링에 특화된 RNN 계열의 LSTM 기반 방법인 bi-LSTM-CRF를 이용한 개체명 인식 연구가 가장 많이 진행되었으며 가장 높은 성능을 보였다. [1, 7-10] 딥러닝 모델은 자질 추출을 위한 노력이 줄어 많이 연구되고 있지만, 고성능의 컴퓨팅 파워가 요구되며 학습 모델의 속도가 느려 실용성이 낮은 단점이 있다.

본 논문에서는 정확률과 처리 속도를 모두 고려하여 딥러닝 방식보다 빠른 속도로 학습 및 분석을 할 수 있는 기계학습 방식을 사용한다. 일반적으로 기계학습 방식을 사용한 개체명 인식 방법의 정확률이 딥러닝을 사용한 방법에 비해 낮다는 점을 보

1) 한국어 어휘의미망 : UWordMap®

완하기 위하여 표준국어대사전을 기반으로 한 한국어 어휘지도에서 단어의 상위어 정보와 고유 명사를 자질로 하여 CRF 모델에 적용하는 방법을 제안한다.

3. 한국어 어휘지도를 활용한 Conditional Random Fields 기반 한국어 개체명 인식

3.1 한국어 어휘지도의 상위어를 이용한 자질

본 논문에서는 개체명이 사전에 학습되지 않은 OOV로 등장하는 문제를 해결하기 위하여 한국어 어휘지도(UWordMap)를 활용하였다. UWordMap은 표준국어대사전을 기반으로 명사, 용언, 부사 등의 어휘들이 의미제약으로 상호 연결된 어휘의미망이다.

UWordMap은 명사의 계층 구조를 담고 있어 명사의 상위어와 하위어 정보를 얻을 수 있다. 상위어 정보는 대상 단어보다 큰 범주의 의미가 있는 단어이므로 상위어 정보를 통해 학습 데이터를 확장하는 역할과 개체명 인식을 위한 키워드 역할을 할 수 있다. 예를 들어, '**베이징**'은 표 1과 같은 계층별 상위어를 가진다. 이를 바탕으로 '베이징'에 대해 1계층 상위 단어인 '수도'를 자질로써 사용하여 학습한다면, 학습 데이터에 등장하지 않은 '하노이(베트남의 수도)'라는 단어가 등장했을 때, '베이징'과 같은 상위어를 가지고 있는 단어이기에 이를 개체명으로 인식할 수 있다.

표 1. '베이징'의 계층별 상위어

계층	단어
1	공간
2	지역
3	도시
4	수도
5	**베이징**

표 2는 3계층 단어가 '**사람**'인 단어들의 계층 구조를 나타낸 것이다. 개체명인지 파악하고자 하는 단어에 대해 주변 단어 중, 표 2와 같이 3계층 단어가 '사람'인 단어가 등장한다면 이는 현재 파악하고자 하는 단어가 사람이라는 개체로 분류될 수 있다는 것을 의미하므로 이를 키워드로써 사용할 수 있다. 예를 들어, '외국인 투수 바르가스가~'라는 문장에서 현재, '바르가스'가 개체명인지 파악하고자 한다면 주변 단어인 '투수'라는 키워드를 통해 인명으로 분류할 수 있다.

상위어를 자질로 사용하는 과정에서 표층형이 같더라도 다른 상위어 계층을 가질 수 있다. 해당 어휘가 동형이의어이거나 다의어일 경우 표층형이 같더라도 의미상으로 다른 상위어를 가지기 때문이다. 본 연구에서는 상위어 자질을 사용할 때, 형태소 및 동형이의어, 다의어 분별 시스템인 UTagger[11]을 통해 각 명사에 대해 다의어까지 분석하여 상위어 자질을 추가하였다.

표 2. 3계층 단어가 '사람'인 단어들의 계층 구조

계층	단어			
1	생물	생물	생물	생물
2	동물	동물	동물	동물
3	**사람**	**사람**	**사람**	**사람**
4	직업인	직업인	구성원	인물
5	체육인	체육인	의원	가공인물
6	선수	선수	국회의원	등장인물
7	운동선수	운동선수		주인공
8	투수	골키퍼		여주인공

3.2 간접적인 의존관계 자질

본 논문은 기존에 의존관계 정보를 개체명 인식에 적용하여 성능 향상을 보였던 연구들[12, 13]을 바탕으로 간접적인 의존관계 정보를 자질로 사용하였다. 간접적인 의존관계 정보 자질을 추가하는 과정은 다음 그림 1과 같다.

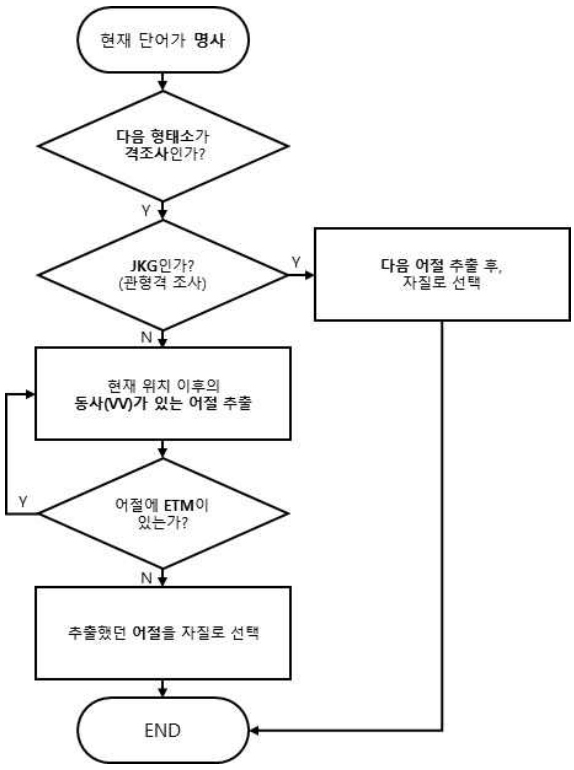

그림 1. 간접적인 의존관계 자질을 추출하는 과정

그림 1의 시작 조건은 현재 단어가 명사일 때이다. 명사인 경우, 명사 다음의 형태소가 격조사인지 확인한다. 이 조건을 만족한다면, 해당 격조사가 관형격 조사(JKG)일 때와 아닐 때로 나뉘어 추출하는 자질이 달라진다. 먼저 관형격 조사일 때는 다음 어절을 지배소 자질로, 현재 명사를 의존소 자질로 선택한다. 추출한 어절에 명사가 포함된 경우, 3.1과 같이 상위어 자질을 추가한다.

관형격 조사가 아닐 때는 현재 위치 이후에 나오는 동사(VV)가 포함된 어절 중, 가장 가까운 어절을 추출한다. 이때, 추출한 어절 내에 관형형 전성 어미(ETM)가 포함된 경우는 제외한다. 선택된 어절에 관형형 전성 어미가 포함되지 않은 경우, 이를 지배소 자질로, 현재 명사를 의존소 자질로 선택한다.

3.3 개체명 인식기의 학습 자질 정보

본 연구에서는 한국어 어휘지도를 활용한 Conditional Random Fields 기반 한국어 개체명 인식기를 학습하기 위해 표 3과 같은 자질을 사용하였다. 기본 학습 자질로는 현재 위치를 기준으로 앞, 뒤 3개의 형태소에 대해서 어휘 정보와 품사 태그의 조합을 학습한다. 기분석 사전 자질은 학습 데이터에서 3회 이상 같은 태그로 분류되는 개체명을 모아놓은 사전이다. 예를 들어, '한국'은 학습 시에 LC(지명)과 OG(기관)으로 분류된 경우가 모두 3번 이상이기 때문에 기분석 사전에 이 정보를 가지고 있게 된다. 유의어 자질은 형태소 분석을 완료한 학습 데이터를 이용하여 word2vec skip-gram으로 학습한 후, 코사인 유사도를 계산하여 유의어 자질로 선택한 것이다. 이때, 유의어 자질로 선택하는 단어는 현재 위치의 단어와 형태소 태그가 같은 단어 중 가장 유사도가 높은 단어로 선택한다. 유의어 기분석 자질은 형태소 태그가 같은 단어 중 코사인 유사도가 가장 높은 5개의 단어를 선택한 후, 기분석 사전에서 검색하여 유의어들이 가장 많이 분류된 개체와 두 번째로 많이 분류된 개체 정보이다. 본 논문에서 제안하는 추가 자질인 상위어와 간접적 의존관계 자질은 각각 3.1과 3.2에서 서술한 정보이다.

표 3. 개체명 인식기 학습 자질 정보

자질 정보	설명
형태소 어휘	(-3, 3) 위치의 형태소 어휘 정보
형태소 품사	(-3, 3) 위치의 형태소 품사 정보
접두, 접미부 음절	현재 형태소의 접두, 접미부 음절
기분석 사전	학습 데이터에서 3회 이상 같은 개체로 분류된 형태소
유의어	학습 셋을 이용한 워드 임베딩 후, 코사인 유사도 계산을 통한 정보
유의어 기분석 자질	유의어를 이용한 기분석 사전 자질
상위어	1계층 상위 단어, 3계층 어휘
간접적 의존관계	지배소와 의존소 정보

4. 실 험

본 논문은 제안한 한국어 어휘지도를 활용한 CRF 기반 한국어 개체명 인식 방법의 성능 평가를 위해 CRF를 빠르고 간단하게 적용할 수 있는 CRFsuite로 구현하였다. 개체명 인식 성능 측정을 위한 데이터로는 엑소브레인 개체명 인식 데이터를 사용하였다. 총 10,000문장 중 중복된 문장을 제거한 후, 7,000문장(약 85%), 1,240문장(약 15%)으로 나누어 각각 학습과 평가에 사용하였다.

표 4. 한국어 개체명 인식 실험 결과(F1-score)

	Baseline	제안 모델	성능 차이
PS	83.43	86.04	+2.61
LC	78.99	83.24	+4.25
OG	78.23	80.15	+1.92
DT	94.96	95.28	+0.32
TI	95.85	97.39	+1.54
micro avg	85.42	87.25	+1.83
macro avg	86.29	88.42	+2.13

표 4는 3.3에서 서술한 기본 자질을 학습한 CRF 모델을 baseline으로 하여 기본 자질과 본 논문에서 제안한 추가 자질을 학습한 제안 모델의 성능을 비교한 결과이다. 실험 결과를 보면 기존 모델에 비해 한국어 어휘지도를 활용한 자질을 학습한 모델의 성능이 전체적으로 상승했음을 알 수 있다. 특히 지명을 나타내는 LC 태그의 성능은 F1 기준 약 4.51% 포인트 향상한 것을 확인할 수 있다. 이는 한국어 어휘지도의 상위어 정보를 사용함으로 학습 데이터를 확장하여 OOV 문제를 보완했음을 설명한다.

표 5. 자질 사용에 따른 성능 비교

	F1 score	Training	Processing
Baseline	85.42	109.87초	2.36초
Baseline + 간접적 의존관계	86.50	114.03초	2.38초
Baseline + 상위어	86.70	121.69초	2.61초
제안 모델	87.25	124.48초	2.65초

표 5는 본 논문에서 제안한 모델을 python으로 구현하여 Intel(R) core(TM) i7-5820K CPU, 32GB의 RAM으로 구성된 환경에서 각 자질에 추가에 따른 성능을 측정한 결과이다. 비교 결과, 간접적 의존관계 자질에 비해 상위어 자질의 기여도가 높음을 확인할 수 있었다. 또한, 본 논문에서 제안한 자질들을 추가함에 따라 Baseline에 비해 1.83% 포인트 성능 향상을 보였으며 평가 셋(1,240 문장)의 처리 시간은 0.29초 증가한 것을 확인할 수 있었다.

표 6은 엑소브레인 개체명 인식 데이터를 사용하여 학습 및 평가를 진행한 기존 모델과 본 논문의 제안 모델의 성능을 비교한 결과이다. [10]은 중복된 문장을 제거하지 않은 10,000개의 문장을 모두 사용하여 5배수로 나눈 뒤 교차 검증 실험을 진행한 결과이다. [14]는 사전 학습된 형태소 단위 BERT의 마지막

레이어 출력 값에 CRF를 연결하여 fine-tuning 하여 개체명 인식의 성능을 측정한 결과이다. 비교 결과, 기존 개체명 인식에서 높은 성능을 보였던 딥러닝 방식의 Bi-LSTM-CRF 모델이 아닌 CRF 단일 모델을 사용했음에도 약 0.08% 포인트만이 차이가 나는 것을 확인할 수 있었다. 하지만 기존 개체명 인식 분야에서 가장 높은 성능을 보이는 [14]와는 4.33% 포인트 떨어진 성능을 보이는데 이는 BERT가 위키피디아 코퍼스를 사전 학습했기 때문에 본 논문에서 제안하는 모델에 비하면 OOV가 등장하는 빈도가 낮기 때문이라고 추측한다. [1]에서 제안한 위키피디아 코퍼스를 이용하여 사전을 구축한 후, 이를 자질로 활용한다면 성능 차이를 줄일 수 있을 것이다.

표 6. 기존 모델과의 성능 비교(F1-score)

모델	F1 score
제안 모델	87.25
Stacked Bi-LSTM-CRF [10]	87.33
BERT [14]	91.58

5. 결 론

본 논문에서는 의미역, 의존관계 분석에 한국어 어휘지도를 이용한 자질을 추가함으로 성능 향상을 보인 기존 연구들을 바탕으로 이를 한국어 개체명 인식에 적용하고 평가하였다. 연구 결과, 제안 자질을 추가로 학습한 모델이 기존 모델보다 전체적으로 성능이 향상됨을 확인할 수 있었다. 또한, 이를 기존 딥러닝을 사용한 모델과 비교한 결과 단일 CRF 모델만을 사용했음에도 그에 준하는 성능을 낼 수 있음을 확인하였다.

Acknowledgement

이 논문은 한국연구재단의 지역대학우수과학자지원사업(NRF-2020R1I1A3070938)과 학제간융합연구지원사업(NRF-2019SA5B6102698)의 지원을 받아 수행된 연구임

참고문헌

[1] 민진우, 나승훈, “문자 기반 LSTM-CRF 한국어 개체명 인식을 위한 사전 자질 활용,” 제 28회 한글 및 한국어 정보처리 학술대회 논문집, p.119-121, 2016

[2] 김완수, 옥철영, “격틀 사전과 하위 범주 정보를 이용한 한국어 의미역 결정,” 정보과학회논문지 43권, p.1376-1384, 2016

[3] 정충선, 신준철, 이주상, 옥철영, “의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템,” 정보과학회논문지 46권, p.1174-1185, 2019

[4] 배영준, 옥철영, “한국어 어휘지도(UWordMap)와 API 소개,” 제 26회 한글 및 한국어 정보처리 학술대회 논문집, p.27-31, 2014

[5] 이창기, 장명길, “Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식,” 인지과학회논문지 21권, p.655-667, 2010

[6] 이태석, 전홍우, 강승식, “CRF를 이용한 특허 개체명 인식,” 한국정보과학회 학술발표논문집, p.612-613, 2014

[7] 유홍연, 고영중, “Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장,” 정보과학회논문지 44권, p.306-313, 2017

[8] 송치윤, 양성민, 강상우, “개선된 워드 임베딩 모델과 사전을 이용한 Bidirectional LSTM CRF 기반의 한국어 개체명 인식기,” 한국정보과학회 학술발표논문집, p.699-701, 2017

[9] 박동주, 안창욱, “개체명 비율 사전을 결합한 Bidirectional LSTM-CRF 기반 개체명 인식,” 한국정보과학회 학술발표논문집, p.721-723, 2019

[10] 장윤정, 민태홍, 이재성, “Stacked Bi-LSTM-CRF 앙상블 모델을 이용한 개체명 인식,” 한국정보과학회 학술발표논문집, p.2049-2051, 2018

[11] 신준철, 옥철영, “기분석 부분 어절 사전을 활용한 한국어 형태소 분석기,” 정보과학회논문지 39권, p.415-424, 2012

[12] Sasano, R., and Kurohashi, S. “Japanese named entity recognition using structural natural language processing,” Proceedings of IJCNLP, p.607-612, 2008

[13] Xiao Ling and Daniel S. Weld, “Fine-Grained Entity Recognition,” Proceedings of AAAI, p.94-100, 2012

[14] 박광현, 나승훈, 신종훈, 김영길, “BERT를 이용한 자연어처리 : 개체명 인식, 감성분석, 의존 파싱, 의미역 결정,” 한국정보과학회 학술발표논문집, p.584-586, 2019