공학석사 학위논문

한국어 어휘 의미망을 활용한 CRF 모델 기반 개체명 인식

CRF based Named Entity Recognition Using A Korean Lexical Semantic Network

울산대학교 대학원

전기전자컴퓨터공학과

박 서 연

한국어 어휘 의미망을 활용한 CRF 모델 기반 개체명 인식

CRF based Named Entity Recognition Using A Korean Lexical Semantic Network

지도교수 옥철영

이 논문을 공학석사 학위논문으로 제출함

2021년 1월

울산대학교 대학원

전기전자컴퓨터공학과

박 서 연

박서연의 공학석사 학위논문을 인증함

심사위원 옥 철 영 (인)

심사위원 권 영 근 (인)

심사위원 김 재 훈 (인)

울산대학교 대학원

2021 년 1 월

[국문 요약]

한국어 어휘 의미망을 활용한 CRF 모델 기반 한국어 개체명 인식

개체명 인식은 주어진 문장 내에서 OOV(Out Of Vocabulary)로 등장하는 고유한 의미가 있는 단어들을 미리 정의된 개체의 범주로 분류하는 작업이다. OOV 문제를 해결하기 위해 단어를 구성하고 있는 문자들을 딥러닝을 활용하여 해당 단어의 임베딩을 CNN, LSTM 네트워크를 통해 합성하는 방식이나 BERT나 ELECTRA와 같은 언어 모델을 학습하여 임베딩한 연구가 진행되었다. 하지만, 이러한 딥러닝 네트워크 혹은 언어 모델을 이용한 임베딩 방식을 사용한 모델은 고성능의 컴퓨팅 파워가 요구되며 학습 모델의 속도가 느려 실용성이 낮다는 문제가 있다. 본 논문에서는 실용성을 목적으로 처리 속도와 정확률을 모두 고려하여 빠른 속도로 학습 및 처리를 할 수 있는 기계학습 방식의 CRF를 기반으로 하여 의미 자질과 구조적 자질을 추가하여 OOV 문제를 보완한 개체명 인식 시스템을 제안한다.

본 논문에서는 한국어 어휘 의미망(UWordMap)을 활용하여 사람의 지식을 기반으로 한 의미 자질을 한국어 개체명 인식에 적용하였다. 대상의 단어보다 큰 범주의 의미를 가지는 상위어를 자질로 사용하므로 학습 데이터를 확장하는 역할을 하여, 개체명 인식 분야에서 가장 큰 문제인 OOV 문제를 보완한다. 또한, 대부분의 기계학습 기반의 개체명 인식 모델에서는 현재 토큰의 제한된 주변 토큰에 대한 정보만을 학습하여 주요 키워드가 멀리 떨어져 있을 경우, 그 정보를 학습하지 못한다는 문제가 있다. 이 점을 보완하기 위해 구조적 자질인 의존관계와 격조사 정보를 학습한다.

국립국어원 모두의말뭉치(개체명 인식 말뭉치)를 사용하여 학습 및 평가한 결과, 한국어 어

휘 의미망을 활용한 의미 자질과 의존관계 정보를 활용한 구조적 자질을 학습한 제안모델은 F1 score 기준 91.05% 포인트의 성능과 초당 약 1,466 문장을 처리하였다. 같은 실험 환경에서 개체명 인식 분야에서 보편적으로 많이 사용된 딥 러닝 모델인 stacked Bi-LSTM-CRF과의 성능 비교 결과, 정확률과 처리 속도, 학습 속도에서 모두 향상된 결과를 보였다. 이를 통해 기계학습 방식의 CRF 모델만을 이용하여 높은 성능과 빠른 학습 및 처리 속도를 모두 고려한 실용성을 높인 개체명 인식 시스템을 구축할 수 있음을 보였다.

차례

그림 차례

표 차례

1 서론

1.1 연구 배경

인공지능은 4차 산업혁명의 혁신적인 변화를 일으킬 미래의 유망기술로 부각되고 있다. 이에 인공지능과 관련된 대화 시스템(Dialog System), 검색 엔진(Search Engine), 질의응답(Question Answering) 시장에 구글, 페이스북, 마이크로소프트 등 여러 글로벌 기업들도 플랫폼 선점에 막대한 투자로 총력을 기울이고 있다. 이러한 인공지능 기반의 자연어처리 분야는 문장의 의도를 이해하기 위해 단어가 가지고 있는 의미를 파악해야 한다. 이러한 단어의 의미를 파악하기 위한 과정에서 필요한 것 중의 하나가 개체명 인식이다. 개체명 인식(Named Entity Recognition)은 주어진 문장 내에서 인명이나 장소, 기관명 등과 같이 고유한 의미가 있는 단어들을 미리 정의된 개체의 범주로 분류하는 작업이다. 정보 검색 엔진에서 개체명은 주요 검색 대상이 되며, 질의응답에서는 주요 질의와 응답 대상이 되는 만큼 자연어처리 분야의 성능 향상을 위한 핵심적인 기반이 되기 때문에 현재까지 국내외에서 다양한 방법으로 연구가 되어왔다.

개체명 인식은 입력된 문장에서, 문장의 단어들을 각각의 개체명 태그로 예측하는 순차적 레이블링(Sequential Labeling) 문제이다. 일반적으로 개체명 태그가 부착된 말뭉치의 문장 성분을 조합하고 효과적인 자질을 찾아내어 학습하는 방식을 이용한다. 이때 사용하는 학습 모델은 순차적 레이블링에 효과적인 Bidircetional LSTM-Conditional Random Fields(BiLSTM-CRF)가 가장 많이 사용되었으며 우수한 성능을 보인다. RNN(Recurrent Neural Network)의 단점을 보완한 LSTM(Long Short-Term Memory) 기반 모델인 BiLSTM-CRF 모델은 입력 문자열을 BiLSTM을 이용하여 양방향으로 은닉 벡터를 얻고 출력 태그 간의 의존성을 CRF로 모델링하여 개체명 태그를 예측한다.

개체명 인식 분야에서는 새롭게 생성되고 사라지는 개체명의 특성 때문에 문장 내에서 OOV(Out Of Vocabulary)의 형태로 등장하는 문제가 발생한다. 이를 해결하기 위해 단어를 구성하고 있는 문자들을 딥 러닝을 활용하여 해당 단어의 임베딩을 CNN, LSTM 네트워크를

통해 합성하는 방식들이 제안되었다. 최근에는 대용량의 말뭉치를 이용해 트랜스포머(Transformer)와 셀프 어텐션 메커니즘(self-attention mechanism)으로 구성된 BERT(Bidirectional Representation from Transformers)나 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)와 같은 언어 모델을 학습하여 임베딩에 활용한 연구가 진행되어 높은 성능을 보이고 있다. 한편, 외부 리소스를 활용하여 OOV 문제를 해결하는 방안도 연구되어, BiLSTM, BERT, ELECTRA와 CRF를 결합한 모델에 개체명 사전과 같은 자질 정보를 추가하는 방식을 많이 사용하고 있다.

하지만 BiLSTM, BERT, ELECTRA와 CRF를 결합한 딥 러닝 모델은 고성능의 컴퓨팅 파워가 요구되며 학습모델의 속도가 느려 실용성이 낮다는 문제가 있다. 최근 연구는 정확률을 중점으로 연구되고 있으나 실용화를 위해서는 처리 속도도 함께 고려되어야 한다. 따라서 본 논문에서는 정확률과 처리 속도를 모두 고려하여 딥 러닝 모델보다 빠른 속도로 학습 및 처리를 할 수 있는 기계학습 방식을 사용한 개체명 인식 모델을 제안한다. 일반적으로 기계학습 방식을 사용한 개체명 인식 방법의 정확률이 딥 러닝을 사용한 방법에 비해 낮다는 점을 보완하기 위하여 사람의 지적 정보를 학습 자질로 활용하였다. 딥 러닝이 사람이 추론하고 생각하는 것에 착안하여 연구된 기계학습 알고리즘이라는 점에 따라 기계학습 알고리즘에 사람의 지적 정보를 학습 자질로 하여 간접적으로 딥 러닝 모델의 역할을 할 수 있도록 하였다. 본 논문에서는 기존 의미역[1], 의존관계[2] 분야에 한국어 어휘 의미망(UWordMap)[3]을 이용한 자질을 추가함에 따라 성능 향상을 보였던 연구들을 바탕으로 한국어 어휘 의미망을 사람의 지적 정보로써 활용하여 단어의 상위어, 고유명사, 의존관계 정보를 학습 자질로 추가한 CRF 기반의 한국어 개체명 인식 모델을 제안한다. 현재 단어에 대한 상위어와 최상위어를 자질을 학습함으로 학습 데이터에 등장하지 않았던 단어를 개체명으로 인식하지 못 하는 OOV 문제를 보완하고 의존관계 자질을 학습함으로 문장의 구조를 이해하여 문맥 정보를 반영한다. 상위어 자질을 의미 자질로, 의존관계 자질을 구조적 자질로 보고 각각의 자질이 개체명 인식의 성능에 어떠한 영향을 주는지 분석한다.

1.2 관련 연구

1.2.1 한국어 개체명 인식

한국어 개체명 인식 문제는 사전기반 및 규칙기반 방식을 사용하는 전통적인 방법과 개체명 태그를 부착한 말뭉치를 학습하여 분석하는 학습기반 방식으로 나뉜다. 규칙기반과 사전기반의 개체명 인식 방법은 정의된 규칙과 사전에 대해서는 정확한 분석이 가능하고 분석 속도가 빠르다는 장점이 있다. 하지만 개체명은 고유명사이거나 학습 데이터에 등장하지 않았던 OOV인 경우가 많으며, 항상 새롭게 만들어지고, 또한 같은 단어라도 사용되는 문맥에 따라 상이한 의미 변화를 보이기 때문에 단순 사전 구축으로는 개체명을 인식하는 것이 어렵다.

학습기반 방식은 개체명 태그가 부착된 말뭉치를 사용하여 문장의 성분들을 조합하고 선택하여 효과적인 자질을 찾아내어 사용한다. 이를 바탕으로 순차적 레이블링(Sequential Labeling)에 효과적인 기계학습 모델에 적용한 개체명 인식 연구가 진행되었다[4, 5]. 이창기, 장명길(2010)의 연구에서는 기계학습 모델인 Structural SVMs 및 Pegasos 알고리즘을 이용한 개체명 인식 모델을 제안하였다[4]. 이태석, 전홍우, 강승식(2014)는 또 다른 기계학습 모델인 CRF를 이용하여 특허 문서에 대한 개체명 인식 태그를 학습 및 평가한 연구를 진행하였다[5].

하지만 최근에는 기계학습 방식은 학습을 위해 다양한 자질을 선별하는 과정에서 비용이 많이 든다는 단점이 있어 순차적 레이블링에 특화된 RNN(Recurrent Neural Network) 계열의 LSTM(Long Short-Term Memory) 기반의 딥 러닝 네트워크를 사용함으로 자질 선정을 위한 비용을 줄인 연구들이 진행되었다[6, 7]. 유홍연, 고영중(2017)은 LSTM 기반의 딥 러닝 모델은 입력이 되는 단어 표상에 의존적이기 때문에 입력이 되는 단어를 잘 표현하기 위하여 단어 표상을 확장하는 방법을 제안하였다[6]. 민진우, 나승훈(2016)은 개체명 인식에서의 OOV 문제를 해결하기 위해 단어 단위의 임베딩 뿐만 아니라 단어를 구성하는 문자로부터 단어 임베딩을 합성해내는 방식을 제안했으며, 위키피디아 사전과 네이버 지식백과, 세종 고유명사 등의 외부 리소스로부터 개체명 사전을 구축하고 최적의 매칭 조합을 찾아 자질로 사

용한 연구를 진행하였다[7].

한편, 대용량의 말뭉치를 이용해 학습한 언어 모델이 다양한 자연어 처리 분야에서 좋은 성능을 보이고 있다. 이를 바탕으로 한국어 개체명 인식에서도 언어 모델을 대용량으로 학습시킨 후 활용한 연구가 진행되었다[8, 9, 10]. 박관형 외(2019)는 트랜스포머와 셀프 어텐션 메커니즘을 이용하여 문장에서 임의로 단어를 마스킹(masking)하고 예측하도록 학습한 BERT를 활용하여 개체명 인식을 수행하였다[8]. 민진우 외(2019)는 BERT와 달리 임의의 단어가 매 학습마다 동적으로 마스킹 되도록 개선한 RoBERTa를 활용한 개체명 인식 모델을 제안하였다[9]. 김홍진 외(2020)은 Generator에서 임의의 단어를 마스킹 하고 예측하도록 학습한 후, Discriminator에서 생성한 단어 열에 대해서 각 단어가 원래 입력과 동일한 것인지 치환된 것인지 예측하도록 학습한 ELECTRA와 LAN(Label Attention Network)를 결합한 음절 단위 개체명 인식을 수행하였다[10].

딥 러닝 모델과 언어 모델을 사용한 개체명 인식은 자질 추출을 위한 노력이 줄어들며 기존 기계학습을 기반으로 한 모델에 비해 높은 성능을 보이고 있어 많이 연구되고 있지만, 고성능의 컴퓨팅 파워가 요구되며 모델의 학습 및 처리 속도가 느려 실용성이 낮은 단점이 있다. 이에 본 논문에서는 실용성을 중점으로 하여 자질 추출을 위한 비용이 들지만 빠른 속도로 학습 및 처리를 할 수 있는 기계학습 방식의 CRF를 기반으로 하여 지적 정보와 구조적 정보를 학습 자질로 사용한 개체명 인식 모델을 제안한다.

1.2.2 CRF (Conditional Random Fields)

CRF (Conditional Random Fields)는 주어진 입력 데이터 열에 대하여 레이블 열의 확률을 이용하는 조건부 모델이다. 일반적인 분류(classification)는 하나의 입력 벡터 x에 대하여 하나의 레이블 값인 y를 예측하는 문제이다. CRF는 입력 X를 길이가 n인 시퀀스이고, $X = x_1, \ldots, x_n$라 할 때, 같은 길이의 레이블 열 $Y = y_1, \ldots, y_n$로 분류할 때 각각의 레이블을 하나씩 예측을 하는 것이 아닌 전체적인 문맥을 고려하여 가능성이 있는 시퀀스 Y 후보를 여러 개 선택한 뒤, 가장 적합한 하나의 레이블 열을 선택하여 Y로 결정함으로 입력 데이터 열 사이의 의존 관계를 반영할 수 있다는 장점이 있다. 일반적으로 CRF 모델은 조건부 확률을 최대화하기 위해 훈련된 비방향성 그래프 모델이나 입력 데이터 열에 레이블 열을 부여하는 문제에는 선형 체인 구조의 CRF 모델이 적합하다.

$X = x_1, \ldots, x_n$을 입력 열에 대한 확률 변수(random variable)라고 하고, $Y = y_1, \ldots, y_n$을 입력 열에 대응하는 레이블 열의 확률 변수라고 한다면, 매개변수 $\Lambda = (\lambda_1, \lambda_2, \ldots, \mu_1, \mu_2, \ldots)$를 갖는 선형 체인 구조의 CRF는 식 (1)과 같은 조건부 확률로 정의된다.

$$P_\Lambda(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_j \sum_{i=1}^{n} \lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_k \sum_{i=1}^{n} \mu_j s_k(y_i, X, i)\right) \quad \textbf{(1)}$$

여기서 $Z(X)$는 입력 데이터 열에 대한 레이블 열의 확률 합이 1이 되도록 하는 정규화 상수이다. $t_j(y_{i-1}, y_i, X, i)$는 전이 자질 함수(transition feature function)이며, $s_k(y_i, X, i)$는 상태 자질 함수(state feature function)이다.

학습 데이터로부터 매개변수 Λ를 구하고 나면, 주어진 입력 데이터 열 X에 대하여 가장 가능성이 높은 레이블 열 $\hat{Y}$은 식 (2)와 같이 구할 수 있다.

$$\hat{Y} = argmax\, P_\Lambda(Y|X) \quad \textbf{(2)}$$

가능성이 가장 높은 레이블 열 $\hat{Y}$은 각 레이블 간의 의존성을 포함하고 있어 개체명 인식에

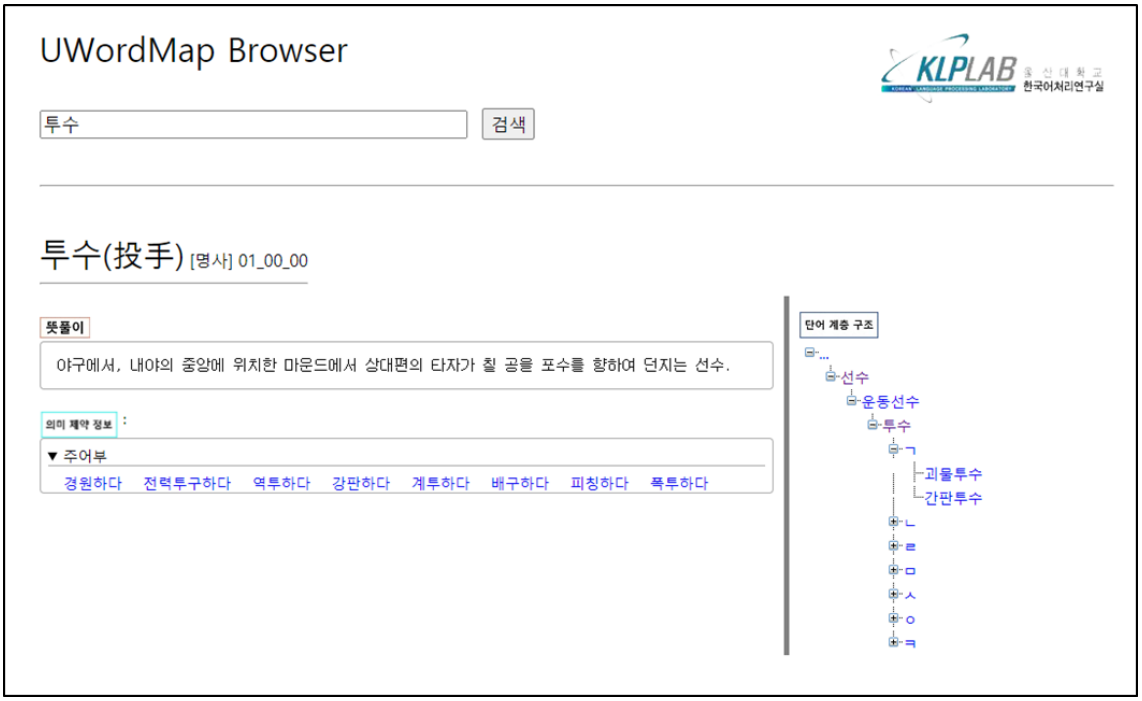
사용하는 BIO 태그의 제약사항을 학습할 수 있다.

입력 데이터 열과 레이블 열 사이의 결합 확률(joint probability)을 이용하는 생성 모델(generative model)인 HMM(Hidden Markov Model)과 같이 입력 데이터 열의 각각에 대해 레이블을 부여하는 모델은 모든 가능한 입력 데이터 열을 나열해야 하며, 상호 작용하는 자질을 표현하기 힘든 단점이 있어 이를 보완한 방법인 CRF는 자연어 처리의 여러 분야에 적용되어 왔다.

2 개체명 인식을 위한 의미 자질

2.1 한국어 어휘 의미망(UWordMap)

본 논문에서는 개체명 인식의 성능 개선을 위하여 한국어 어휘 의미망(UWordMap)[1]을 사용한다. UWordMap은 표준국어대사전을 기반으로 명사, 용언, 부사 등의 어휘들이 의미제약으로 상호 연결된 다의어 수준의 어휘 의미망이다. UWordMap에는 명사의 계층 구조를 담고 있어 명사에 대한 상위어와 하위어 정보를 얻을 수 있다.

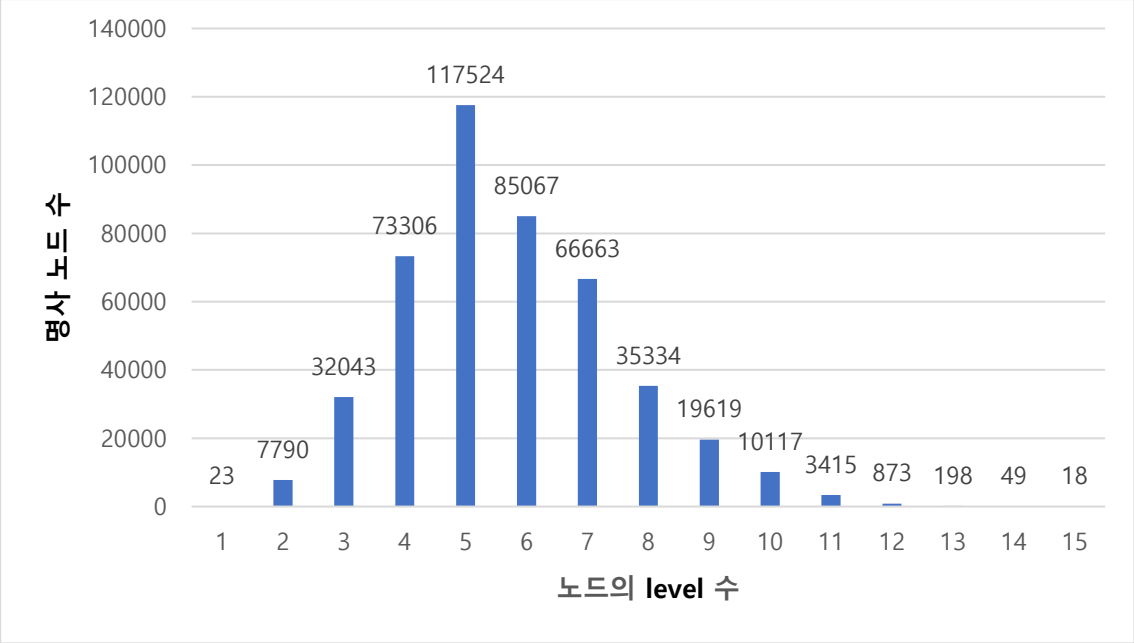


[그림 1] UWordMap에서 '투수(pitcher)' 검색 결과

[그림 1]은 실제 UWordMap에서 '투수'를 검색하여 '투수__01(pitcher)'을 선택한 결과이다. 검색 결과, '투수__01'에 대하여 뜻풀이, 의미제약, 단어 계층 구조 등의 정보를 얻을 수 있다. UWordMap은 최상위 계층 단어 root인 UWIN 노드가 있으며, 최상위어로 총 23개의

[1] UWordMap 브라우저 : http://nlplab.ulsan.ac.kr:7070

단어를 가진다. [그림 2]는 UWordMap의 각 계층별 단어 분포도를 나타낸 것이다. 단어 계층 구조에서 상위어는 현재 단어보다 의미적으로 큰 범주를 가지며, 하위어는 해당 단어보다 의미적으로 작은 범주를 가진다. 상위어 정보를 학습 자질로 사용하여 현재 단어의 형제어를 하나의 범주로 인식할 수 있다.



[그림 2] UWordMap의 계층별 분포도

2.2 한국어 어휘 의미망의 상위어를 이용한 자질

형태소 단위의 개체명 인식 모델은 개체명의 특성상 사전에 학습되지 않은 OOV로 등장하는 경우가 많기 때문에 재현율이 낮을 수 있다. 본 논문에서는 OOV 문제를 보완하여 재현율을 높이기 위해 상위어 정보를 학습 자질로 사용하였다. 어휘의 상위어 정보는 동형이의어 및 다의어 수준의 어휘 의미망인 울산대학교의 UWordMap을 사용한다. [그림 3]은 실제 UWordMap에 '운동선수(athlete)'를 검색하여 얻은 결과이다. 오른쪽 단어 계층 구조를 통해서 '운동선수'의 1계층 상위어는 '선수(player)'이며, '운동선수'에 대한 1계층 하위어에는 '키퍼(keeper)', '타자(hitter)', '투수(pitcher)' 등이 있음을 알 수 있다. 단어의 계층 구조를 통하여 상위어 정보를 학습 자질로 사용함으로 학습 데이터를 확장하는 역할과 개체명 인식을 위한 키워드 역할을 할 수 있다.

[그림 3] UWordMap에서 '운동선수'를 검색한 결과

2.2.1 학습 데이터의 확장

첫 번째로 상위어 자질은 학습 데이터를 확장하는 역할을 한다. 상위어는 현재 단어보다 큰 의미를 가지는 단어이다. 이를 학습 자질로 사용을 하면 같은 상위어를 가지는 단어들 모두를 하나의 범주로 인식할 수 있게 된다.

〈표 1〉 1계층 상위어가 '수도__09(capital)'인 단어들의 계층별 구조

계층	단어			
1	공간__05(space)			
2	지역__03(area)			
3	도시__03(city)			
4	수도__09(capital)			
5	**베이징**	**런던**	**도쿄**	**하노이**

〈표 1〉은 1계층 상위어가 '수도__09(capital)'인 단어의 계층별 구조를 나타낸 것이다. 예를 들어, 중국의 수도를 의미하는 '베이징'은 현재 5계층에 존재하는 단어이므로 이의 1계층 상위어는 4계층 단어인 '수도__09'이며, 최상위어는 1계층 단어인 '공간__05(space)'이다.

〈표 2〉 상위어 자질 추가 예시

원문	**베이징**은 극심한 대기 오염으로 악명이 높기 때문이다.
상위어 자질	**수도**는 극심한 대기 오염으로 악명이 높기 때문이다.
데이터 확장	**하노이**는 극심한 대기 오염으로 악명이 높기 때문이다.

이를 바탕으로 '베이징'에 대해 1계층 상위 단어인 '수도__09'를 학습 자질로 사용한다면, 같은 1계층 상위 단어를 가지는 어휘인 '런던(영국의 수도)', '도쿄(일본의 수도)', '하노이(베트남의 수도)' 등을 하나의 범주로 인식할 수 있다. 특히나 학습 데이터에는 등장하지 않았던 '하노이'에 대해서도 '베이징'과 동일한 1계층 상위 단어를 가지기 때문에 지명으로 인식할 수 있다.

2.2.2 개체명 인식을 위한 키워드

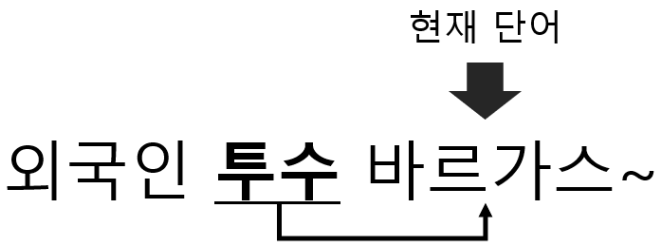
두 번째로 상위어 자질은 개체명 인식을 위한 키워드의 역할을 한다. 학습 데이터에 등장하는 단어 중에는 상위어가 개체의 범주인 '사람', '지역'과 같은 단어로 추상화를 할 수 있는 것들이 존재한다. 주변 단어를 추상화 했을 때, '사람'이나 '지역'과 같은 개체명의 범주를 나타내는 단어가 등장한다면 이를 해당 범주로 인식할 수 있도록 키워드의 역할을 할 수 있다.

〈표 3〉 3계층 단어가 '사람(person)'인 단어들의 계층별 구조

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

〈표 3〉은 3계층 단어가 '사람(person)'인 단어들인 '투수(pitcher)', '골키퍼('goalkeeper)', '국회의원(member of congress)', '여주인공(heroine)'의 계층 구조를 나타낸 것이다. 투수, 골키퍼, 여주인공은 8계층 단어이며 국회의원은 6계층 단어이다. 해당 단어가 존재하는 계층 레벨은 다르지만 모두 3계층 단어가 사람으로 동일하다. 개체명인지 파악하고자 하는 단어의 주변 단어 중, 〈표 3〉과 같이 3계층 단어가 '사람'인 단어가 등장한다면 이는 현재 파악하고자 하는 단어가 사람이라는 개체로 분류될 수 있다는 것을 의미하므로 이를 키워드로 사용할 수 있다

예를 들어, 학습 데이터에 '외국인 **투수** 바르가스~'라는 문장이 등장했으며 현재 '바르가스'가 개체명인지 파악하고자 한다면 주변 단어인 '**투수**'라는 키워드를 통해 인명으로 분류할 수 있다.



[그림 4] 키워드 역할을 하는 상위어의 예시

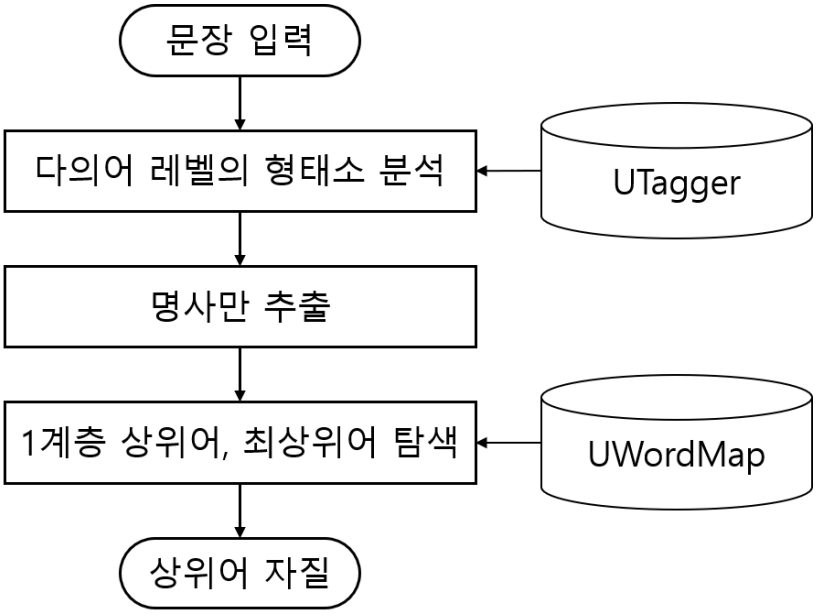
2.2.3 상위어 자질 추가 알고리즘

상위어를 자질로 사용하는 과정에서 표층형이 같더라도 다른 상위어 계층을 가질 수 있다. 해당 어휘가 동형이의어 혹은 다의어일 경우, 표층형이 같더라도 의미상으로 다른 상위어를 가지기 때문이다. 〈표 4〉는 '한국'이라는 명사를 UWordMap에 검색하여 얻은 일부의 동형이의어와 다의어를 나타낸 것이다. 일반적으로 개체명 인식 말뭉치에 등장하는 '한국'은 〈표 4〉의 '한국__0502(Republic of Korea)'일 것이다. 하지만 이를 분별하지 않고 상위어 자질을 추가하면 잘못된 상위어를 학습 자질로 사용하게 된다. 또한, 같은 동형이의어 내에서도 다의어 번호에 따라 다른 상위어를 가질 수 있다. 예시의 '한국__0501(Korean Empire)'과 '한국__0502(Republic of Korea)'은 같은 동형이의어 번호를 가지고 있지만 각각 '제국(empire)'과 '공화국(republic)'으로 다른 상위어 계층 구조 가진다.

〈표 4〉 '한국'의 동형이의어와 다의어

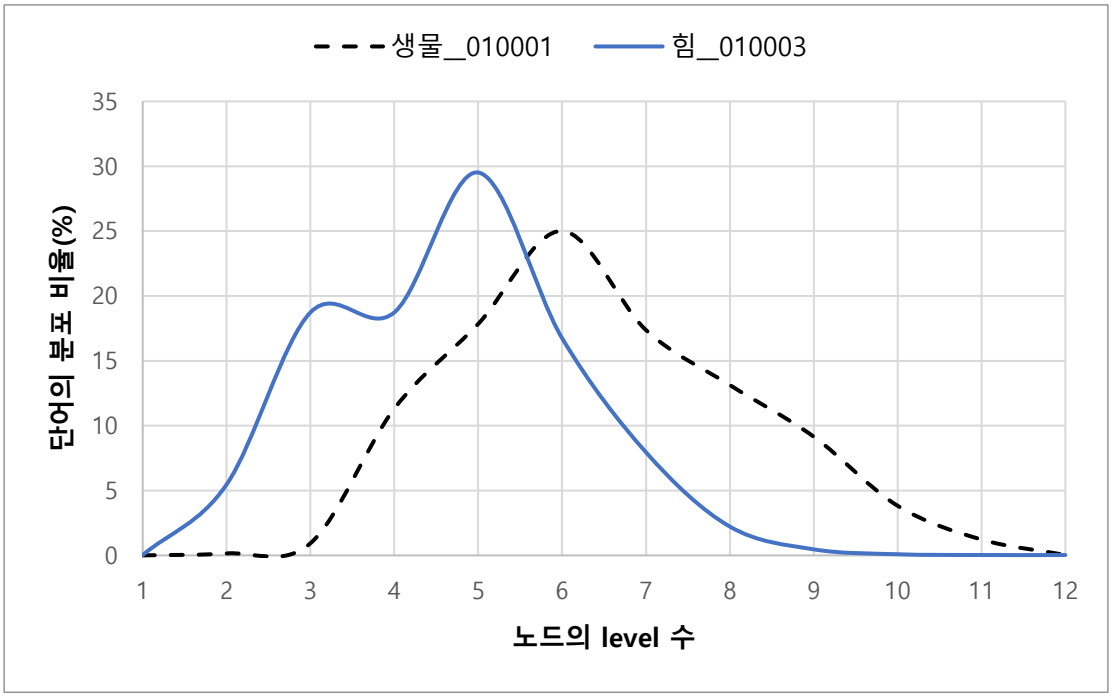
단어	뜻풀이	1계층 상위어
한국_0401	국화과의 재배 식물의 하나. 11월에 노란 꽃이 피고 총포가 길다.	국화
한국_0402	겨울에 피는 국화. 대개 12월에서 다음 해 1월에 걸쳐 노란 꽃이 핀다.	국화
한국_0501	'대한 제국'을 줄여 이르는 말.	제국
한국_0502	대한민국 - 아시아 대륙 동쪽에 있는 한반도와 그 부속 도서로 이루어진 공화국.	공화국

본 논문은 잘못된 상위어를 자질로 사용하는 것을 방지하기 위하여 형태소 분석 및 동형이의어, 다의어 분별 시스템인 UTagger[11]를 이용하여 동형이의어와 다의어 태그가 부착된 명사만을 추출한 후, UWordMap의 명사 계층 구조에서 각 명사에 대한 1계층 상위어와 최상위어를 찾아 자질로 추가하였다.



[그림 5] 상위어를 학습 자질로 추가하는 과정

각 명사에 대한 최상위어 자질을 추가할 때, UWordMap의 단어 계층에서 UWIN(최상위 단어)을 기준으로 바로 하위 단계의 1계층 단어를 자질로 학습한다. 하지만 2계층 단어를 최상위어 자질로 선택하는 경우도 존재한다. 이는 각각의 1계층 단어에 대해 하위 단어들의 분포가 다르기 때문에 1계층 단어를 자질로 사용할 경우, 과도한 추상화를 하여 의미를 충분히 전달하지 못할 수 있기 때문이다. 아래 [그림 6]은 UWIN을 제외한 최상위어가 “생물__0101(life)”인 명사와 “힘__0103(power)”인 명사들의 계층 분포를 나타낸 것이다. UWIN을 제외한 최상위어가 “생물__0101”인 단어는 6계층에 가장 많이 분포하고 있으며 “힘__0103”을 최상위어로 가지는 단어는 5계층에 가장 많이 분포하고 있다.



[그림 6] 최상위어가 '생물(life)'와 '힘(power)'인 단어 계층 분포

$$level_{avg} = \sum_{i=1}^{n} i * ratio_i \quad \textbf{(3)}$$

식 (3)은 각 최상위어에 대한 명사의 평균 level을 구하는 식이다. 수식의 i는 현재 노드의 level 수를 나타내며 $ratio_i$는 현재 노드의 level에 대한 명사 분포 비율을 나타낸다. 최상위어에 대한 각 단어의 분포 비율을 식 (3)에 적용해보면 "생물__0101"의 평균적으로 6.49층, "힘__0103"은 4.69층임을 알 수 있다. 이처럼 각각의 최상위어에 대해 각 명사들의 계층 분포가 다르기 때문에 "생물__0101"과 같이 평균계층이 6층 이상인 경우, 최상위어 자질로 1계층 단어가 아닌 2계층 단어를 선택한다. 이 경우에 학습 데이터 내에서 1계층 단어를 "생물__0101"을 가지는 2계층 단어가 등장한다면, 상위어 자질과 최상위어 자질로 현재 토큰의 표층형을 사용한다. 예외로 현재 명사에 대한 상위 계층에 "사람(3계층 단어, 최상위어_생물)", "지역(2계층 단어, 최상위어_공간)", "조직(2계층 단어, 최상위어_집단)" 등의 개체 범주에 해당하는 단어가 존재한다면 평균 계층과 무관하게 범주를 의미하는 단어를 최상위어 자질로

사용한다.

〈표 5〉 최상위어 별 평균 계층

최상위어	평균 계층	최상위어	평균 계층	최상위어	평균 계층
공간_0502	5.05	방법_0001	5.99	재료_0101	4.51
과정_0300	5.31	범위_0001	4.92	정도_1101	5.46
관계_0501	4.68	생물_0101	6.49	존재_0001	6.99
기호_0100	4.81	성질_0002	3.64	종류_0201	4.22
단위_0201	5.45	시간_0401	5.09	집단_0000	5.93
대상_1101	5.21	요소_0401	5.88	행위_0001	5.99
모양_0201	5.68	인지_0803	6.06	힘_0103	4.69
물건_0001	5.99	작용_0101	4.69		

3 개체명 인식을 위한 구조적 자질

대부분의 기계학습 모델을 사용한 한국어 개체명 인식 연구에서는 지도 학습을 기반으로 하여, 현재 토큰에서 얻은 자질, 선행 토큰 2개, 후속 토큰 2개 등의 제한된 주변 정보만을 활용한 자질을 사용한다. 하지만, 개체명은 같은 단어라도 문맥에 따라 상이한 의미 변화를 가지므로 문맥을 파악하지 않고 제한된 주변 정보만을 학습하면 개체명을 인식하는 것이 불가능한 경우가 존재한다.

〈표 6〉 문맥에 따라 달라지는 개체명의 예시

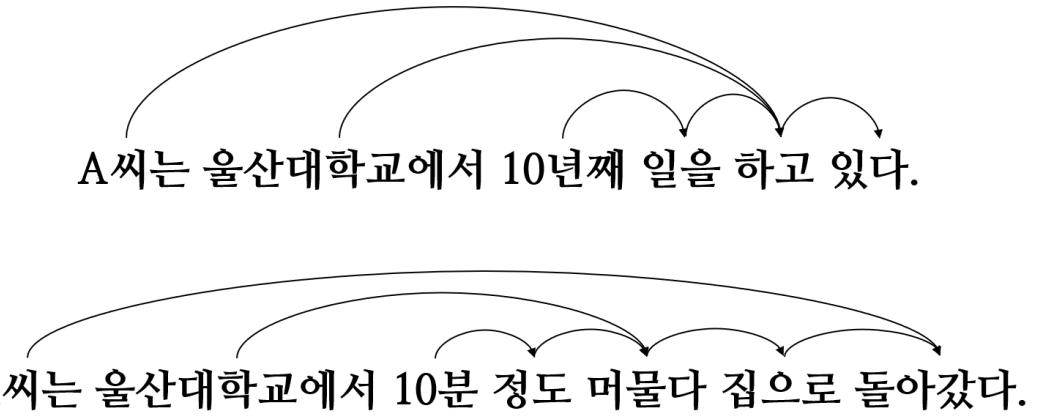
(1) A씨는 〈울산대학교:OG〉에서 10년째 일을 하고 있다. (2) A씨는 〈울산대학교:LC〉에서 10분 정도 머물다 집으로 돌아갔다.

예를 들어, 〈표 6〉의 (1)에 나오는 “울산대학교”는 A 씨가 근무하는 기관을 의미한다. 반면에 (2)의 “울산대학교”는 특정 위치를 의미하고 있기 때문에 지명으로 분류가 된다. 같은 단어임에도 문맥에 따라 분류가 되어야 하는 범주가 달라지고 있지만, 현재 파악하고자 하는 단어인 “울산대학교”의 (1)과 (2)에서 모두 선행 토큰 2개는 “A/SL”, “씨/NNB”이며 후속 토큰 2개는 “에서/JKB”와 “10/SN”이기 때문에 제한된 주변 토큰만을 이용하여서는 이를 정확한 개체명 범주로 분류하는 것이 어렵다.

정확한 개체명 범주로 분류하기 위해서는 문맥을 파악하여 해당 정보를 사용하는 것은 필수적이다. 이를 보완하기 위해 국외에서는 의존관계 자질을 추가하여 개체명 인식 분야에 성능향상을 보인 연구들이 존재한다[12, 13]. 본 논문에서는 이를 한국어 개체명 인식에 적용하기 위해 문장의 구조를 파악하여 의존관계 자질과 격조사 자질을 추가하였다.

3.1 의존관계 자질

문장 성분의 관계를 결정하는 작업은 언어 처리 단계 중 구문 분석 단계에 해당하며 의존관계 결정을 통해 이루어진다. 본 논문에서는 문맥을 파악하여 해당 정보를 한국어 개체명 인식 분야에 적용하기 위하여 의존관계 자질을 추가하였다. 의존관계 자질은 문장의 성분들 간의 의존(dependent) – 지배(governor) 관계로 문장을 표현하여 현재 형태소가 포함된 어절을 의존소로 보고, 해당 어절에 대한 지배소를 찾아 키워드로 활용한 것이다.

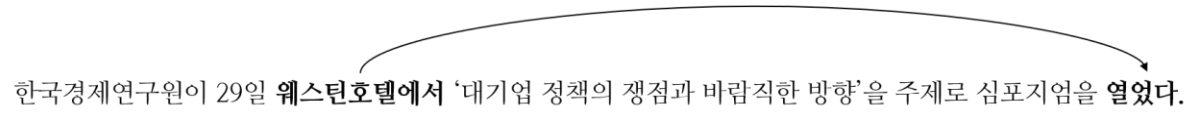


[그림 7] 의존관계가 표시된 한국어 문장

[그림 7]은 〈표 4〉의 (1)과 (2)에 대해 각각의 의존관계를 표시한 것이다. 기존 제한된 주변 정보만을 자질로 활용할 경우, 두 문장의 선행 토큰 2개와 후속 토큰 2개가 모두 동일했지만 의존관계를 통해 (1) 문장의 “울산대학교”는 “하다/VV”를 수식하고 있고 (2) 문장의 “울산대학교”는 “머물다/VV”를 수식하고 있음을 알 수 있다. 이를 자질로 활용할 경우, 주변 정보가 아닌 제한된 범위 밖의 멀리 떨어진 키워드를 찾을 수 있다.

[그림 8]은 의존관계 자질을 이용함으로 키워드를 찾을 수 있는 또 다른 예시이다. 예시의 “웨스틴호텔”은 특정 위치를 나타내는 지명이다. “웨스틴호텔”이 지명임을 인식할 수 있는 가장 중요한 키워드는 격조사인 “에서”와 동사인 “열다”이다. 현재 예문의 “웨스틴호텔”과 키워드인 “열다/VV”는 형태소 단위를 기준으로 17 번째 후속 토큰이다. 이 때문에 단순히 제한

된 주변 정보만을 반영한다면 “열다/VV”라는 중요 키워드를 찾을 수 없게 된다. 하지만 의존관계 정보를 반영한다면, [그림 8]처럼 “웨스틴호텔”의 지배소 어절에 “열다/VV”가 있기 때문에 해당 키워드를 찾아내어 이를 지명으로 인식할 수 있다.



[그림 8] 의존관계가 표시된 한국어 문장

의존관계 정보를 자질로 활용하기 위하여 의존관계가 모두 분석이 되어 있는 국립국어원의 모두의 말뭉치(구문 분석 말뭉치)를 사용하였다. 국립국어원의 모두의 말뭉치에는 각 문장을 구분하기 위한 문서번호인 id가 존재한다. 이를 이용하여 개체명 말뭉치에 존재하는 문장에 대해 의존관계가 태깅 되어 있는 문장을 구문 분석 말뭉치로부터 찾을 수 있었다. 이 때, 개체명 말뭉치에 존재하는 모든 문장이 구문 분석 말뭉치에 존재하진 않는다. 아래 〈표 7〉은 국립국어원의 구문 분석 말뭉치를 UTagger를 통해 형태소 분석을 하여 가공한 예시이다. dependency에는 원문에 대한 의존관계를 나타낸 것이다. 각 줄의 첫 번째 숫자는 현재의 어절이 문장 내에서 몇 번째 어절인가를 나타내며 두 번째 숫자는 현재의 어절이 문장 내의 몇 번째 어절을 수식하는가를 나타낸다. 예시의 “[횡설수설/권순활]北” 어절은 첫 번째 어절이며 세 번째 어절인 “뜯어먹기”를 수식함을 의미한다.

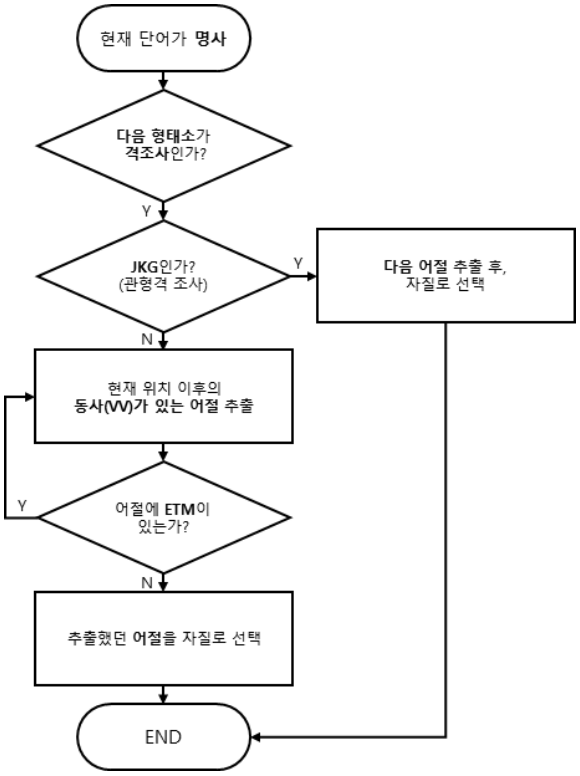
〈표 7〉 구문 분석 말뭉치의 예시

text	[횡설수설/권순활]北 ‘외화벌이’ 뜯어먹기
id	NWRW1800000029.315.1.1
dependency	1 3 [/SS+횡설수설/NNG+//SP+권순활/NNP+]/SS+北/SH 2 3 ‘/SS+외화벌이NNG+’/SS 3 0 뜯어먹/VV+기/ETN

3.2 간접적 의존관계 자질

의존관계 자질은 입력된 문장의 의존관계가 분석되어 있지 않다면 해당 자질을 추가할 수 없다는 제약 사항을 가지고 있다. 이러한 제약사항을 보완하기 위하여 간접적으로 의존관계를 분석하여 사용하는 간접적 의존관계 자질을 추가하였다. 본 논문에서 사용하는 간접적 의존관계 자질이란 명사 다음 격조사가 올 경우, 그에 대한 지배소를 찾아 자질로 사용하는 것을 의미한다. 이는 각 개체에 따라 결합되는 격조사와 지배소가 해당 개체를 정확한 범주로 인식하기 위한 중요 키워드인 경우가 많기 때문에 의존관계가 분석되어 있지 않은 문장에 대해서도 이 키워드를 찾아낼 수 있도록 한다.

간접적 의존관계 자질을 추가하는 과정은 다음 [그림 9]와 같다. 간접적 의존관계 자질의 추가 조건은 현재 단어가 명사이고, 그 명사 다음의 형태소가 격조사인 경우이다. 이 조건을 만족한다면 해당 격조사가 관형격 조사(JKG)일 때와 아닐 때로 나누어 추출하는 자질의 종류가 달라진다.



[그림 9] 간접적 의존관계 자질을 추가하는 과정

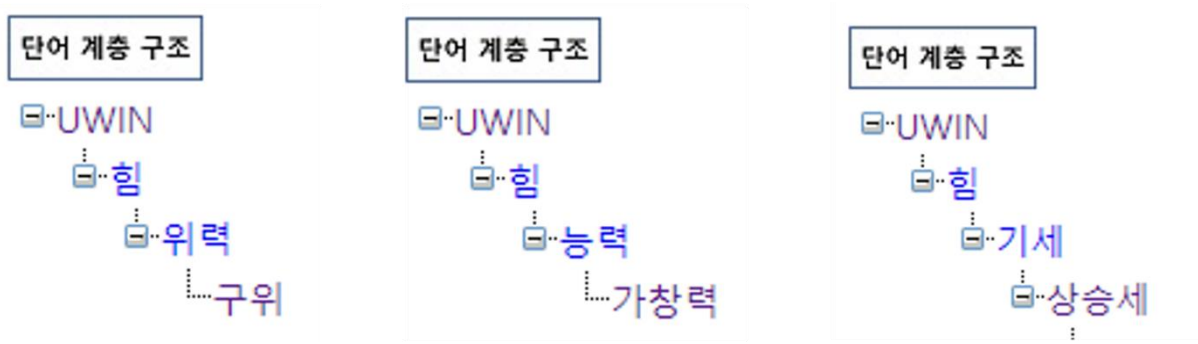
3.2.1 관형격 조사(JKG)

명사 다음의 형태소가 관형격 조사(JKG)일 때는 현재 명사와 관형격 조사가 결합된 어절 바로 다음 어절을 지배소 자질로 선택한다. 선택된 어절 내에 명사가 포함되었다면 상위어 자질을 추가한다. 〈표 8〉은 명사와 관형격 조사가 결합된 어절에 대해 간접적 의존관계 자질을 추출한 예시이다. 예문 '정재훈의 구위에 문제가 있었던 것은 아니다.'에는 고유명사 '정재훈'과 관형격 조사 '의'가 결합된 어절이 존재한다. [그림 9]의 간접적인 의존관계 자질을 추가하는 과정에 따라 바로 다음 어절인 '구위__04/NNG' + '에/JKB'가 지배소 자질로 선택된다. 선택된 지배소 자질 내에는 명사인 '구위__04(야구에서 투수가 던지는 공의 위력)'가 포함되어 있으므로 앞선 상위어 자질 추가 방식을 통해 '위력__0101(상대를 압도할 만큼 강력한 힘)'이라는 1계층 상위어와 '힘__0103(어떤 일을 할 수 있는 능력이나 역량)'이라는 1계층 단어를 각각 상위어 자질과 최상위어 자질로 사용한다.

〈표 8〉 명사+관형격 조사에 대한 격조사 자질 추출 예시

〈정재훈:PS〉의 구위에 문제가 있었던 것은 아니다. 이날 **〈박찬호:PS〉의 구위에** 높은 점수를 주었다. 그러나 팬텀은 **〈아이비:PS〉의 가창력을** 발견했고 **〈김종국:PS〉의 상승세에** 밀려 2위로 내려왔다.

[그림 10]은 〈표 8〉의 예문에서 추출한 지배소 자질 내에 존재하는 명사인 '구위_04'와 '가창력(노래를 부르는 능력)', '상승세(위로 올라가는 기세)'의 계층 구조를 나타낸 것이다. '구위', '가창력', '상승세'는 모두 1계층 단어(최상위어)가 '힘__0103(power)'이기 때문에 각각의 문장을 모두 '〈사람 개체〉의 힘'이라고 추상화를 할 수 있다.



[그림 10] '구위_04', '가창력', '상승세'의 계층 구조

3.2.2 관형격 조사 이외의 격조사

명사 다음의 형태소가 관형격 조사(JKG)가 아닌 다른 격조사인 경우에는, 현재 위치 이후에 나오는 동사(VV)가 포함된 어절 중 가장 가까운 어절을 추출한다. 이 때, 추출한 어절 내에 관형형 전성 어미(ETM)가 포함된 경우는 제외한다. 선택된 어절 내에 관형형 전성 어미가 포함되지 않은 경우, 이를 지배소로 선택한다. 또한 관형격 조사가 아닌 다른 격조사가 등장한 경우에는 앞선 3.2.1의 관형격 조사인 경우와 다르게 명사 다음의 격조사를 자질로 활용한다. 이는 개체 범주에 따라 결합되는 격조사의 종류와 지배소가 달라지므로 정확률을 높이는 역할을 한다.

〈표 9〉는 명사 다음 관형격 조사 이외의 격조사가 등장하여 간접적 의존관계 자질을 추출한 예시이다. 예문 '곧바로 귀국해 9일 제주도로 향한다.'에는 고유명사 '제주도'와 부사격 조사 '로'가 결합된 '제주도로'라는 어절이 존재하므로 간접적 의존관계 자질을 추가하기 위한 시작 조건을 만족한다. 자질 추가 과정을 통해 '제주도로'라는 어절을 기준으로 그 이후에 나오는 동사인 '향하다'가 지배소 자질로 선택된다. 추가로 부사격 조사 '로'를 격조사 자질로 선택한다. 예문처럼 부사격 조사 '로'가 고유명사와 결합된 경우에는 해당 고유명사가 개체명일 확률이 높으며 사람 개체일 가능성보다는 지역 혹은 조직 개체일 확률이 높아진다.

〈표 9〉 명사+격조사(JKG 이외)에 대한 격조사 자질 추출 예시

곧바로 귀국해 9일 **〈제주도:LC〉로 향한다.** 치료를 위해 **〈독일:LC〉로 향했던** 김남일이 성공적으로 수술을 마쳤다. 12일 **〈삼성전자:OG〉에 따르면** 갤럭시 S2는 독일 출시 2주만에 **〈부산도시공사:OG〉에 따르면** 일본 노무라종합연구소에서 새로운 마스터플랜을

〈표 10〉은 개체에 따라 간접적 의존관계 자질을 추출한 예시이다. 추출된 간접적 의존관계 자질이 '을/를 앞세우다'인 경우에는 현재 단어가 '하승진', '홍영현', '피어스'와 같은 사람 개체일 가능성이 높다. 조직 개체에 대해 추출된 간접적 의존관계 자질은 '와의 경기'이다. 추출된 격조사 자질에는 부사격 조사 '와'와 관형격 조사 '의'가 모두 등장하였다. 이 경우에는 명사 바로 다음 격조사인 부사격 조사를 기준으로 하지 않고, 예외적으로 해당 어절 내의 마지막 격조사인 '의'를 기준으로 하여 다음 어절을 지배소 자질로 추출한다.

〈표 10〉 개체에 따른 격조사 자질 예시

개체	추출된 격조사 자질	개체명 예시
PS	을/를 앞세우다	하승진, 홍영현, 피어스
LC	로/으로 떠나다	독일, 헬싱키, 캘리포니아
OG	와의 경기	레드삭스, 롯데, 현대
TI	에 마감되다	0시, 0분, 0분 전
DT	보다 증가/하락하다	전년도, 지난해, 전일

4 한국어 어휘 의미망을 활용한 한국어 개체명 인식

4.1 학습 자질

본 논문은 형태소 단위의 CRF 모델을 기반으로 한 개체명 인식 시스템을 제안한다. 〈표 11〉은 한국어 어휘 의미망을 활용한 형태소 단위의 CRF 기반 개체명 인식 모델의 학습 자질을 나타낸 것이다. 개체명 인식 모델에서 사용하는 자질은 크게 기본 자질, 의미 자질, 구조 자질로 구분한다.

〈표 11〉 개체명 인식 모델의 학습 자질

자질	구분	설명
형태소 어휘	기본 자질	(-2, 2) 위치의 형태소 어휘 정보
형태소 품사	기본 자질	(-2, 2) 위치의 형태소 품사 정보
어휘 형태	기본 자질	입력 토큰의 lexical form
어절	기본 자질	(-2, 2) 위치의 어절 정보
접두, 접미부	기본 자질	현재 형태소의 접두, 접미부 음절
기분석 사전	기본 자질	학습 데이터에서 3회 이상 같은 개체로 분류된 형태소
상위어	의미 자질	2. 참고
의존관계	구조 자질	3.1 참고
간접적 의존관계	구조 자질	3.2 참고

기본 학습 자질은 문장을 구성하는 형태소와 음절, 어절을 조합한 자질이다. 현재 위치를 기준으로 앞, 뒤 2개의 형태소에 대해서 어휘 정보와 품사 태그의 조합한 형태소 어휘, 형태소 품사 자질과 현재 입력 토큰의 어휘 형태를 나타낸 어휘 형태 자질, 현재 형태소가 속한 어절과 해당 어절의 문장 내의 인덱스, 현재 형태소의 어절 내의 인덱스와 앞, 뒤 2개의 어절 정보를 나타낸 어절 자질, 기분석 사전 자질이 기본 자질로 분류된다. 기분석 사전 자질은 학습 데이터에서 3회 이상 같은 태그로 분류되는 개체명을 모아 놓은 사전을 이용한 자질이다. 예를 들어 '한국__0502/NNP'은 학습 시에 LC(지명)과 OG(기관)으로 분류된 경우가 모두 3

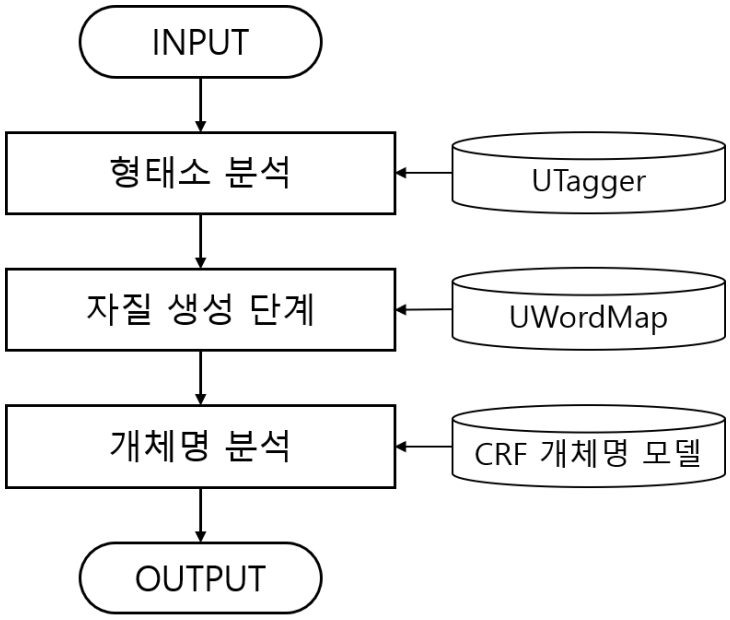
회 이상이기 때문에 기분석 사전에 이 정보를 가지고 있으며 이를 학습 자질로 활용한다.

〈표 12〉 개체명 인식 모델의 학습 자질 예시

국제빙상경기연맹(ISU)은 올 시즌 회전수가 부족한 점프에 대한 판정을 두 단계로 나눴다	
형태소 어휘	국제빙상경기연맹, (, ISU,), 은, @SP@, 올
형태소 품사	NNP, SS, SL, SS, JX, @SP@, NNG
어휘 형태	한글 → '가', 영문 → 'A', 숫자 → '0', 기호 → '.'
어절	국제빙상경기연맹(ISU)은, 올, 시즌, 회전수가
접두, 접미부	국, 국제, 맹, 연맹
기분석 사전	국제빙상경기연맹 : OG
상위어	시즌 → 시기(상위어), 시간(최상위어)
의존관계	국제빙상경기연맹(의존소) → 나누/VV(지배소)
간접적 의존관계	생략

〈표 12〉는 예문 "국제빙상경기연맹(ISU)은 올 시즌 회전수가 부족한 점프에 대한 판정을 두 단계로 나눴다"에 대해 각각의 학습 자질 생성 예시를 나타낸 것이다. 현재 예문에 대한 의존관계 분석 결과가 존재하기 때문에 간접적 의존관계 자질을 생략된 것을 확인할 수 있다.

4.2 한국어 개체명 인식 시스템



[그림 11] 제안 모델의 전체 시스템 구조

[그림 11]은 본 논문에서 제안하는 개체명 인식 모델의 전체 구조이며 입력 문장을 형태소 분석 단계, 개체명 자질 생성 단계, 개체명 분석 단계를 거쳐 개체명 인식 결과를 출력한다.

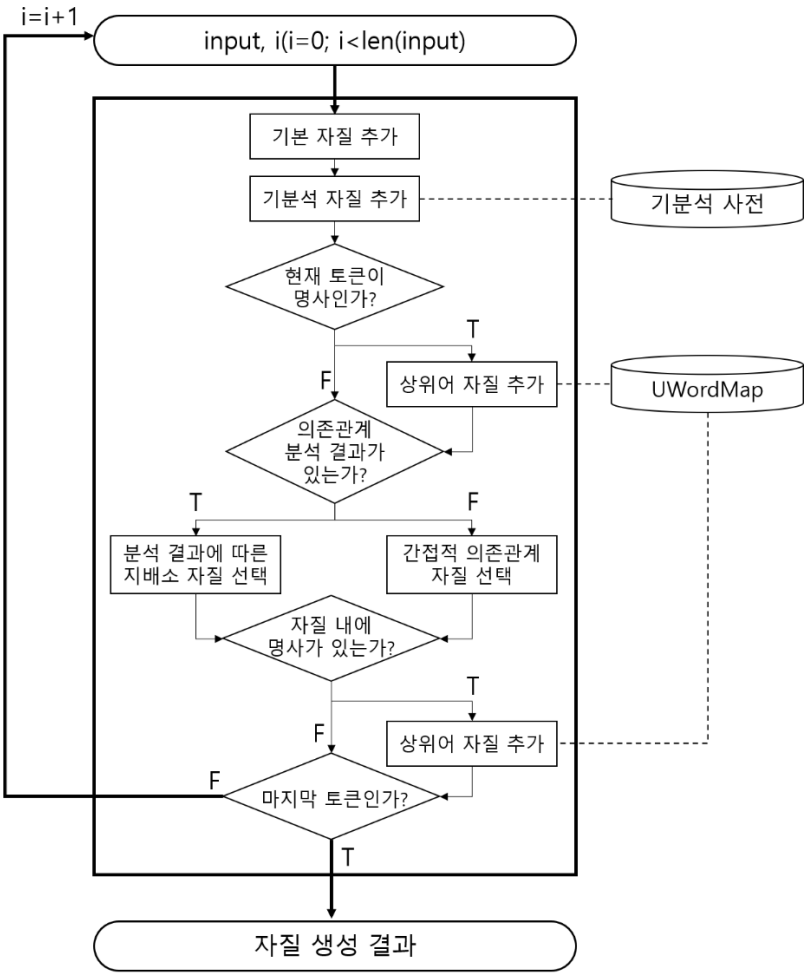
4.2.1 형태소 분석 단계

형태소 분석 단계에서는 울산대학교의 동형이의어, 다의어 및 형태소 분석기 UTagger를 이용하여 입력 문장에 대한 형태소 분석 결과와 다의어 분별 결과를 얻는다. 형태소 및 다의어 분별 결과는 형태소 단위로 표층형, 다의어 번호, 품사 태그가 부착되어 출력된다. 분석된 결과를 각각의 토큰으로 분리한 후, 어절과 어절 사이에 “@SP@” 기호를 추가하여 어절 단위를 구분하여 〈표 13〉과 같이 가공을 하여 자질 생성 단계의 입력으로 사용한다.

〈표 13〉 예문의 형태소 분석 결과 및 가공 결과

입력 문장	곧바로 귀국해 9일 제주도를 향한다
형태소 분석 결과	곧바로/MAG 귀국하/VV+여/EC 9/SN+일__07/NB 제주도__01/NNP+로/JKB 향하__000001/VV+ㄴ다/EF
가공된 입력 결과	곧바로/MAG @SP@ 귀국하/VV 여/EC @SP@ 9/SN 일__07/NB @SP@ 제주도__01/NNP 로/JKB @SP@ 향하__000001/VV ㄴ다/EF

4.2.2 자질 생성 단계



[그림 12] 자질 생성 단계의 구조

[그림 12]는 입력 문장에 대해 형태소 및 다의어 분석이 완료된 형태소 열을 이용한 개체명 인식을 위한 자질 생성 단계를 나타낸 것이다. 자질 생성 단계에서는 〈표 13〉의 가공된 입력 결과와 현재 형태소의 토큰 열에서의 인덱스를 입력으로 한다. 인덱스를 통해 현재 토큰 정보를 가져오며 현재 토큰에 대해 기본 자질을 추가하고 현재 토큰이 기분석 사전에 존재한다면 3회 이상 같은 분류되었던 범주에 대한 정보를 학습한다. 현재 토큰이 명사라면 울산대학교의 어휘 의미망 UWordMap의 정보를 이용하여 상위어와 최상위어 자질을 추가한다. 상위어와 최상위어는 다의어를 기준으로 추상화를 하며, 1계층 단어에 대한 평균 계층에 따라 최상위어 자질을 1계층 단어 혹은 2계층 단어로 지정하여 사용한다. 지배소 자질은 현재 문장에 대해 의존관계가 분석된 결과가 존재할 경우 의존관계 분석 결과에 따라 지배소를 선택하며, 분석 결과가 존재하지 않을 경우에는 간접적인 의존관계 추가 과정을 통해 지배소 자질을 선택한다. 선택된 지배소 자질에 명사가 존재할 경우, 지배소에 대한 상위어와 최상위어 자질을 추가한다. 이후, 현재 토큰이 입력 문장의 마지막 형태소라면 지금까지 생성한 자질을 반환한다.

4.2.3 개체명 분석 단계

개체명 분석 단계에서는 자질 생성 단계에서 반환된 자질을 학습 말뭉치로 학습을 완료한 CRF 기반 개체명 인식 모델의 입력으로 사용하여 개체명 태그를 예측하는 단계이다. 가능성이 있는 시퀀스 레이블 후보들에 대해 생성된 자질에 대해 점수를 계산하여 가장 적합한 하나의 시퀀스 레이블 열을 선택하는 방식으로 개체명 태그를 예측하여 결과를 출력한다.

5 실험 및 평가

본 논문에서 제안한 한국어 어휘 의미망을 활용한 CRF 모델 기반 한국어 개체명 인식 방법의 성능 평가를 위한 데이터로는 국립국어원의 개체명 분석 말뭉치와 한국전자통신연구원(ETRI) 엑소브레인 언어분석 말뭉치의 개체명 말뭉치를 사용하였다. 국립국어원 개체명 분석 말뭉치는 ETRI의 '세부분류 개체명 가이드라인 2018' 지침에 준하여 개체명의 경계를 인식하고 15개 의미 분류 체계에 따른 태그를 부착한 말뭉치이다. ETRI의 개체명 인식 말뭉치는 PS(Person, 사람), LC(Location, 장소), OG(Organization, 기관), TI(Time, 시간), DT(Date, 날짜)로 5개의 개체명 범주로 분류하여 부착한 말뭉치이다. 본 논문에서는 〈표 14〉와 같이 국립국어원 개체명 말뭉치와 ETRI 개체명 말뭉치 모두에 등장하며 개체명 인식 분야에서 가장 보편적으로 사용되는 5개의 개체명 범주로 제한하여 실험을 진행하였다.

〈표 15〉는 국립국어원의 개체명 인식 말뭉치의 예시이다. id는 말뭉치 내의 문장을 구분하기 위한 문서 번호를 의미하고 이 값을 이용하여 의존관계 분석이 완료된 문장을 국립국어원의 구문 분석 말뭉치에서 찾아 자질로 사용한다. from은 원문, word는 원문을 어절 단위로 구분한 값을 가지고 있으며 NE는 문장 내의 개체명을 모두 나타낸다. 본 논문은 형태소 단위를 입력으로 하는 개체명 인식 시스템이므로 〈표 15〉와 같은 말뭉치를 〈표 16〉과 같은 형태로 가공하여 학습에 사용하였다. 각각의 형태소 단위 토큰과 개체명 태그는 띄어쓰기로 구분을 하며 원문 내의 공백은 @SP@ 태그를 추가하여 구분하였다.

〈표 14〉 개체명 종류

	개체명 분류	표기
1	인물(PERSON)	PS
2	지역(LOCATION)	LC
3	기관(ORGANIZATION)	OG
4	시간(TIME)	TI
5	날짜(DATE)	DT

〈표 15〉 개체명 인식 말뭉치의 예시


```
{
        "id": "NWRW1800000029.315.1.1",
        "form": "[횡설수설/권순활]北 ‘외화벌이’ 뜯어먹기",
        "word": [
            {
                "id": 1,
                "form": "[횡설수설/권순활]北",
                "begin": 0,
                "end": 11
            },
            {
                "id": 2,
                "form": "‘외화벌이’",
                "begin": 12,
                "end": 18
            },
            {
                "id": 3,
                "form": "뜯어먹기",
                "begin": 19,
                "end": 23
            }
                ],
        "NE": [
            {
                "id": 1,
                "form": "권순활",
                "label": "PS",
                "begin": 6,
                "end": 9
            },
            {
                "id": 2,
                "form": "北",
```

```
                "label": "LC",
                "begin": 10,
                "end": 11
            }
        ]
},
```

〈표 16〉 가공된 개체명 인식 말뭉치의 예시

id	NWRW1800000029.315.1.1
tokens	[/SS 횡설수설/NNG //SP 권순활/NNP]/SS 北/SH @SP@ ‘/SS 외화벌이/NNG ’/SS @SP@ 뜯어먹/VV 기/ETN
labels	O O O B-PS O B-LC O O O O O O O

총 150,082 문장으로 구성된 국립국어원의 개체명 말뭉치를 형태소 분석기 UTagger를 사용하여 형태소 태그를 부착 후, 실험에 제한한 5개의 범주가 포함되지 않은 문장은 삭제하여 10배수 교차 검증 실험을 진행하였다. 본 실험을 진행한 환경은 다음 〈표 17〉**과** 같다.

〈표 17〉 실험 환경

운영체제	Windows 10
CPU	Intel® Core™ i7-5820K
RAM	32GB
개발 언어 및 도구	Python, python-crfsuite, Visual Studio Code

5.1 실험 결과

5.1.1 기본 자질 성능 비교

기본 자질만을 학습한 CRF 개체명 모델을 baseline으로 하여 기본 자질과 본 논문에서 제안한 추가 자질을 학습한 제안 모델의 개체명 인식 정확률(f1 score)과 처리 속도를 측정하고 10배수 교차 검증을 통한 평균값으로 성능을 비교하였다.

〈표 18〉 Baseline과 제안 모델의 개체명 인식 성능 비교(F1 score)

fold	baseline		제안 모델	
	F1-score	처리 속도	F1-score	처리 속도
1	90.37	5.46 sec	91.11	7.73 sec
2	90.51	5.53 sec	91.11	7.03 sec
3	90.34	5.46 sec	91.01	7.06 sec
4	90.23	5.42 sec	90.93	7.00 sec
5	90.58	5.56 sec	91.23	7.15 sec
6	90.24	5.40 sec	90.95	7.39 sec
7	90.47	5.55 sec	91.16	7.20 sec
8	90.52	5.50 sec	91.07	7.00 sec
9	90.11	5.47 sec	90.64	7.12 sec
10	90.58	5.47 sec	91.34	6.98 sec
평균	90.40	5.482 sec	91.05	7.065 sec

실험 결과, 기존 모델에 비해 본 논문의 추가 자질인 의미 자질(상위어, 최상위어)와 구조적 자질(의존관계, 격조사)를 추가로 학습한 모델이 모든 fold에서 f1-score가 향상했음을 확인할 수 있다. 반면에 테스트 데이터에 대한 개체명 인식 태그를 부착하는 처리 속도는 baseline에 비해 약 1.583초 늘어 속도 면에서의 성능은 하락했다.

〈표 19〉 Baseline과 제안 모델의 개체명 인식 성능 비교(F1 score)

	baseline	제안 모델	성능 차이
PS	92.47	93.74	+1.27
LC	85.43	86.69	+1.26
OG	83.91	84.56	+0.65
DT	96.21	96.33	+0.12
TI	93.64	93.61	−0.03
micro avg	90.40	91.05	+0.40

〈표 19〉는 각각의 개체명 범주에 대한 baseline과 제안 모델의 성능을 나타내고 있다. 실험 결과를 보면 기존 모델에 비해 한국어 어휘 의미망을 활용한 의미 자질을 추가 학습한 모델의 성능이 시간을 나타내는 TI 태그를 제외한 모든 태그에 대해 향상했음을 알 수 있다. 특히 인명을 나타내는 PS 태그는 F1 score 기준 약 1.27% 포인트, 지명을 의미하는 LC 태그의 성능은 F1 score 기준 약 1.26% 포인트 향상했다. 이는 한국어 어휘 의미망의 상위어 정보를 사용하므로 학습 데이터를 확장하여 OOV 문제를 보완했으며 문장 내의 의존관계 정보를 통한 격조사와 지배소 자질을 사용함으로 개체명 인식을 위한 키워드를 찾아내어 사용했음을 알 수 있다.

5.1.2 자질 별 개체명 인식 성능 비교

〈표 21〉은 〈표 18〉의 실험 결과, 평균 F1 score와 평균 처리 시간에 가장 가까웠던 fold 8을 대상으로 하여 각 자질의 조합에 따른 개체명 인식 성능, 학습 및 처리 속도를 비교한 결과이다. 학습 시간은 CRF 기반 개체명 인식 모델의 max iteration을 130으로 설정하여 학습한 결과를 측정한 것이다. 〈표 20〉은 실험에 사용한 자질 조합의 종류를 나타내고 있다. 기본 자질만을 학습한 Baseline을 (0)으로 표기하여, (1)부터 (6)은 기본 자질과 번호에 따른 자질

조합을 학습한 모델을 나타낸다. 마지막 (7)은 본 논문에서 제안하는 모든 자질을 학습한 제안 모델을 나타낸다.

〈표 20〉 실험에 사용한 자질 조합

표기	자질 조합
(0)	Baseline(기본 자질)
(1)	간접적 의존관계 자질
(2)	의존관계 자질
(3)	상위어 및 최상위어 자질
(4)	(1) + (2)
(5)	(1) + (3)
(6)	(2) + (3)
(7)	제안 모델

실험 결과 간접적 의존관계 자질을 추가함에 따라 Baseline에 비해 0.15% 포인트 성능이 향상했다. 의존관계 자질을 추가했을 때는 0.31% 포인트, 상의어 자질을 추가했을 때는 0.36% 포인트 성능 향상을 보였다. 비교 결과, 상위어 및 최상위어, 의존관계, 간접적 의존관계 자질 순으로 개체명 인식 성능에 대한 기여도가 높음을 확인할 수 있었다. 반면, 학습 및 처리 시간은 Baseline이 가장 짧았으며, 각각의 자질을 추가함에 따라 늘어남을 확인하였다.

〈표 21〉 자질 추가에 따른 성능 비교

	F1 score	학습 시간	처리 시간
Baseline	90.52	1140.19 sec	5.50 sec
(1)	90.67	1277.80 sec	5.96 sec
(2)	90.83	1320.89 sec	6.23 sec
(3)	90.88	1324.70 sec	6.30 sec
(4)	90.94	1332.41 sec	6.29 sec
(5)	90.94	1373.98 sec	6.32 sec
(6)	91.00	1420.38 sec	6.48 sec
제안 모델	91.07	1460.46 sec	7.00 sec

5.1.3 기존 모델과의 성능 비교

〈표 22〉는 개체명 인식 분야의 기존 모델과 본 논문의 제안 모델의 성능을 비교한 결과이다. 본 논문에서는 학습 및 평가를 위한 말뭉치로 국립국어원 모두의말뭉치(개체명 인식 말뭉치)를 선정하였으나, 기존 모델과의 성능 비교를 위해 ETRI의 엑소브레인 개체명 인식 말뭉치를 이용한 성능 평가 또한 진행하였다. ETRI의 엑소브레인 개체명 인식 말뭉치는 총 10,000개의 문장이 있으며, 본 논문에서 제한한 5개의 개체명 범주에 대해서 분류되었다. 실험을 위해 울산대학교의 UTagger를 이용하여 형태소 분석을 한 후, 중복된 문장을 모두 제거하여 7,000문장(약 85%), 1,240문장(약 15%)으로 나누어 각각 학습 및 평가에 사용하였다. 제안 모델의 학습 자질 중 하나인 의존관계 자질은 입력 문장에 대한 의존관계가 분석이 완료된 경우에만 추가할 수 있으나, ETRI의 엑소브레인 개체명 인식 말뭉치에 대한 의존관계 분석이 완료된 데이터는 찾을 수 없어서 이를 제외한 자질만을 학습한 결과이다.

〈표 22〉 기존 모델과의 개체명 인식 성능 비교

모델	F1 score
제안 모델	87.36
Stacked BiLSTM-CRF [14]	87.33
LSTM-CRF(문자기반)+개체명 사전 [7]	89.34
BERT-CRF [8]	91.58
ELECTRA-LAN [10]	92.78

[14]는 ETRI의 엑소브레인 개체명 인식 말뭉치의 중복된 문장을 제거하지 않은 10,00개의 문장을 모두 사용하여 5배수로 나눈 뒤 교차 검증 실험을 진행한 결과이다. [7]은 ETRI의 범용 개체명 인식 데이터를 사용하여 학습 및 평가를 했으며 한국어 위키 백과를 이용한 개체명 사전을 구축하여 자질로 사용한 결과이다. [8]은 사전 학습된 형태소 단위 BERT의 마지막 레이어 출력 값에 CRF를 연결하여 fine-tuning 후, 엑소브레인 개체명 인식 말뭉치를 사용

하여 성능을 측정한 결과이다. [10]은 20GB의 한국어 위키피디아, 뉴스 데이터를 사용하여 사전 학습한 음절 단위 ELECTRA 모델에 LAN 계층을 연결한 후, 엑소브레인 언어분석 말뭉치를 사용하여 성능을 평가하였다.

기존 모델과의 개체명 인식 성능을 비교한 결과, 기존 개체명 인식 분야에서 가장 많이 사용되며 준수한 성능을 보이던 딥 러닝 방식의 stacked BiLSTM-CRF 모델이 아닌 CRF 단일 모델을 사용했음에도 더 높은 성능을 보이는 것을 확인할 수 있었다. 제안 모델은 위키피디아를 이용한 개체명 사전을 사용한 [7]과 언어 모델을 이용하여 임베딩을 한 [8], [10]에 비해서는 낮은 성능을 보이고 있다. [7]과 [10]은 학습 및 평가에 사용한 말뭉치가 다르기 때문에 정확한 비교가 힘드나, 동일한 데이터를 사용한 [8]에 비해 성능이 떨어지는 것은 [8]의 모델이 위키피디아 코퍼스를 사전 학습을 했기 때문에 본 논문에서 제안하는 모델에 비해 OOV가 등장하는 빈도가 낮기 때문이라고 추측한다. 이는 [7]에서 제안한 위키피디아 코퍼스를 이용한 개체명 사전을 구축 후, 자질로 활용한다면 CRF 모델의 개체명 인식 성능과 언어 모델을 사용한 개체명 인식의 성능 차이를 줄일 수 있을 것이다. 또한, 제안 모델에서 학습하는 의존관계 자질을 제외한 학습 결과이기 때문에 같은 말뭉치에 대해 의존관계 자질을 추가로 학습할 경우 성능이 더 향상될 것이라고 예상한다.

개체명 인식 분야에서 학습 시간이나 처리 속도가 공개된 연구가 소수이며 정확한 속도 비교를 위해서는 동일한 실험 환경에서 진행하는 것이 중요하다. 따라서, 동일한 실험 환경에서의 학습 시간 및 처리 속도를 측정하기 위해서 개체명 인식 분야에서 가장 보편적으로 많이 사용된 stacked Bi-LSTM-CRF 모델을 구현하여 성능 비교를 진행하였다. 〈표 23〉은 비교 모델로 선정한 stacked Bi-LSTM-CRF 모델의 하이퍼파라미터이다. 평가셋으로는 제안 모델의 평균 성능에 가장 가까웠던 fold 8을 선택했으며, 전체적인 실험 성능은 가장 좋은 성능을 보인 16 epoch로 평가하였다.

〈표 23〉 비교 모델의 하이퍼파라미터에 따른 값

하이퍼파라미터	값
LSTM 층 수(n layers)	2
배치 사이즈(batch size)	128
학습률(learning rate)	0.01
드롭아웃 확률(dropout rate)	0.5
임베딩 크기(word embedding size)	300
LSTM 유닛 사이즈(LSTM unit size)	128

동일한 실험 환경에서의 성능 비교 결과, 제안 모델이 딥 러닝 기반의 stacked Bi-LSTM-CRF 모델보다 개체명 인식 정확률, 학습 속도, 처리 속도 모두 더 높은 성능을 보였다. 개체명 인식의 정확률 부분에서는 약 0.16% 포인트 더 높았으며, 학습 시간은 6시간 이상 차이가 났으며 평가셋을 모두 처리하는 시간은 2분 이상 차이가 났다. 이는 제안 모델이 학습 속도는 약 16배 빨랐으며 평가셋의 처리 속도는 약 18배 빠르다는 결과를 나타낸다. 이를 통해 기계학습 방식의 CRF 모델만을 이용하여 높은 성능과 빠른 학습 및 처리 속도를 모두 고려한 실용성을 높인 개체명 인식 시스템을 구축할 수 있음을 보였다.

〈표 24〉 딥러닝 개체명 인식 모델과의 비교

	정확률	학습 시간	처리 시간
Stacked BiLSTM-CRF	90.91	23754.32 sec	132.65 sec
제안 모델	91.07	1460.46 sec	7.00 sec
△	+0.16	-22293.86 sec	-125.65 sec

5.2 오류 분석

실험에서 나타난 오류는 크게 잘못된 개체명 범주가 부착된 유형과 개체명으로 인식하지 못 하는 유형으로 나눌 수 있다. 잘못된 개체명 범주가 부착된 유형은 다시 두 가지 경우로 나눌 수 있다. 첫 번째는 하나의 단어가 문맥에 따라 그 뜻이 상이해 잘못된 개체명 범주가 부착된 경우이다. 이 경우에는 '지명(LC)'과 '기관(OG)'를 혼동하여 부착한 경우가 가장 많았다. 〈표 25〉는 'LC'와 'OG'를 혼동하여 부착한 가장 대표적인 예이다.

〈표 25〉 개체명 범주를 혼동하여 부착한 예시

미국은 그동안 이라크전에 직•간접적으로 총 3조달러의 예산을 투입하였다.		
형태소	예측	정답
미국/NNP	B-OG	B-LC

靑 관계자 "X밴드 레이더가 中본토 탐지하는지 확인 원해"		
형태소	예측	정답
中/SH	B-LC	B-OG

미국/NNP'과 '中/SH"와 같은 어들은 문맥에 따라 장소를 의미하는 LC로 분류되거나 정부 혹은 나라의 대표팀 등의 의미를 가져 기관 및 조직을 의미하는 OG로 분류되어야 한다. 학습에서 주변 정보를 이용하여 문맥을 파악하고 정확한 개체로 분류할 수 있도록 하나 주변 정보가 부족하면 해당 의미를 명확히 파악하지 못해 잘못된 개체명 범주가 부착되는 경우가 다수 존재한다. 또한 이러한 오류에는 하나의 문장만으로 해당 형태소가 LC의 의미를 갖는지 OG의 의미를 갖는지 명확하지 않아 작업자의 주관적인 판단으로 이를 결정한 경우도 존재한다.

잘못된 개체명 범주가 부착되는 두 번째 경우는 국립국어원 모두의 말뭉치(개체명 인식 말뭉치)는 기존 15개의 의미 분류 체계에 따른 표지를 부착한 말뭉치이나 본 논문에서는 5개의 범주(인명, 지명, 기관명, 날짜, 시간)로 제한하여 생긴 경우이다. 예를 들어, 기존 15개의 의미 분류 체계에는 인공물을 나타내는 AF(ArtiFacts)가 존재한다. 이는 사람에 의해 창조된 대

상물을 말하며 문화재, 건물, 악기, 도로, 무기, 운송 수단, 작품명 등을 포함한다. 본 논문에서는 AF 태그를 제거하고 실험함에 따라 이를 OG 혹은 LC로 분류하는 경우가 다수 존재한다. 〈표 26〉은 범주를 제한함에 따라 생기는 오류의 예시이다.

〈표 26〉 개체명 범주를 제한함에 따라 생기는 오류의 예시

과거 정부에서 청와대 핵심 참모를 지낸 한 인사는 대통령이 바라보는 현실과 바깥에서 바라보는 현실의 괴리를 '임기 3년차 증후군'으로 설명했다.

형태소	예측	정답
청와대/NNP	B-OG	B-AF(O)

실험에서 나타난 오류의 두 번째 유형은 개체명을 인식하지 못한 오류이다. 개체명을 인식하지 못한 오류는 형태소 분석의 오류에 의한 것이 대표적이다. 이는 형태소 분석 오류에 의해 입력 토큰에 개체명에 해당하는 형태소가 존재하지 않는 경우이다. 예를 들어 "서울대공원장 자리를 알아봐주겠다며~"라는 문장에 대해 "서울대공원장"은 "서울대공원/NNP+장__41/XSN"으로 분석되어야 하며, 장소를 뜻하는 "서울대공원/NNP"를 지명(LC)로 분별해야 한다. 하지만 "서울대공원장/NNG"으로 분석이 되어 입력 토큰 내에 개체명에 해당하는 형태소인 "서울대공원/NNP"가 존재하지 않아 이를 지명으로 인식하지 못하게 된다.

6 결론

본 논문에서는 기존 딥 러닝 방식의 개체명 인식 모델이 고성능의 컴퓨팅 파워를 필요로 하며 학습 및 처리 속도가 떨어진다는 점을 개선하기 위하여 기계학습 방식의 CRF를 기반으로 한 한국어 개체명 인식 모델을 제안하였다. 딥 러닝 모델이 사람이 추론하고 생각하는 것에 착안하여 연구된 기계학습 알고리즘이라는 점에 따라 기계학습 알고리즘에 사람의 지적 정보를 학습 자질로 하여 간접적으로 딥 러닝 모델의 역할을 할 수 있도록 하였다.

본 논문에서 선정한 지적 정보는 기존 의미역, 의존관계 분석에 한국어 어휘 의미망(UWordMap)을 이용한 자질을 추가하여 성능 향상을 보인 연구들을 바탕으로 한국어 어휘 의미망의 명사 계층을 이용한 상위어와 최상위어 자질을 추가하였다. 대상 단어보다 큰 범주의 의미가 있는 단어인 상위어를 자질로 하여 학습 데이터를 확장하였으며 상위어를 개체명 인식을 위한 키워드로 사용함으로 성능 향상을 보였다.

개체명의 특성상 문맥에 따라 그 의미가 상이하다는 점에서 정확한 개체명 범주로 분류하기 위해서는 문맥을 파악하여 해당 정보를 사용하는 것이 필수적이라는 점에서 문장의 구조를 파악하여 그 정보를 반영하는 구조적 자질인 의존관계 자질과 격조사 자질을 추가하였다. 기존 기계학습 기반의 개체명 인식 모델에서는 현재 토큰을 기준으로 제한된 범위의 선행, 후행 토큰만을 이용하여 중요 키워드가 현재 토큰에서 일정 거리 이상 떨어져 있을 때 이 정보를 반영하지 못하는 문제가 있었으나 의존관계 및 격조사 자질을 학습함으로 이 점을 보완하여 개체명 인식의 성능 향상을 보였다.

제안한 개체명 인식 모델은 성능 평가를 위해 국립 국어원 모두의말뭉치(개체명 인식 말뭉치)를 형태소 및 동형이의어, 다의어 분별 시스템인 UTagger를 이용하여 형태소 태그를 부착하여 실험하였다. 실험 결과, F1 score 기준 91.05% 포인트의 성능을 보였다. 이는 기본 자질만을 학습한 Baseline에 비해 0.65% 포인트 향상한 성능이며 기존 딥 러닝을 사용한 모델에 준하는 성능임을 확인하였다. 또한, 같은 실험 환경에서 개체명 인식 분야에서 가장 보편적으로 사용되는 딥 러닝 모델인 stacked Bi-LSTM-CRF과의 비교 결과, 학습 속도는 약 16배,

처리 속도는 약 19배 빠른 결과를 보였다. 이를 통해 기계학습 방식의 CRF 모델만을 이용하여 높은 성능과 빠른 학습 및 처리 속도를 모두 고려한 실용성을 높인 개체명 인식 시스템을 구축할 수 있음을 보였다.

참고 문헌

[1] 김완수, 옥철영, “격틀 사전과 하위 범주 정보를 이용한 한국어 의미역 결정”, 정보과학회논문지, Vol43. pp.1376-1384, 2016

[2] 정충선, 신준철, 이주상, 옥철영, “의미 추상화를 이용한 전이 기반 한국어 의존관계 분석 시스템”, 정보과학회논문지, Vol46, pp.1174-1185, 2019

[3] 배영준, 옥철영, “한국어 어휘지도(UWordMap)와 API 소개”, 한국정보과학회 언어공학연구회 제26회 한글 및 한국어 정보처리 학술대회 논문집, pp.27-31, 2014

[4] 이창기, 장명길, “Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식”, 인지과학회 논문지, vol.21, pp.655 – 667, 2010

[5] 이태석, 전홍우, 강승식, “CRF를 이용한 특허 개체명 인식”, 한국정보과학회 학술발표논문집, pp.612 – 613, 2014

[6] 유홍연, 고영중, “Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장”, 정보과학회논문지, Vol.44, pp.306 – 313, 2017

[7] 민진우, 나승훈, “문자 기반 lSTM-CRF 한국어 개체명 인식을 위한 사전 자질 활용”, 한국정보과학회 언어공학연구회 제28회 한글 및 한국어 정보처리 학술대회 논문집, pp.119 – 121, 2016

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

[9] 민진우, 나승훈, 신종훈, 김영길, “RoBERTa를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱”, 한국정보과학회 한국소프트웨어종합학술대회 논문집, pp.407 – 409, 2019

[10] 김홍진, 오신혁, 김학수, “ELECTRA와 Label Attention Network를 이용한 한국어 개체명 인식”, 한국정보과학회 언어공학연구회 제32회 한글 및 한국어 정보처리 학술대회 논문집, pp.333-336, 2020

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

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

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

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

[Abstract]

CRFs based Named Entity Recognition Using A Korean Lexical Semantic Network

NER (Named Entity Recognition) is the task of classifying words with unique meanings that often appear as OOV (Out Of Vocabulary) within a given sentence into categories of predefined objects. To solve the OOV problem, researches have been conducted using deep learning to synthesize the words' embedding via CNN, LSTM networks or learning language models such as BERT or ELECTRA. However, models using these deep learning network or language model require high performance computing power and have low practicality due to slow speed of the learning model. For practicality, this paper proposes named recognition system based on CRFs, which adds features using external resources to supplement OOV problems.

In this paper, using Korean Lexical Semantic Network (UWordMap), the semantic features based on human knowledge were applied in the Korean named entity recognition system. By using hypernym as feature, it serves to expand training data to complement OOV problems, which are the biggest problem in Named Entity Recognition. In addition, most machine learning-based Named Entity Recognition models only reflect information about limited peripheral tokens in current token. Therefore, if a keyword is far away from current token, the model cannot reflect that information. To make up for this point, our model applied the structural features of dependence and investigation information.

As a result of the experiment, our model showed 91.05% named entity recognition accuracy, 1,466 sentences/sec processing speed. In the same experimental environment, performance comparison with the stacked Bi-LSTM-CRF, a deep learning model commonly used in named

entity recognition, resulted in improved accuracy, throughput, and training speed. The experiment showed that the machine learning based CRF model could be used to build a named entity recognition system that increases practicality considering both high performance and fast speed.