기계학습 기반 개체명 인식을 위한 사전 자질 생성

Feature Generation of Dictionary for Named-Entity Recognition based on Machine Learning

김 재 훈* · 김 형 철** · 최 윤 수***
Jae-Hoon Kim · Hyung-Chul Kim · Yun-Soo Choi

차 례

초 록

오늘날 정보 추출의 한 단계로서 개체명 인식은 정보검색 분야 뿐 아니라 질의응답과 요약 분야에서 매우 유용하게 사용되고 있다. 개체명은 일반 단어와 달리 다양한 문서에서 꾸준히 생성되고 변화되고 있다. 이와 같은 개체명의 특성 때문에 여러 응용 시스템에서 미등록어 문제가 야기된다. 본 논문에서는 이런 미등록어 문제를 해결하기 위해 기계학습 기반 개체명 인식 시스템을 위한 새로운 자질 생성 방법을 제안한다. 일반적으로 기계학습 기반 개체명 인식 시스템은 단어 단위의 자질을 사용하므로 구절 단위의 개체명을 그대로 자질로 사용할 수 없다. 이 문제를 해결하기 위해 본 논문에서는 새로운 구절 단위의 정보를 단어 단위의 자질로 변환하는 자질 생성 방법을 제안하였다. 이 방법으로 개체명 사전과 WordNet을 개체명 인식의 자질로 사용할 수 있었다. 그 결과 영어 개체명 시스템은 F1 점수의 약 6%가 향상되었고 오류의 약 38%가 줄어들었다.

키 워 드

개체명 인식, 워드넷, 개체명 사전, 자질 생성

* 한국해양대학교 컴퓨터공학과 교수
(Professor, Dept. of Computer Engineering, Korea Maritime University, jhoon@hhu.ac.kr)
** 한국해양대학교 컴퓨터공학과 석사과정
(Graduate Student, Dept. of Computer Engineering, Korea Maritime University, yhdosu@nate.com)
*** 한국과학기술정보연구원 정보기술연구실 선임연구원(교신저자)
(Corresponding Author, Senior Researcher, Dept. of Information Technology Research, KISTI, armian@kisti.re.kr)
• 논문접수일자: 2010년 3월 8일
• 최종심사일자: 2010년 3월 26일
• 게재확정일자: 2010년 3월 26일

ABSTRACT

Now named-entity recognition(NER) as a part of information extraction has been used in the fields of information retrieval as well as question-answering systems. Unlike words, named-entities(NEs) are generated and changed steadily in documents on the Web, newspapers, and so on. The NE generation causes an unknown word problem and makes many application systems with NER difficult. In order to alleviate this problem, this paper proposes a new feature generation method for machine learning-based NER. In general features in machine learning-based NER are related with words, but entities in named-entity dictionaries are related to phrases. So the entities are not able to be directly used as features of the NER systems. This paper proposes an encoding scheme as a feature generation method which converts phrase entities into features of word units. Futhermore, due to this scheme, entities with semantic information in WordNet can be converted into features of the NER systems. Through our experiments we have shown that the performance is increased by about 6% of F1 score and the errors is reduced by about 38%.

KEYWORDS

Named Entity Recognition, WordNet, NER Dictionary, Feature Generation

1. 서 론

개체명(Named-entity: NE)이란 문서에서 나타나는 고유한 의미를 가지는 명사나 숫자 표현과 같이 고유한 성질의 표현을 말하며 인명(Person: PER), 지명(Location: LOC), 기관명(Organization: ORG)과 같은 이름 표현, 날짜나 시간과 같은 시간 표현, 금액이나 퍼센트와 같은 수치 표현으로 구분할 수 있다(Chinchor et al. 1999). 대부분 하나 이상의 단어가 결합하여 개체명을 구성하게 된다.

개체명 인식(Named Entity Recognition: NER)은 문서에서 이러한 개체명을 추출하고 추출된 개체명의 종류를 결정하는 작업을 말한다(Nadeau and Sekine 2007). 1990년대에 정보추출(Information Extraction)의 목적으로 개최되었던 MUC(Message Understanding Conference)에서 개체명 인식을 정보추출의 일환으로 본격적으로 연구되기 시작하였으며(Grishman and Sundheim 1996), 그 후에도 많은 연구들이 활발하게 진행되어 왔다. 또한 정보추출시스템 외에 다른 자연언어처리 분야에서도 개체명 인식기를 핵심 구성 요소로 사용함에 따라 개체명 인식 시스템

의 개발 필요성 및 성능 향상에 대한 관심이 증대하게 되었다. 특히 최근 자연언어처리와 정보검색 분야에서 활발하게 연구되고 있는 질의응답(question answering) 시스템의 성능 개선을 위해서는 개체명 인식기의 성능향상이 필수적이다(Nadeau and Sekine 2007).

개체명 인식이 어려운 이유는 새로운 개체명이 꾸준히 만들어지고 있기 때문에 사전에 모든 개체명을 등록할 수 없다는 점과 같은 단어로 구성된 개체명이 문맥에 따라 다른 개체명으로 해석될 수 있는 중의성이 발생할 수 있다는 점이다. 이러한 문제점을 고려하여 개체명을 인식하기 위하여 예전에는 규칙에 기반한 방법을 많이 사용하였으며(Ravin and Wacholder 1996; Brin, 1998; Liu et al. 2006) 현재는 학습 말뭉치를 이용한 기계학습 방법을 많이 사용한다(Bikel et al. 1997; Borthwick 1998; Asahara and Matsumoto 2003; McCallum and Li 2003).

규칙 기반의 방법의 경우 미리 만들어 둔 사전에 등록된 개체명들은 대부분 100% 신뢰하고 개체명으로 인식하지만, 문맥상 개체명이 아닌 단어에도 사전에 등록되어 있다는 이유로 개체명으로 인식될 수 있는 문제점을 가지고 있으며, 또한 앞서 기술한 중의성 등의 문제가 발생하기 쉽다. 그에 비해 기계학습 방법은 주변 문맥 등 말뭉치에 존재하는 패턴을 이용하여 개체명을 인식하기 때문에 중의성 등에 대해 조금 더 안전하지만, 말뭉치에 존재하지 않는 개체명에 대해서는 그 정확도가 낮은 편이다. 이러한 문제를 해결하기 위하여 말뭉치와 독립적으로 개체명 사전을 구성하여 기계학습의 자질 중의 하나로 사용한다면, 인식률의 향상을 가져올 수 있다(Cohen 2004; Egorov et al. 2004). 또한 인식된 개체명을 분류할 때 개체명을 이루고 있는 단어 각각의 속성들을 이용한다면 좀 더 나은 인식률을 기대할 수 있을 것이다. 하지만, 단어 단위의 학습이 이루어지는 기계학습의 특징상 구절 단위로 이루어져 있는 개체명 사전 정보를 쉽게 적용할 수는 없다. 이 문제를 개선하기 위해 본 논문에서는 구절 단위의 사전 정보를 이용해서 단어 단위의 자질을 생성하는 방법을 제안한다. 이와 같은 방법으로 생성된 자질의 유용성을 보이기 위해 기계학습 기반 개체명 인식 시스템을 구성하고 생성된 자질을 이용해서 성능이 향상됨을 보였다. 성능 향상을 위하여 직접 구성한 개체명 사전과 WordNet을 이용하였다.

2장에서는 관련 연구에 대하여 간단히 설명하고, 3장에서는 기본 개체명 인식 시스템의 구성과 개체명 사전과 WordNet의 자질 생성 방법에 대해 서술한다. 4장에서는 3장에서 제시된 사전 정보를 이용한 개체명 인식 시스템을 실험을 통하여 그 유용성을 살펴본다. 끝으로 5장에서는 향후 연구과제에 대하여 생각해보고 결론을 맺고자 한다.

2. 관련 연구

2.1 개체명 인식 기술

개체명은 고유명사(복합명사 포함)와 시간 등을 나타내는 수식 표현을 말하나(Chinchor et al. 1999), 본 논문에서는 고유명사에 포함되는 인명(PER), 지명(LOC), 기관명(ORG)을 대상으로 하며 아래의 예는 개체명이 표시된 XML 형식의 문장이다.

〈PER〉Wolff〈/PER〉, currently a journalist in 〈LOC〉Argentina〈/LOC〉, played with 〈PER〉Del Bosque〈/PER〉 in the final years of the seventies in 〈ORG〉 Real Madrid〈/ORG〉.

〈표 1〉 BIO 태그로 표현된 개체명 인식

단어	BIO 태그	비고
Wolff	B-PER	
,	O	
currently	O	
a	O	
journalist	O	
in	O	
Argentina	B-LOC	지명
played	O	
with	O	
Del	B-PER	인명
Bosque	I-PER	
in	O	
the	O	
final	O	
years	O	
of	O	
the	O	
seventies	O	
in	O	
Real	B-ORG	기관명
Madrid	I-ORG	

개체명 인식은 정보 추출의 한 분야로서 문서 내에서 개체명을 추출하고 추출된 개체명의 종류(인명, 지명, 기관명 등)를 결정하는 작업을 말한다. 개체명 인식에 관한 연구는 MUC-6(Message Understanding Conferences)[1]에서 유래되어 최근에는 생의학 분야(Ananiadoua et al. 2004) 등에서 널리 사용되고 있다. MUC-6에 참가한 많은 시스템은 특정 언어에 제한된 규칙과 자신만의 입출력 방법을 사용하여 다른 언어나 다른 영역에 쉽게 적용할 수 없었다. MUC-6 이후 개체명에 대한 연구가 꾸준히 진행되었으며 CoNLL (Conference on Computational Natural Language Learning) 2002[2]와 2003[3]을 통해서 더욱 많은 발전이 있었다. 이 대회에 참가한 대부분의 시스템은 기계학습 방법을 이용하였으며 영어의 경우에 약 89%의 정확률을 보인다. 기계학습 방법에서는 주로 BIO 태그(B: 개체명의 시작, I: 개체명의 중간, O: 관계없음)(Ramshaw, L. A. and Marcus 1995)를 이용하는데 BIO 태그로 표현된 개체명 인식의 예는 〈표 1〉에서 나타내었다.

개체명 인식은 매우 다양한 방법으로 시도되었다. 가장 간단한 방법으로는 규칙 기반 방

1) http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
2) http://www.cnts.ua.ac.be/conll2002/ner/
3) http://www.cnts.ua.ac.be/conll2003/ner/

법이며 이 방법에서는 주로 정규표현(Regular Expression)(Brin 1998)이나 사전 탐색(Liu et al. 2006)을 주로 사용하지만 기계학습을 통해서 규칙을 학습하는 방법도 연구되었다(Black and Vasilakopolos 2002). 최근에 주로 이용되는 방법은 통계기반의 기계학습 방법이며, 대표적인 방법으로는 HMM(Hidden Markov Model)(Bikel et al. 1997), MEM (Maximum Entropy Model)(Borthwick 1998), SVM(Support Vector Machines) (Asahara and Matsumoto 2003), CRF (Conditional Random Fields)(McCallum and Li 2003) 등이 있다.

개체명 인식 연구가 시작된 90년대 말에는 주로 영어만을 대상으로 이루어졌으나 최근에서 영어뿐만 아니라 한국어(이창기 외 2006), 일본어(Utsuro et al. 2002), 중국어(Fu and Luke 2005), 불어(Poibeau 2003), 그리스어(Boutsis et al. 2000) 등 매우 다양한 언어에 대해서 개체명 인식 시스템이 개발되었으며 시스템의 응용 분야도 정보추출 뿐 아니라 정보검색, 질의응답 시스템 등 매우 다양한 분야로 확대되고 있다(Nadeau and Sekine 2007).

2.2 WordNet

WordNet은 프린스턴 대학 조지 밀러(George A. Miller)교수가 주도하는 대규모의 영어 어휘 데이터베이스이다(Miller 1995). WordNet은 영어 어휘를 명사(noun), 동사(verb), 형용사(adjective), 부사(adverb)로 크게 나누고, 이들 어휘의 동의어 집합(synset)을 정의하고, 이들 동의어 집합 간의 의미적 상관관계를 포함하고 있다. 이들 상관관계는 상하관계(a kind of) 또는 부분관계(a part of)로 정의되며 상하관계에는 상위어(hypernym)와 하위어(hyponym) 관계가 있고 부분관계에서 전체어(holonym)와 부분어(meronym)관계가 있다. 이 외에 반의어(anatonym), 양태어(troponym), 등위어(coordinate terms)의 관계가 포함되어 있다. 현재 WordNet은 15만5,287개의 어휘와 11만7,659개의 동의어 집합과 20만6,941개의 관계로 구성된 대규모 어휘사전이며[4] 현재에도 계속 개발 중이다.

2.3 조건부 확률장 모델(Conditional Random Fields Model)

은닉 마르코프 모델(Hidden Markov Model: HMM)은 관찰열(observation sequence)과 정답열(tag sequence)의 확률을 이용한 생성모델(Generative Model)이다(Rabiner 1989). 이 모델은 크게 두 가지의 문제가 있다. 첫째, 다양한 자질들을 자유롭게 사용할 수 없다. 왜냐하면 이 모델의 매개변수는 어휘 확률(Observation Probability)과 전이 확률(Tran-

4) http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

sition Probability)로 구성되며 다양한 자질들이 이들 확률로 모델링될 수 있어야 한다. 그러나 다양한 자질들이 이 두 확률로 모델링되기 위해서는 가능한 자질의 종류가 기하급수적으로 늘어나기 때문에 계산 속도가 크게 증가될 뿐 아니라 계산이 불가능할 수도 있다. 둘째, 다양한 문맥 정보를 이용할 수 없다. 문맥 정보는 전이 확률로 모델링되는데 일반적으로 사용되는 문맥이 이전의 한 개 혹은 두 개의 문맥 정도만 사용된다. 그러나 언어 현상 중에서 장거리 의존관계를 모델링하기에는 여러 가지의 제약을 가지고 있다.

이러한 문제를 해결하기 위해 조건부 모델(Conditional Model)이 적용된 최대 엔트로피 모델(Maximum Entropy Markov Model: MEMM)이 개발되었다(Ratnaparkhi 1997). MEMM은 HMM의 문제를 해결하였고, 대부분의 경우, HMM보다 좋은 성능을 보였다. 그러나 유한 상태 모델을 사용함으로써 label bias 문제를 야기하였다(Lafferty et al. 2001). 최근에는 label bias 문제를 해결하면서 MEMM 이상의 성능을 보이는 조건부 확률장 모델(Conditional Random Fields Model: CRF)이 개발되어 널리 사용되고 있다(Lafferty et al. 2001).

CRF는 품사 부착과 같은 연속적인 자료의 라벨(label)을 결정하는데 매우 유용한 분별 확률 모델(Discriminative Probability Model)이며(〈식 1〉 참조), 즉 주어진 입력 벡터 x에 대해서 조건부 확률 $p(y|x)$를 최대로 하는 라벨 y*를 선택하는 비방향성 그래프 모델이다(Lafferty et al. 2001).

$$y^* = \mathrm{argmax}_y p(y|x) \qquad \langle 식\ 1 \rangle$$

여기서 $p(y|x)$는 CRF의 종류에 따라 다양하게 정의될 수 있다. 품사 부착의 경우에는 선형 연쇄(Linear Chain) 모델이 적합하며, 〈식 2〉와 같이 구한다(Sutton and McCallum 2001).

$$p(y|x) = \frac{1}{Z(x)} exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t) \right) \qquad \langle 식\ 2 \rangle$$

여기서 $f_k(\cdot)$는 자질 함수(Feature Function)이며, 자질 k에 따른 특성 함수(Characteristic Function)이다. 즉 주어진 입력 y_{t-1}, y_t, x에 자질 k가 포함되어 있으면 1을 반환하고 그렇지 않으면 0을 반환한다. λ_k는 매개변수이며 자질 k의 가중치가 된다. λ_k의 학습 방법은 일반적으로 기울기 하강 알고리즘(Gradient Descent Algorithm: Generalized Iterative Scaling(GIS), Improved Iterative Scaling(IIS))과 준뉴턴 방법(Quasi-Newton Method: Limited Memory BFGS(L-BFGS))을 주로 사용한다. $Z(x)$는 정규화 요소이다.

3. 사전 자질 생성을 통한 영어 개체명 인식

본 절에서는 CRF를 이용한 영어 개체명 인

식 시스템에 대해서 살펴보고, 새로운 자질 생성 방법에 대해서 기술한다. 본 논문에서 제안된 자질 생성 방법은 개체명의 단어 단위 자질 생성과 WordNet 기반 의미 정보의 자질 생성 방법을 기술한다.

3.1 CRF를 이용한 영어 개체명 인식 시스템

2장에서 기술했던 것처럼 개체명 인식 시스템을 구현하기 위한 여러 가지 방법들이 시도되었으나 현재는 기계학습을 이용한 방법이 가장 널리 사용되고 있으며, 본 논문에서도 기계학습 기반의 개체명 인식 시스템을 사용한다. 기계학습을 이용하기 위해서는 정답이 부착된 말뭉치가 필요하다. 개체명 인식을 위해 말뭉치들에는 MUC-6&7(Grishman and Sundheim 1996), OntoNote 2(Hovy et al. 2006) 등이 있다. 이들 말뭉치들은 개체명은 부착되어 있지만, 문장 분리나 단어 분리 및 품사 정보가 포함되어 있지 않다. 그러나 일반적으로 개체명 인식은 개체명 인식 단위가 하나의 문장 안에서 결정되고, 품사 정보 또한 성능에 직접적인 영향을 미치는 자질 중에 하나이다(Baluja et al. 2000). 본 논문에서는 앞에서 언급한 모든 말뭉치를 사용하지 않고 OntoNote 말뭉치에서 WSJ 부분만 사용하였다. 왜냐하면 OntoNote(WSJ)는 Penn Treebank에 문장 분리, 토큰 분리, 품사 등의 정보가 부착되어 있기 때문에 정확한 정보를 사용할 수 있기 때문이다. 학습 알고리즘은 앞서 기술했던 CRF를 이용하였으며 본 논문에서 사용된 자질들은 〈표 2〉와 같다.

자질 Word는 단어 그 자체를 사용하며 적용 범위는 현재 단어(w_i)를 중심으로 이전/이

〈표 2〉 개체명 인식의 자질 집합

자질 이름	적용 범위	비 고
Word	w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, w_{i-1}/w_i, w_i/w_{i+1}	단어 그대로의 자질
POS	p_{i-2}, p_{i-1}, p_i, p_{i+1}, p_{i+2}, p_{i-1}/p_i, p_i/p_{i+1}, $p_{i-2}/p_{i-1}/p_i$, $p_{i+1}/p_i/p_{i+1}$, $p_i/p_{i+1}/p_{i+2}$	단어들의 품사
Word/POS	w_i/p_i	단어와 품사의 조합
BNP	n_{i-2}, n_{i-1}, n_i, n_{i+1}, n_{i+2}, n_{i-1}/n_i, n_i/n_{i+1}, $n_{i-2}/n_{i-1}/n_i$, $n_{i+1}/n_i/n_{i+1}$, $n_i/n_{i+1}/n_{i+2}$	기저구
Suffix3	w_i	단어의 접미문자 3개
Prefix2		
InitCap		
NE_Dic		

후 두 단어이다. 또한 현재 단어 주변의 연속 두 단어(w_{i-1}/w_i과 w_i/w_{i+1})도 매우 중요한 자질로 사용하였다. 자질 POS는 자질 Word에 대응하는 품사 정보이며 Word와 마찬가지로 적용 범위는 현재 품사(p_i)를 중심으로 이전/이후 두 단어이다. 품사의 경우에는 단어에 비해서 자료 부족 현상(Data Sparseness)이 심하지 않기 때문에 다양한 형태의 연속적인 품사 정보를 사용한다. 자질 Word/POS는 현재 단어와 품사 정보의 쌍이다. 자질 BNP는 기저구(Base-noun Phrase) 정보이고 Word와 마찬가지로 적용 범위는 현재 기저구(n_i)를 중심으로 이전/이후 두 기저구이다. 자질 Suffix3, Prefix2, InitCap, NE_Dic의 적용 범위는 모두 현재 단어이며 Suffix3는 길이가 3인 접미사이고 Prefix2는 길이가 2인 접두사이며, InitCap는 단어의 시작이 대문자로 시작하는지 유무이다. 접미사와 접두사의 길이는 실험을 통해서 결정되었으며 대문자에 관련된 다양한 정보 또한 실험을 통해서 결정되었다(김형철 외 2009). NE_Dic는 개체명 사전 정보이며 다음 절에서 자세히 설명할 것이다.

3.2 개체명 사전의 자질 생성

기계학습의 특성상 개체명 인식 시스템은 단어 단위로 자질이 구성되어야 한다. 그러나 개체명은 구절 단위로 이루어져 있기 때문에, 이를 해결하기 위하여 2장에서 기술한 것처럼 BIO 태그를 이용하여 정답을 인코딩하였다(Ramshaw and Marcus 1995). 이런 문제는 NER 사전을 학습 자질로 사용하고자 할 경우도 그대로 발생된다. 사전 자질은 학습이 토큰 단위로 이루어지지만 개체명 사전은 구절 단위로 이루어져 있기 때문이다. 또한 개체명은 단어의 형태는 같지만 그 종류가 다른 경우(Washington의 경우 인명과 지명 모두 포함)가 발생하기 때문에 쉽게 적용할 수가 없다. 이와 같은 중의성을 자질에 정확하게 표현하지 않으면 단어 단위의 자질로 표현할 수 없다. 본 논문에서는 〈표 3〉과 같이 구절 단위의 개체명에서 발생되는 모든 중의성을 단어

〈표 3〉 개체명 사전의 자질 생성

개체명 사전		개체명 사전의 자질	
개체명	개체명 태그	단어	개체명 단어의 자질
franklin delano roosevelt	PER	franklin	BIP__
		delano	_IP__
		roosevelt	_IP___
abc news	ORG	abc	B___O
		news	_I__O
white house	LOC	white	BIPL_
		house	_I_L_
ronnie white	PER	ronnine	B_P__
		linda	B_P__
linda franklin	PER		

단위의 자질에 표현하는 방법을 제안하여 개체명 사전의 모든 개체명을 자질로 사용할 수 있도록 하였다. 〈표 3〉은 개체명 사전을 자질 생성의 예를 보여준다.[5] 개체명에 속한 단어(개체명 단어)는 5개의 문자(BIPLO)로 자질값을 표현하였다. 각 문자의 의미는 〈표 4〉에서 자세히 설명하고 있다. 〈표 3〉에서 '_'는 개체명 단어가 해당하는 의미로 사용되지 않았음을 의미한다.

3.3 WordNet의 자질 생성

WordNet에는 많은 의미 정보를 포함하고 있으며(Miller 1995), 의미 정보는 개체명 인식에 많은 도움을 준다(Wattarujeekrit 2005; Han and Zhao 2009). 본 논문에서는 WordNet의 의미 정보를 개체명의 자질로 사용한다. WordNet을 개체명 인식에 사용한 연구는 몇 개 있었으나(Magnini et al 2002; Negri and Magnini 2004) 본 연구에서처럼 직접 자질로 이용한 경우는 없었다.[6] 본 연구에서 WordNet에 있는 모든 단어(14만7,816개)의 의미를 28개의 기본 의미 집합으로 구분하였다. 기본 의미 집합을 결정한 기준은 WordNet의 각 노드에 포함된 단어 수를 기준으로 하였으며 추출된 노드에 포함된 하위 단어의 수는 대체로 골고루 분포되어 있다. 그 결과 WordNet의 3단계 혹은 4단계에 속하는 의미 집합이 선정 되었으며, 〈표 5〉는 본 논문에서 선정된 기본 의미 집합이다. 이와 같은 기본 의미 집합을 이용하여 WordNet에 포함된 모든 단어에 대하여 자질을 생성하였다. WordNet 단어의 자질은 어떤 기본 의미가 포함되는지를 표시하며 하나의 단어는 여러 개의 의미를 동시에 포함할 수 있다. 따라서 각 단어의 WordNet 자질값은 길이가 28인 {+, -}로 이루어진 문자열이다. 즉 +는 단어가 해당 의미를 포함함을 의미하고 -는 포함하지 않음을

〈표 4〉 개체명의 자질 표현에 사용된 문자의 의미

문 자	의 미
B	기본 개체명 사전에 첫 단어로 출현
I	기본 개체명 사전에 첫 단어가 아닌 단어로 출현
P	기본 개체명 사전에 인명으로 출현
L	기본 개체명 사전에 지명으로 출현
O	기본 개체명 사전에 기관명으로 출현

5) 참고로 〈표 3〉에서 개체명의 표제어는 대소문자를 구분하지 않았다.
6) 사전에 포함 여부를 사용하는 경우는 있었다(Cohen 2004).

의미한다. 각 의미의 위치는 〈표 5〉에 나타난 순서에 따라 결정되며, 〈표 6〉은 WordNet 단어의 자질을 생성한 예를 보이고 있다.

〈표 5〉 WordNet 말뭉치에서 추출된 기본 의미 집합

구 분	분 류 명
3 단계 계층(6개)	thing
	object, physical object
	causal agent, cause, causal agency
	substance, matter
	process, physical process
	abstraction
4 단계 계층 중 가장 많이 나오는 (22 개)	change
	freshener
	horror
	jimdandy, jimhickey, crackerjack
	security blanket
	stinker
	whacker, whopper
	living thing, animate thing
	psychological feature
	whole, unit
	group, grouping
	attribute
	communication
	location
	measure, quantity, amount
	part, piece
	relation
	agent
	material, stuff
	food, nutrient
	compound, chemical compound
	solid

〈표 6〉 WordNet 사전의 자질 생성

단 어	워드넷 검색 결과	태그
caftan	- object, physical object - whole, unit	_+_____+_____
eyes	- abstraction - psychological feature	_____+_____+_____
glaze	- object, physical object - abstraction - whole, unit - attribute	_+___+_____+_+_____

4. 실험 및 고찰

본 실험의 목적은 두 가지이다. 하나는 본 논문에서 제안된 자질 생성 방법이 개체명 인식 시스템에서 얼마나 유용한지를 살펴보는 것이고, 다른 하나는 의미 속성이 개체명 인식에 어떤 영향을 미치는지를 살펴보는 것이다. 본 절에서는 실험 환경으로 말뭉치의 구성과 성능 척도에 대해 살펴보고 제안된 자질 생성 방법의 유용성을 평가할 것이다.

4.1 말뭉치

본 연구에서는 3장에서 기술한 것처럼 OntoNote2(WSJ) 말뭉치를 사용하였다. 전체 말뭉치는 30만8,736 어절로 구성되어 있으며, 문장 분리, 단어 분리, 품사, 기저구, 개체명 정보가 부착되어 있다. 이 말뭉치를 2:1의 비율로 나누어서 각각 학습과 실험 말뭉치로 사용하였다. 학습 말뭉치는 20만3,715 단어로 구성되었으며 약 8,600개의 개체명이 포함되어 있다. 이 크기는 실용적인 개체명 인식 시스템을 구현하기에는 충분한 크기가 아니라고 생각한다.[8] 실험 말뭉치는 10만5,021 단어로 구성되었으며 약 3,900개의 개체명이 포함되어 있다.

개체명 사전을 만들기 위한 개체명은 크게 2가지 방법을 이용하여 수집하였다. 첫 번째 방법은 개체명이 부착된 말뭉치에서 정규표현식 등을 이용하여 수집하였으며, 두 번째는 Wikipedia[9]에 존재하는 개체명들을 수작업으로 직접 수집하였다. 〈표 7〉에 수집된 개체명 수가 표시되어 있다.

4.2 성능 척도

개체명 인식 시스템의 성능은 주로 재현율(recall)과 정확률(precision)을 이용한다. 재

〈표 7〉 수집된 개체명 수

말뭉치	PER	LOC	ORG
OntoNote2[7]	4,141	2,192	3,894
MUC	2,888	1,157	2,759
ACE	7,579	976	4,592
Wikipedia	18,424	16,010	14,111

7) 실험 말뭉치에 포함된 개체명을 포함되지 않았다.

8) CoNLL-2003에서 사용한 학습 말뭉치의 크기와 비슷하며 이 대회에서 가장 좋은 성능을 보인 시스템은 약 89%의 F1-점수를 보였다(Kim Sang and de Meulder 2003). 이는 약 10개의 개체명마다 1개의 오류를 포함한다.

9) http://wikipedia.org/

현율은 문서에 포함된 개체명 중 시스템이 정확히 찾은 개체명 수의 비율을 말하고 정확률은 시스템이 찾아낸 개체명 중 정확히 찾은 개체명 수의 비율을 말한다. 응용 분야에 재현율이 높은 개체명 인식기를 선호하는 경우도 있고 정확률이 높은 개체명 인식기를 선호할 수도 있다. 따라서 시스템의 성능이 얼마나 향상되는지를 살펴보기 위해서는 일반적으로 F1 점수(F1 Score: F1)를 많이 사용한다. F1 점수는 재현율과 정확률의 조화 평균으로 구하며 아래의 식과 같다.

$$F1 = 2 \times \frac{pricision \times recall}{pricision + recall}$$

4.3 성능 평가

본 논문에서 제안된 영어 개체명 인식기는 CRF를 기반으로 구현하였다. 〈표 7〉은 본 논문에서 제안된 영어 개체명 인식기의 성능을 보이고 있다. 〈표 7〉의 "기본 시스템"은 〈표 2〉에서 소개한 기본 자질을 이용한 개체명 인식 시스템의 성능이다. 이 시스템은 약 85%의 F1-점수를 보였으며 재현율보다 정확률이 더 좋은 시스템이다. 〈표 8〉의 "개체명 사전 적용"은 〈표 2〉의 기본 자질과 개체명 사전의 자질을 이용한 개체명 인식 시스템의 성능이다. 이 성능은 F1-점수의 약 2%가 향상되었으며 오류의 약 27%가 줄었다. 이는 본 논문에서 제안된 자질 생성 방법이 합리적이며 개체명 인식 시스템에서 매우 유용한 자질로 사용되고 있음을 알 수 있었다. 〈표 8〉의 "WordNet 사전 적용"은 〈표 2〉의 기본 자질과 개체명 사전의 자질 뿐 아니라 WordNet의 의미 정보 자질을 추가한 개체명 인식 시스템의 성능이다. 이 성능은 기본 시스템에 대해 F1-점수의 약 6%가 개선되었으며 오류의 약 38%가 줄었다. 또한 이 성능은 개체명 사전만 추가했을 경우에 비해 F1-점수의 약 2%가 개선되었

〈표 8〉 사전 자질에 의한 성능 향상

구 분		기본 시스템	개체명 사전 적용	WordNet 사전 적용
실제 NE의 수(개)		3,986		
시스템이 낸 NE의 수(개)		3,556	3,729	3,817
맞은 NE의 수(개)		3,229	3,418	3,527
재현율(%)		81.00	85.75	88.48
정확률(%)		88.32	91.66	92.40
F1-점수(%)		84.50	88.60	90.40
기본 시스템에 대한 개선율(%)	F1-점수	-	4.10	5.90
	오류율	-	26.45	38.06

으며 오류의 약 16%가 줄었다. 이 결과는 WordNet의 의미 정보가 개체명 인식 시스템에서 매우 유용한 자질로 사용될 수 있음을 알 수 있었다.

5. 결론

본 논문에서는 기계학습 기반 개체명 인식 시스템을 위한 새로운 자질 생성 방법을 제안한다. 이 자질 생성 방법은 여러 가지 중의성을 가진 다양한 자질을 생성할 수 있었으며 본 논문에서는 두 가지 영역에 적용해 보았다. 하나는 개체명 사전의 개체명을 단어 단위의 자질로 생성하였고 또 다른 하나는 WordNet의 의미 정보에 적용하여 단어 단위의 자질을 생성하였다. 그 결과 영어 개체명 인식 시스템에서 F1점수의 약 6%가 향상되었고 오류의 약 38%가 줄어들었다. 따라서 본 논문에서 제안된 자질 생성 방법은 개체명 인식에 매우 적합한 방법임을 알 수 있었으며 개체명 사전과 WordNet의 의미 정보가 개체명 인식에 매우 유용한 자질임을 알 수 있었다.

앞으로 실용적인 개체명 인식 시스템으로 개발하기 위해서 다양한 형식의 학습 말뭉치 확장 방법이 연구되어야 할 것이다. 또한 WordNet의 의미 정보를 좀 더 많은 단어에 적용하기 위한 방법이 연구되어야 할 것이다. 예를 들면 여러 가지의 굴절(inflection) 현상에 무관하게 같은 자질이 생성되어야 할 것이다. 생/의학 분야나 화학 분야와 같은 분야에서는 다양한 기호가 어휘에 포함된다. 이와 같이 같은 의미를 가진 단어가 매우 다양한 형식으로 표현될 경우 어휘의 정규화가 개체명 인식에 많은 도움을 줄 수 있을 것으로 생각된다.

참고문헌

김형철, 김재훈, 최윤수. 2009. 접사 정보를 이용한 영어 미등록어의 품사부착 성능개선. 『한글 및 한국어 정보처리 학술대회 발표논문집』, 21(2009): 186-190.

이창기, 황이규, 오효정, 임수종, 허정, 이충희, 김현진, 왕지현, 장명길. 2006. Conditional Random Fields를 이용한 세부 분류 개체명 인식. 『한글 및 한국어 정보처리 학술대회 발표논문집』, 18(2006): 268-272.

최윤수, 정창후, 최성필, 류범종, 김재훈. 2009. 대용량 자원 기반 과학기술 핵심개체 탐지에 관한 정보추출기술 통합에 관한 연구. 『정보관리연구』, 40(4): 1~22.

Ananiadoua, S., Friedman, C., and Tsujii, J. 2004. "Introduction: named entity recognition in biomedicine." *Journal of Biomedical Informatics,* 37(6): 393-395.

Asahara, M. and Matsumoto, Y. 2003. "Japanese named entity extraction with redundant morphological analy-

sis." *Proceedings of the Human Language Technology Conference – North American chapter of the Association for Computational Linguistics,* 8–15.

Baluja, S., Mittal, V. and Sukthankar, R. 2000. "Applying machine learning for high performance named–entity extraction." *Proceedings of the Conference of the Pacific Association for Computational Linguistics,* 365–378.

Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R. 1997. "Nymble: a High–performance learning name–finder." *Proceedings of the Conference on Applied Natural Language Processing,* 194–201.

Black, W. and Vasilakopoulos, A. 2002. "Language independent named entity classification by modified transformation–based learning and by decision tree induction." *Proceedings of the 6th Conference on Natural Language Learning,* 159–162.

Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. 1998. "NYU: Description of the MENE named entity system as used in MUC–7." *Proceedings of the 7th Message Understanding Conference.*

Boutsis, S., Demiros, I., Giouli, V., Liakata, M., Papageorgiou, H. and Piperidis, S. 2000. "A system for recognition of named entities in Greek." *Lecture Notes in Computer Science,* 1835: 424–435.

Brin, S. 1998. "Extracting patterns and relations from the World Wide Web." *Proceedings of WebDB Workshop at 6th International Conference on Extending Database Technology,* 172–183.

Chinchor, N., Brown, E., Ferro, L. and Robinson, P. 1999. *Named Entity Recognition Task Definition,* version 1.4.

Cohen, W. 2004. "Exploiting dictionaries in named entity extraction: Combining semi–Markov extraction processes and data integration methods." *Proceedings of KDD,* 89–98.

Egorov, S., Yuryev, A. and Daraselia, N. 2004. "A simple and practical dictionary–based approach for identification of proteins in medline abstracts." *The Journal of the American Medical Informatics Association,* 11(3): 174–178.

Fu, G. and Luke, K.–K. 2005. "Chinese named entity recognition using lexi–

calized HMMs." *ACM SIGKDD Explorations Newsletter,* 7(1): 19-25.

Grishman, R. and Sundheim, B. 1996. "Message understanding conference - 6: A brief history." *Proceedings of the 16th International Conference on Computational Linguistics,* 466 -471.

Han, X. and Zhoa, J. 2009. "Named entity disambiguation by leveraging wikipedia semantic knowledge." *Proceeding of the 18th ACM conference on Information and Knowledge Management,* 215-224.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L. and Weischedel, R. 2006. "OntoNotes: The 90% solution." *Proceedings of Proceedings of the Human Language Technology Conference of the NAACL,* 57-60.

Kim Sang, E. F. T. and de Meulder, F. 2003. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." *Proceedings of the seventh conference on Natural Language Learning,* 142-147.

Lafferty, J., McCallum, A. and Pereira, F. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." *Proceedings of the 18th International Conference on Machine Learning,* 282-289.

Liu, H., Hu, Z. Z., Torii, M., Wu, C., and Friedman, C. 2006. "Quantitative assessment of dictionary-based protein named entity tagging." *Journal of the American Medical Informatics Association,* 13(5): 497-507.

Magnini, B., Negri, M., Prevete, R., and Taney H. 2002. "A WordNet-based approach to named entities recognition." *Proceedings of the International Conference On Computational Linguistics(on SEMANET: Building and Using Semantic Networks),* 1-7.

McCallum, A. and Li, W. 2003. "Early results for named entity recognition with conditional random fields, features induction and web-enhanced lexicons." *Proceedings of the Conference on Computational Natural Language Learning,* 188-191.

Miller, G. A. 1995. "WordNet: A lexical database for English." *Communications of the ACM,* 38(11): 39-41.

Nadeau, D. and Sekine, S. 2007. "A survey of named entity recognition and classification." *Journal of Linguisticae Investigationes,* 30(1): 3-26.

Negri, M. and Magnini, B. 2004. "Using

WordNet predicates for multilingual named entity recognition." *Proceedings of The Second Global WordNet Conference*, 169-174.

Poibeau, T. 2003. "The multilingual named entity recognition framework." *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics*, 155-158.

Rabiner. L. R. 1989. "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE*, 77(2): 257-286.

Ramshaw, L. A. and Marcus, M. P. 1995. "Text chunking using transformation-based learning." *Proceedings of the Third ACL Workshop on Very Large Corpora*, 82-94.

Ratnaparkhi, A. 1997. *A Simple Introduction to Maximum Entropy Models for Natural Language Processing*. University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-97-08.

Ravin, Y. and Wacholder, N. 1996. *Extracting Names from Natural-Language Text*. IBM Research Report RC 2033.

Lise Getoor and Ben Taskar. 2007. *Introduction to Statistical Relational Learning*. Cambridge, Mass: MIT Press.

Utsuro, T., Sassano, M. and Uchimoto, K. 2002. "Combining outputs of multiple Japanese named entity chunkers by stacking." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 281-288.

Wattarujeekrit, T. 2005. *Exploring Semantic Roles for Named Entity Recognition in the Molecular Biology Domain*. Ph.D. diss., Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies.