ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 43, No. 6, pp. 678-685, 2016. 6
http://dx.doi.org/10.5626/JOK.2016.43.6.678

Word Embedding 자질을 이용한 한국어 개체명 인식 및 분류

(Korean Named Entity Recognition and Classification using Word Embedding Features)

최 윤 수 † (Yunsu Choi)　　차 정 원 †† (Jeongwon Cha)

요 약 한국어 개체명 인식에 다양한 연구가 있었지만, 영어 개체명 인식에 비해 자질이 부족한 문제를 가지고 있다. 본 논문에서는 한국어 개체명 인식의 자질 부족 문제를 해결하기 위해 word embedding 자질을 개체명 인식에 사용하는 방법을 제안한다. CBOW(Continuous Bag-of-Words) 모델을 이용하여 word vector를 생성하고, word vector로부터 K-means 알고리즘을 이용하여 군집 정보를 생성한다. word vector와 군집 정보를 word embedding 자질로써 CRFs(Conditional Random Fields)에 사용한다. 실험 결과 TV 도메인과 Sports 도메인, IT 도메인에서 기본 시스템보다 각각 1.17%, 0.61%, 1.19% 성능이 향상되었다. 또한 제안 방법이 다른 개체명 인식 및 분류 시스템보다 성능이 향상되는 것을 보여 그 효용성을 입증했다.

키워드: 자연어 처리, 개체명 인식 및 분류, 단어 표현, CBOW 모델

Abstract Named Entity Recognition and Classification (NERC) is a task for recognition and classification of named entities such as a person's name, location, and organization. There have been various studies carried out on Korean NERC, but they have some problems, for example lacking some features as compared with English NERC. In this paper, we propose a method that uses word embedding as features for Korean NERC. We generate a word vector using a Continuous-Bag-of-Word (CBOW) model from POS-tagged corpus, and a word cluster symbol using a K-means algorithm from a word vector. We use the word vector and word cluster symbol as word embedding features in Conditional Random Fields (CRFs). From the result of the experiment, performance improved 1.17%, 0.61% and 1.19% respectively for TV domain, Sports domain and IT domain over the baseline system. Showing better performance than other NERC systems, we demonstrate the effectiveness and efficiency of the proposed method.

Keywords: natural language processing, named entity recognition and classification, word embedding, continuous bag-of-words model

· 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No. R0101-16-0054, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)
· 이 논문은 제42회 동계학술발표회에서 'Word Embeddings 자질을 이용한 한국어 개체명 인식 및 분류'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 창원대학교 친환경해양플랜트FEED공학과
cyunsu77@changwon.ac.kr
†† 종신회원 : 창원대학교 컴퓨터공학과 교수
(Changwon National Univ.)
jcha@changwon.ac.kr
(Corresponding author임)

논문접수 : 2016년 2월 15일
(Received 15 January 2016)
논문수정 : 2016년 4월 4일
(Revised 4 April 2016)
심사완료 : 2016년 4월 6일
(Accepted 6 April 2016)

정보과학회논문지 제43권 제6호(2016. 6)

1. 서 론

개체명(Named Entity)이란 문서에서 특정한 의미를 가지고 있는 단어 또는 어구를 말한다. 정보 검색에서 개체명은 주요 검색 대상이 되며, 질의/응답에서는 주요 질의/응답 대상이 된다. 이러한 개체명을 추출하기 위해 자연어 처리 분야에서 개체명 인식 및 분류(Named Entity Recognition and Classification) 연구가 활발하게 진행되었다.

개체명 인식 및 분류에 관한 연구는 영어권에서 먼저 발전하였다. 영어권에서는 대문자나 호칭 기호 자질 등 영어에서 나타나는 특징을 이용하여 높은 개체명 인식 및 분류 성능을 보였다[1,2].

한국어 개체명 인식 및 분류에 관한 연구로 반지도 학습인 Co-Training 기법을 변형한 규칙 기반 방식이 있었다[3]. 지도 학습 방법으로는 CRFs(Conditional Random Fields)를 이용한 방법과 Structural SVMs (Support Vector Machines)과 Pegasos 알고리즘을 이용한 방법이 있었다[4,5]. 최근에는 딥 러닝을 이용한 개체명 인식에 대한 연구가 있었다[6]. 또한 개체명 인식 및 분류 성능 향상을 위해 개체명 사전을 구축하고 확장하는 방법이 있었다[7,8]. 하지만 한국어 개체명 인식 및 분류에 대해 많은 연구가 있었으나, 영어에서 나타나는 대문자와 같은 특정 자질의 부재로 개체명을 인식하기 어려운 점이 있다.

본 논문에서는 한국어 개체명 인식 및 분류를 위한 자질 부족 문제를 극복하기 위해, word embedding 자질을 개체명 인식 및 분류를 위한 자질로 사용하는 방법을 제안한다. CBOW 모델과 K-means를 이용하여 각각 형태소 단위의 word vector와 word cluster symbol을 생성하고, 이를 CRFs의 자질로 사용하여 개체명 인식 및 분류 실험을 수행하였다. 실험 결과 word embedding 자질을 개체명 인식 및 분류에 사용할 경우 의미 있는 성능이 향상이 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 기존 영어와 한국어 개체명 인식 및 분류에 대한 연구와, word embedding에 대하여 기술한다. 3장에서는 word embedding을 한국어 개체명 인식 및 분류에 적용한 방법에 대하여 설명한다. 4장에서는 제안된 방법을 이용한 다양한 실험에 대해 기술하고 실험 결과에 대한 분석과 토의를 한다. 마지막으로 5장에서 결론을 도출하고 향후 과제를 기술한다.

2. 관련 연구

2.1 개체명 인식 및 분류

개체명 인식 및 분류에 관한 연구는 영어권에서 먼저 발전하였다. 초기 개체명 인식은 HMM(Hidden Markov Model)을 이용하여 사람, 단체, 지역, 시간, 날짜, 백분율, 금액, NOT-A-NAME 총 8개의 범주에 대하여 개체명을 부착하였다[1]. 이 연구에서는 대문자나 호칭 기호 등 영어에서 나타나는 문자의 특징을 자질로 사용하여 93%의 높은 성능을 보였다. 최근에는 트위터 글을 분석하여 개체명을 인식하는 실험이 있었다[2]. 트위터 글은 오타나 축약어, 신조어 등의 사용으로 단어의 원형을 복원하는 작업이 필요하다. 예를 들어 'tomorrow'라는 단어를 트위터에서는 '2morrow'나 'tmrw' 등으로 사용하기 때문에 이를 정규화 하는 작업과 함께 개체명을 인식하는 방법이다. 이 방법은 트위터 글을 학습하고 개체명 인식 실험을 수행하여 83.6%의 성능을 보였다.

한국어에서의 개체명 인식 및 분류에 대해서는 다음과 같은 연구가 있었다. 개체명 인식을 위한 학습 중 반지도 학습인 Co-Training 기법을 변형한 규칙 기반의 방식이 있었다[3]. 그리고 지도학습 방법으로 CRFs와 최대 엔트로피 모델(Maximum Entropy Model)을 이용하는 방법이 있었다[4]. CRFs로 개체명의 경계만을 인식하고 최대 엔트로피 모델을 이용하여 개체명을 분류하는 방법으로 83.4%의 성능을 보였다. 또한 Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식 방법이 있었다[5]. 이 방법은 CRFs를 이용한 방법[4]보다 높은 성능을 유지하면서 학습 시간은 4% 줄일 수 있었다. 최근에는 딥 러닝을 이용한 개체명 인식 또한 연구 되었는데, 영어에 비해 자질이 부족한 한국어에 자질 튜닝 작업에 들어가는 시간과 노력을 줄이면서 기존의 개체명 인식기 성능과 큰 차이가 없음을 보였다[6]. 또한 다른 방법으로 개체명 인식을 위해 개체명 사전을 이용하는 방법이 있다[7,8]. 개체명 인식 성능 향상을 위해 위키피디아를 이용하여 개체명 사전을 구축하고 확장하는 방법이다.

이와 같이 개체명 인식 및 분류에 대한 다양한 연구가 있었지만, 성능 개선을 위해 추가로 사용할 수 있는 자질이 부족한 문제를 가지고 있다. 이러한 자질 부족 문제를 해결하기 위해 최근 word embedding 자질을 개체명 인식 및 분류뿐만 아니라 다양한 자연어 처리 기술에 이용하는 연구가 진행되고 있다.

2.2 Word Embedding

Language model은 문장을 이루는 단어들의 확률분포이다. 음성 인식, 기계 번역, 형태소 분석, 필체 인식, 정보 분석 등의 분야에서 매우 중요한 정보로 사용되고 있다. Word embedding이란 language model의 하나로서 문장 속 단어들 사이의 관계를 비지도 학습(Unsupervised Learning)방식으로 분석하여 특징화 하는 것이다. 최근에는 다양한 word embedding 방법을 이용하여 영어의 chunking과 개체명 인식을 수행하고 각각의 성능

을 비교하는 연구가 있었다[9].

Word embedding 방법 중 인공 신경망을 이용하는 NNLM(Neural Network Language Model)은 뛰어난 성능을 나타내어 많은 연구에 참고 되었다[10]. 최근 새로운 word embedding 방법으로 CBOW 모델이 제안되었다[11]. CBOW 모델은 현재 word의 문맥을 이루는 vector들의 합으로 현재 word의 vector를 결정하는 모델이다. NNLM의 구조를 변경해 은닉층(Hidden Layer) 대신 투영층(Projection Layer)을 사용함으로써 학습시간을 100배 이상 단축시켰으며, NNLM보다 의미 정확도는 1%, 구문 정확도는 11% 높은 성능을 보였다.

본 논문에서는 NNLM보다 높은 성능을 보이는 CBOW 모델을 이용하여 word embedding을 수행하고, word embedding으로 생성한 자질을 한국어 개체명 인식 및 분류에 사용하는 방법을 제안한다.

3. Word Embedding 자질을 이용한 한국어 개체명 인식 및 분류

3.1 한국어 개체명 인식 및 분류 시스템

이 장에서는 기본 한국어 개체명 인식 및 분류 시스템에 대하여 기술한다.

우선 개체명 인식 및 분류 시스템은 인식 및 분류하고자 하는 범주를 결정한다. 본 연구에서는 ETRI에서 정의한 개체명 범주 중 상위 15개를 사용하고, 이 중 Study Field와 Theory는 그 의미가 유사하다고 판단하여 하나의 범주로 합쳐 표 1과 같이 총 14개(인명, 학술분야 및 이론, 인공물, 기관, 지역, 문명/문화 관련 명칭, 날짜, 시간, 수량 표현, 이벤트, 동물, 식물, 물질, 용어)의 개체명 범주를 사용하였다[4].

개체명 인식 및 분류는 인식과 분류의 문제를 2단계로 나누어 해결할 수도 있다. 본 연구에서는 인식과 분류의 문제를 한 번에 해결하는 방법을 사용하였다. 또한 개체명 인식 및 분류 단위로서 형태소 단위의 개체명 인식 및 분류 방법을 사용하였다. 형태소 단위로 개체명을 인식 및 분류하는 경우, 개체명의 형태소 경계를 구분해야하는 문제가 발생한다.

이와 같은 두 가지 문제를 해결하기 위해 본 연구에서는 표 1의 개체명 범주 태그에 B/I/O 형태를 결합한 개체명 태그를 사용하였다. B/I/O 형태는 개체명 시작(BEGIN), 개체명의 중간 혹은 마지막(INSIDE), 개체명이 아닌 것(OUTSIDE)으로 구성된다. 표 2에서 형태소 단위로 B/I/O형태의 개체명 태그를 부착하는 예를 보여준다.

본 연구에서는 개체명 인식 및 분류를 수행하기 위하여 기계학습 기반 방법 중 하나인 CRFs를 사용하였다. CRFs의 기본 자질은 기존 한국어 개체명 인식 및 분류 연구를 참고하여 표 3과 같이 사용하였다[4,5].

표 1 개체명 범주
Table 1 Named Entity Categories

	Category	Tag
1	PERSON	PER
2	FIELD	FLD
3	ARTIFACTS_WORKS	AFW
4	ORGANIZATION	ORG
5	LOCATION	LOC
6	CIVILIZATION	CVL
7	DATE	DAT
8	TIME	TIM
9	NUMBER	NUM
10	EVENT	EVT
11	ANIMAL	ANM
12	PLANT	PLT
13	MATERIAL	MAT
14	TERM	TRM

표 2 개체명 B/I/O 태그 예
Table 2 Example of NE B/I/O Tag

Morpheme	Morpheme Tag	NE B/I/O Tag
박지성	NNP	PER_B
은	JX	O
맨체스터	NNP	ORG_B
유나이티드	NNG	ORG_I
FC	SL	ORG_I
소속	NNG	O
축구	NNG	CVL_B
선수	NNG	CVL_I
이	VCP	O
다	EF	O
.	SF	O

표 3 한국어 개체명 인식 및 분류를 위한 자질
Table 3 Features for Korean NERC

Feature
Morpheme (-2 ~ +2)
POS tag (-2 ~ +2)
Length of morpheme
Position of morpheme in Eojeol
If last morpheme of Eojeol is a particle. Then last morpheme/POS tag in Eojeol. Otherwise '-'
Existing or not in NE dictionary (0/1)
Existing or not in common noun dictionary (0/1)

표 3에서 일반 명사 사전이란 명사들을 모두 모은 후, 개체명이 될 수 있는 명사를 제외시키고 남은 명사로 생성한 사전이다. 계속해서 생성되고 삭제되는 개체명의 특성상 개체명 사전만으로는 개체명을 인식하기 어렵기 때문에, 일반 명사 사전 존재 유무를 개체명 인식 및 분류의 자질로 사용한다. 일반 명사 사전에 존재한다는 것은 개체명이 아닐 가능성이 높다는 의미이며 일반 명사

사전에 존재하지 않는다는 것은 개체명일 가능성이 높다는 의미로 해석할 수 있다.

3.2 Word Embedding 자질을 이용한 개체명 인식 및 분류

이 장에서는 word embedding 자질을 개체명 인식 및 분류에 이용한 방법에 대하여 설명한다. 우선 word embedding을 수행하기 위해 대량의 원시 문서를 준비한다. 본 연구에서는 형태소 단위의 개체명 인식 및 분류 방법을 사용하였다. 또한 한국어의 경우 의미를 가지는 최소 단위가 형태소이므로, 형태소 단위의 word embedding을 수행한다. 따라서 원시 문서에 형태소 분석 및 품사 부착 단계를 수행하여 대량의 학습 말뭉치를 생성한다.

대량의 학습 말뭉치를 생성하고 이를 CBOW 모델을 사용하여 형태소 단위의 word embedding을 수행한다. CBOW 모델을 사용하였기 때문에, 현재 형태소의 주변 형태소가 가진 vector값으로 현재 형태소의 vector값이 결정된다. Word embedding을 수행하면 각각의 형태소는 d차원의 실수 값으로 이루어진 word vector가 생성되고, 이 word vector를 구성하는 d개의 실수 값을 CRFs의 자질로 사용하였다.

또한 형태소 단위로 생성된 word vector의 실수 값을 K-means를 사용하여 clustering을 수행한다. K개의 class로 clustering을 수행하면 각각의 형태소는 class 번호를 가진다. 이 class 번호가 word cluster symbol이 되고 각 형태소마다 하나의 CRFs 자질로 사용하였다.

Word embedding으로 생성된 word vector는 대량의 문서로부터 형태소가 가진 특징이나 의미 정보를 표현한 것이다. 따라서 word vector를 자질로 사용할 경우 형태소의 어휘나 품사 정보로는 얻을 수 없는 정보를 개체명 인식 및 분류에 사용할 수 있다.

CBOW 모델은 문맥을 정보를 이용하므로, 같은 개체명 범주를 가지는 형태소는 비슷한 문맥을 가질 수 있고 이는 비슷한 vector값을 가질 수 있다는 것을 의미한다. 따라서 K-means는 vector간 거리를 이용하므로, 같은 개체명 범주를 가지는 형태소는 같은 word cluster 정보를 가질 수 있다.

4. 실험 및 토의

4.1 실험환경

제안된 방법의 효용성을 보이기 위해서 다양한 실험을 진행하였다. 우선 개체명 인식 및 분류와 word embedding을 수행하기 위해 형태소 분석이 필요하다. 형태소 분석은 창원대학교 적응지능연구실에서 공개한 Espresso를 사용하여 수행하였다[12]. 또한 CRFs를 이용한 개체명 인식 및 분류를 수행하기 위해 CRF++ Tool[1]을 이용하였다.

한국어 개체명 인식 및 분류 시스템의 성능을 측정하기 위해서 ETRI 개체명 말뭉치의 TV 도메인과 스포츠 도메인, 그리고 IT 도메인 문서를 사용하였다. 실험을 위해서는 5-fold 교차 평가를 수행하였다. TV 도메인에서는 104,929문장을 학습 데이터로 사용하고 4,000문장을 테스트 데이터로 사용하였다. 스포츠 도메인에서는 42,809문장을 학습 데이터로 사용하고 4,000문장을 테스트 데이터로 사용하였다. 마지막으로 IT 도메인에서는 14,075문장을 학습 데이터로 사용하고 1,000문장을 테스트 데이터로 사용하였다.

Word embedding 자질을 이용하였을 때 개체명 인식 및 분류 성능을 알아보기 위해, 각 도메인에서 기본 개체명 인식 및 분류 시스템 성능과, word vector 자질과 word cluster 자질을 추가로 사용한 개체명 인식 및 분류 성능을 비교 분석하였다. 그리고 기본 시스템에 word vector 자질과 word cluster 자질을 모두 사용하였을 때의 성능을 비교 분석하였다.

Word embedding에는 총 2억 8천만 개 형태소를 학습에 사용하고 50차원의 실수로 이루어진 569,589개 word vector를 생성하였다. Word cluster 자질은 앞서 생성된 word vector와 K-means를 이용하여 clustering을 수행하였다. 각각 200, 300, 400, 500개의 class로 clustering을 수행하여 class 번호를 word cluster 자질로 사용하였다.

4.2 실험 결과

4.2.1 Word Vector를 사용한 성능 평가

표 4, 표 5와 표 6은 TV 도메인과 Sports 도메인 그리고 IT 도메인에서의 한국어 개체명 인식 및 분류 실험 결과이다. 각 자질별로 5개의 테스트 셋을 이용한 실험 결과로 나온 성능의 평균을 측정하였다. 우선 CRFs를 사용한 기본 시스템에 word vector 자질을 추가로 사용하고 실험을 수행하였다.[2] Word vector 자질을 추가로 사용하였을 때 TV 도메인에서는 89.4%로 기본 시스템보다 0.76% 성능이 향상되었다. Sports 도메인에서는 89.68%로 기본시스템보다 0.25% 성능이 향상되었고, IT 도메인에서는 85.49%로 0.4% 성능이 향상되었다. 이 실험 결과를 통해 word vector 자질을 추가로 사용하였을 경우 한국어 개체명 인식 및 분류 성능을 향상 시킬 수 있음을 알 수 있다. 하지만 TV 도메인에 비해 다른 두 도메인에서는 성능의 향상 폭이 상대적으로 작았다. 따라서 word embedding 자질로 word vector 자질 외에 추가로 다른 자질을 사용할 필요성이 있음을 알 수 있다.

---

1) https://taku910.github.io/crfpp/

2) Word vector는 실수 값으로 소수점 여섯째자리에서 반올림하여 소수점 다섯째자리로 사용하였다.

표 4 Word embedding 자질을 이용한 한국어 개체명 인식 및 분류 실험 결과 (TV 도메인)
Table 4 Experimental Result of Korean NERC using Word Embedding Features (TV Domain)

System	Precision(%)	Recall(%)	F1 Score(%)
Baseline	89.54	87.76	88.64
Baseline + Word Vector	89.80	89.01	89.40
Baseline + Word Cluster(Class 200)	89.81	87.97	88.88
Baseline + Word Cluster(Class 300)	89.84	88.03	88.93
Baseline + Word Cluster(Class 400)	89.96	88.10	89.02
Baseline + Word Cluster(Class 500)	89.83	88.06	88.94
Baseline + Word Vector + Word Cluster(Class 400)	90.22	89.42	**89.81**

표 5 Word embedding 자질을 이용한 한국어 개체명 인식 및 분류 실험 결과 (Sports 도메인)
Table 5 Experimental Result of Korean NERC using Word Embedding Features (Sports Domain)

System	Precision(%)	Recall(%)	F1 Score(%)
Baseline	89.87	89.00	89.43
Baseline + Word Vector	89.91	89.45	89.68
Baseline + Word Cluster(Class 200)	90.04	89.14	89.59
Baseline + Word Cluster(Class 300)	90.23	89.26	89.74
Baseline + Word Cluster(Class 400)	90.34	89.36	89.85
Baseline + Word Cluster(Class 500)	90.22	89.28	89.75
Baseline + Word Vector + Word Cluster(Class 400)	90.36	89.72	**90.04**

표 6 Word embedding 자질을 이용한 한국어 개체명 인식 및 분류 실험 결과 (IT 도메인)
Table 6 Experimental Result of Korean NERC using Word Embedding Features (IT Domain)

System	Precision(%)	Recall(%)	F1 Score(%)
Baseline	85.81	84.38	85.09
Baseline + Word Vector	85.93	85.05	85.49
Baseline + Word Cluster(Class 200)	86.33	84.97	85.65
Baseline + Word Cluster(Class 300)	86.35	85.05	85.69
Baseline + Word Cluster(Class 400)	86.65	85.23	85.93
Baseline + Word Cluster(Class 500)	86.39	85.07	85.72
Baseline + Word Vector + Word Cluster(Class 400)	86.77	85.80	**86.28**

4.2.2 Word Cluster를 사용한 성능 평가

Word cluster 자질을 만들기 위해 앞서 생성한 word vector와 K-means를 사용하였다. K-means는 clustering 하고자 하는 class의 개수를 지정한다. 표 4, 표 5와 표 6에서 TV 도메인과 Sports 도메인, IT 도메인의 class 개수 별 성능을 보여준다. 모든 도메인에서 word cluster 자질을 추가로 사용하였을 때 기본 시스템보다 성능이 향상되었으며, class 400개로 생성한 word cluster 자질을 사용하였을 때 class 개수 별 성능 중 가장 높은 성능을 보였다. TV 도메인에서는 89.02%로 기본 시스템보다 0.38% 성능이 향상되었고, Sports 도메인에서는 89.85%로 0.42% 성능이 향상 되었다. IT 도메인에서는 85.93%로 기본 시스템보다 0.84% 성능이 향상 되었다. Word cluster 자질을 사용하였을 때, TV 도메인에서는 word vector 자질을 사용하였을 때보다 성능의 향상 폭이 작았다. 하지만 Sports 도메인과 IT 도메인에서는 word cluster 자질을 사용하였을 때 word vector 자질을 사용한 것 보다 더 높은 성능 향상이 있었다.

4.2.3 자질들을 조합했을 때의 성능 평가

마지막으로 기본 시스템에 word vector 자질과 word cluster 자질을 모두 사용하여 실험을 수행하였다. Word vector 자질은 4.2.1장에서 사용한 자질과 동일하게 사용하였다. Word cluster 자질은 4.2.2장의 실험에서 성능이 가장 높게 나온 class 400개로 clustering을 수행한 word cluster 자질을 사용하였다.

표 4, 표 5와 표 6에서 기본 시스템에 추가로 word vector 자질과 word cluster 자질을 모두 사용하였을 때 성능을 보여준다. TV 도메인에서는 89.81%로 기본 시스템보다 1.17% 성능이 향상 되었다. Sports 도메인에서는 90.04%로 0.61% 성능이 향상 되었으며, IT 도메인에서는 86.28%로 1.19% 성능이 향상되었다. 모든 도메인에서 word vector 자질과 word cluster 자질을 모두 사용하였을 때 가장 높은 성능을 보였다.

표 7 Baseline + Word Vector + Word Cluster (Class 400) Confusion Matrix (TV 도메인)

Table 7 Baseline + Word Vector + Word Cluster (Class 400) Confusion Matrix (TV Domain)

G \ S	PER	FLD	AFW	ORG	LOC	CVL	DAT	TIM	NUM	EVT	ANM	PLT	MAT	TRM	None	All
PER	1,254	1	6	0	3	6	0	0	1	0	2	0	1	1	133	1,408
FLD	1	404	3	1	5	3	0	0	2	1	0	0	1	3	84	508
AFW	10	4	457	2	6	9	3	1	6	0	5	0	2	3	142	650
ORG	2	2	1	232	15	2	0	0	2	1	0	0	0	1	45	303
LOC	5	1	3	4	1,006	7	0	0	1	2	2	0	0	2	88	1,121
CVL	4	5	4	1	6	1,990	1	0	7	2	1	1	0	4	298	2,324
DAT	2	1	1	0	5	1	2,029	0	12	2	0	0	0	0	19	2,072
TIM	0	0	0	0	0	0	2	282	8	0	0	0	0	0	1	293
NUM	2	1	1	1	0	4	25	2	3,557	1	0	0	0	0	51	3,645
EVT	1	1	0	2	4	1	1	0	2	50	0	0	0	1	19	82
ANM	2	0	3	0	3	3	0	0	2	0	705	0	0	4	95	817
PLT	0	0	0	0	0	0	0	0	0	0	0	120	0	0	19	139
MAT	1	1	3	0	0	0	0	0	1	0	0	0	166	2	31	205
TRM	2	4	4	1	2	3	0	0	2	0	2	0	1	682	184	887
None	94	46	73	37	37	216	55	19	312	10	61	8	16	100	-	1,084
All	1,380	471	559	281	1,092	2,245	2,116	304	3,915	69	778	129	187	803	1,209	

표 8 Baseline + Word Vector + Word Cluster (Class 400) Confusion Matrix (Sports 도메인)

Table 8 Baseline + Word Vector + Word Cluster (Class 400) Confusion Matrix (Sports Domain)

G \ S	PER	FLD	AFW	ORG	LOC	CVL	DAT	TIM	NUM	EVT	ANM	PLT	MAT	TRM	None	All
PER	2,661	0	0	6	3	8	0	0	1	1	1	0	0	1	164	2,846
FLD	0	14	0	1	0	0	0	0	0	0	0	0	0	1	12	28
AFW	2	0	107	4	8	2	0	0	0	3	0	0	0	0	31	157
ORG	6	0	2	2,388	38	10	0	0	3	24	0	0	0	1	139	2,611
LOC	4	0	3	46	1,254	3	0	0	0	17	0	0	0	1	50	1,378
CVL	6	0	1	13	6	2,775	2	0	12	12	1	0	0	4	380	3,212
DAT	0	0	0	0	0	0	1,143	1	86	2	0	0	0	0	28	1,260
TIM	0	0	0	0	0	0	0	208	2	0	0	0	0	0	5	215
NUM	1	0	0	0	0	6	2	1	3,829	3	0	0	0	1	94	3,937
EVT	3	0	2	31	20	16	3	0	12	752	0	0	0	2	213	1,054
ANM	0	0	0	0	0	1	0	0	0	0	168	0	0	2	30	201
PLT	0	0	0	0	0	0	0	0	0	0	0	4	0	0	5	9
MAT	0	0	0	0	0	0	0	0	0	0	0	0	2	0	4	6
TRM	2	0	0	4	1	4	0	0	2	1	3	0	0	704	207	928
None	143	4	17	127	28	371	29	6	119	194	19	1	1	177	-	1,236
All	2,828	18	132	2,620	1,358	3,196	1,179	216	4,066	1,009	192	5	3	894	1,362	

표 9 Baseline + Word Vector + Word Cluster (Class 400) Confusion Matrix (IT 도메인)

Table 9 Baseline + Word Vector + Word Cluster (Class 400) Confusion Matrix (IT Domain)

G \ S	PER	FLD	AFW	ORG	LOC	CVL	DAT	TIM	NUM	EVT	ANM	PLT	MAT	TRM	None	All
PER	610	0	1	3	1	0	0	0	0	0	0	0	0	0	57	672
FLD	1	302	0	5	1	3	0	0	3	0	0	0	0	16	75	406
AFW	3	4	127	6	5	2	0	0	2	1	0	0	0	4	44	198
ORG	5	10	3	1,395	15	9	0	0	1	4	0	0	0	7	112	1,561
LOC	1	1	2	6	556	3	0	0	0	2	0	0	0	2	40	613
CVL	1	8	1	14	2	928	0	0	4	2	0	0	0	6	141	1,107
DAT	0	0	0	0	0	0	607	0	26	0	0	0	0	0	7	640
TIM	0	0	0	0	0	0	0	21	0	0	0	0	0	0	1	22
NUM	0	0	0	0	0	1	0	0	712	0	0	0	0	0	12	725
EVT	0	5	2	8	3	4	0	0	2	45	0	0	0	6	33	108
ANM	0	0	0	0	0	0	0	0	0	0	8	0	0	0	6	14
PLT	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2
MAT	0	0	0	0	0	0	0	0	0	0	0	0	6	0	4	10
TRM	1	21	2	11	3	3	0	0	3	3	0	0	0	460	139	646
None	46	75	26	99	32	134	5	1	31	21	3	0	3	123	-	599
All	668	426	164	1,547	618	1,087	612	22	784	78	11	1	9	624	672	

4.3 오류 분석

실험에서 나타난 오류는 크게 두 가지 유형으로 나눌 수 있다. 하나는 잘못된 개체명 범주가 부착된 유형이다. 잘못된 개체명 범주가 부착된 유형은 다시 두 가지 경우로 나눌 수 있다. 첫 번째로 애매성에 의해 잘못된 개체명 범주가 부착된 경우이다. 예를 들어 '청와대'의 경우 '장소(LOC)'와 '기관(ORG)' 두 가지 범주에 모두 속하게 된다. 애매성이 발생할 경우 학습에서 주변 정보를 이용하여 애매성을 해결하게 된다. 하지만 주변 정보가 부족하면 애매성을 해결하지 못하고 잘못된 개체명 태그가 부착되는 오류가 발생한다. 잘못된 개체명 범주가 부착되는 두 번째 경우는 과 분석에 의한 오류로 개체명이 아닌 것에도 개체명 범주를 부착하는 것이다.

표 7, 표 8과 표 9는 각 도메인에서 가장 높은 성능을 보인 실험 결과의 confusion matrix이다. 표에서 열은 시스템(System) 출력 결과를 나타내고 행은 실제 정답(Gold)을 나타낸다. 표를 살펴보면 개체명이 아닌데 개체명을 부착한 경우가 상당히 많음을 알 수 있다. 또한 애매성이나 주변 정보에 의해 잘못된 개체명 범주가 부착된 오류의 수를 확인할 수 있다.

다른 하나의 유형은 개체명을 인식하지 못한 오류이다. 개체명을 인식하지 못한 오류는 다시 다음과 같은 경우로 나눌 수 있다. 첫 번째는 형태소 분석 오류에 의해 발생하는 개체명 인식 오류이다. 예를 들어 사람 '인명(PER)'으로 분류되는 개체명인 '박지성'은 형태소 분석에서 '박지성/NNP'으로 분석되어야 한다. 하지만 '박지/NNG+성/XSN'으로 잘못 분석될 경우 개체명 인식 오류를 발생시킬 수 있다. 두 번째는 2어절 이상의 개체명의 경우 어절 모두를 개체명으로 인식하지 못한 오류이다. 예를 들어 '웨인 루니'나 '크리스티아누 호날두'같은 개체명에서 '웨인', '크리스티아누' 등이 잡히지 않는 오류이다. 표 7, 표 8과 표 9를 살펴보면 개체명 범주별 성능에서 '이론 및 기술(FLD)', '인공물(AFW)', '사건(EVT)'의 성능이 다른 개체명 범주 성능보다 떨어진다. 이는 세 개의 개체명 범주에서 다 어절 개체명이 많이 나타나 개체명으로 인식하지 못하는 오류가 많이 발생하기 때문이다. 이론 및 기술에서는 복잡한 기술명이나 학문명이, 인공물에서는 책 제목과 프로그램 제목 등이 다 어절 개체명으로 많이 나타난다. 그리고 사건에서는 역사적 사건 이름 등이 다 어절 개체명으로 많이 나타난다.

4.4 토의

표 10은 word embedding 자질 사용에 의해 나타난 정답과 오류 예이다. Word embedding 자질을 사용함으로써 기본 시스템에서 개체명으로 인식하지 못했던 형태소에 정답과 같은 개체명 태그를 부착할 수 있었다. 실험 결과인 표 4, 표 5와 표 6을 살펴보면 word embedding 자질을 사용하였을 때 정밀도(precision)에 비해 재현율(recall)의 성능이 많이 향상된 것을 볼 수 있다. 또한 기본 시스템에서 정답과 다른 개체명 태그가 부착 된 형태소에 대해서도 일정 부분 정답 개체명 태그를 부착할 수 있었다.

반면 기본 시스템에서 정답과 같은 개체명 태그를 부착하였지만 word embedding 자질을 사용함으로써 잘못된 개체명이 부착되거나 개체명으로 인식하지 못하는 오류가 나타나기도 했다. 또한 word embedding 자질을 추가로 사용하면서 개체명이 아닌 형태소에 개체명 태그를 부착하는 오류가 나타나는 경우도 있었다.

표 11은 다른 한국어 개체명 인식 및 분류 시스템들과 본 논문에서 제안하는 방법의 성능을 비교한 것이다. Structural SVM을 사용한 S-SVM3 방법과 FFNN, CNN을 사용한 방법보다 제안 방법이 더 높은 성능을 보여 그 효용성을 입증한다.

표 10 Word Embedding 자질 사용에 의해 나타난 정답과 오류 예

Table 10 Examples of Correct and Error using Word Embedding Features

	Morpheme	Gold	Baseline	Using Word Embedding Features
Correct	버즈	PER_B	O	PER_B
	제주	LOC_B	O	LOC_B
	해리포터	AFW_B	PER_B	AFW_B
Error	LA	LOC_B	LOC_B	ORG_B
	티아라	PER_B	PER_B	O
	나사	O	O	ORG_B

표 11 한국어 개체명 인식 및 분류 시스템 성능 비교 (TV 도메인)

Table 11 Comparative Results of each System (TV Domain)

System	F1 Score(%)
S-SVM3 (C. LEE, 2014)[3]	89.03
FFNN (C. LEE, 2014)	87.74
CNN (C. LEE, 2014)	88.57
Our System	89.81

3) ETRI의 TV 도메인 개체명 학습데이터 셋 중에서 인물/기관/지역 태그 셋을 사용하였다.

5. 결론 및 향후 과제

본 논문에서는 한국어 개체명 인식 및 분류에 word embedding 자질을 이용하는 방법을 제시하였다. Word embedding 자질을 이용하는 방법으로 형태소 단위의 word vector와 word cluster symbol을 CRFs의 자질로 사용하였다. TV 도메인과 Sports 도메인, IT 도메인에서 실험을 수행한 결과 기본 시스템 성능보다 각각 1.17%, 0.61%, 1.19%의 성능이 향상되었으며, 최신의 개체명 인식 및 분류 시스템보다도 높은 성능을 보여 그 효용성을 입증했다.

향후에는 word embedding을 형태소 단위가 아닌 개체명 단위로 수행하고, 그 word vector를 사용하여 개체명 인식 및 분류를 수행하는 실험을 수행할 것이다. 그리고 word embedding에 더 많은 학습 문서를 사용하면 개체명 인식 및 분류 성능 향상에도 도움이 될 것으로 기대된다.

References

[1] DM. Bikel, S. Miller, R. Schwartz, R. Weischedel, "Nymble: a High-Performance Learning Name-finder," *Proc. of the 5th Conference on Applied Natural Language Processing*, pp. 194-201, 1997.

[2] X. Liu, M. Zhou, F. Wei, Z. Fu and X. Zhou, "Joint Inference of Named Entity Recognition and Normalization for Tweets," *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 526-535, 2012.

[3] E. Chung, H. Lee, Y. Hwang and B. Yun, "Korean Name Entity Detection using Co-Training Methods," *Proc. of the Human Computer Interaction 2003*, pp. 1289-1293, 2003.

[4] C. Lee, et al., "Fine-Grained Named Entity Recognition using Conditional Random Fields for Question Answering," *Proc. of the 18th Annual Conference on Human & Cognitive Language Technology*, pp. 268-272, 2006.

[5] C. Lee and M. Jang, "Named Entity Recognition with Structural SVMs and Pegasos algorithm," *Journal of The Korean Society for Cognitive Science*, Vol. 21, No. 4, pp. 655-667. Dec. 2010.

[6] C. Lee, J. Kim, J. Kim and H. Kim, "Named Entity Recognition using Deep Learning," *Proc. of the 41th KIISE Winter Conference*, pp. 423-425, 2014.

[7] S. Bae and Y. Ko, "Automatic Construction of Class Hierarchies and Named Entity Dictionaries using Korean Wikipedia," *Journal of KIISE : Computing Practices and Letters*, Vol. 16, No. 4, pp. 492-496, Apr. 2010.

[8] Y. Song, S. Jeong and H. Kim, "A Constructing Method of Named Entity Dictionary using Wikipedia Based on Information Retrieval Method," *Proc. of the KIISE Korea Computer Congress 2015*, pp. 648-650, 2015.

[9] J. Turian, L. Ratinov and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 384-394, 2010.

[10] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, Vol. 3, pp. 1137-1155, 2003.

[11] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *ICLR Workshop*, 2013.

[12] J. Hong and J. Cha, "A New Korean Morphological Analyzer using Eojeol Pattern dictionary," *Proc. of the KIISE Korea Computer Congress 2008*, pp. 279-284, 2008.

최 윤 수

2014년 창원대학교 정보통신공학과 졸업(학사). 2016년 창원대학교 친환경해양플랜트FEED공학과정(컴퓨터·정보통신공학전공) 졸업(석사). 관심분야는 자연어처리, 기계학습, 딥 러닝

차 정 원

숭실대학교(학사). 포항공과대학교(석사, 박사). USC/ISI(박사후연수). 2004년~현재 창원대학교 컴퓨터공학과 교수. 관심분야는 자연어처리, 기계학습, 정보검색