공학석사학위논문

Word Embedding 자질을 이용한 한국어 개체명 인식

2015年 12月

창원대학교 대학원

친환경해양플랜트FEED공학과

최 윤 수

공학석사학위논문

Word Embedding 자질을 이용한 한국어 개체명 인식

Korean Named Entity Recognition Using Word Embedding Features

지도교수 차 정 원

이 논문을 공학석사학위논문으로 제출함.

2015年 12月

창원대학교 대학원

친환경해양플랜트FEED공학과

최 윤 수

최윤수의 석사학위 논문을 인준함.

심 사 위 원 장 이 종 근 ㊞

심 사 위 원 김 한 경 ㊞

심 사 위 원 차 정 원 ㊞

2015年 **12**月 日

창원대학교 대학원

목차

그림 목차

표 목차

제 Ⅰ 장

서론

개체명(Named Entity)이란 인명, 기관명, 지명 등과 같이 문서나 문장에서 특정한 의미를 가지고 있는 단어 또는 어구를 말한다[1].

- 인명: 최윤수, 박태호
- 기관명: 청와대, 창원대학교
- 지명: 부산광역시, 경상남도 창원시

정보 검색에서 개체명은 주요 검색 대상이 된다. 이러한 개체명을 추출하기 위해 자연어 처리 분야에서 개체명 인식(Named Entity Recognition)에 대한 연구가 발전했다.

개체명은 다음과 같은 특징들을 가진다. 첫 번째로 개체명의 대부분은 고유명사로서 미등록어인 경우가 많고 신조어와 같이 계속해서 생성되거나 삭제되는 경우가 많다. 두 번째로 개체명은 애매성을 가진다. 애매성이란 같은 단어라도 문맥에 따라 다른 개체명을 가지는 것이다. 아래 예문을 살펴보면 '창원대학교'가 첫 번째 문장에서는 기관명을 의미하지만, 두 번째 문장에서는 지명을 의미한다.

- 최윤수는 2014년 3월 3일 **창원대학교**에 입학했다.
- 윤수와 태호는 **창원대학교** 앞에서 만나기로 약속했다.

이와 같은 개체명의 특징들 때문에 사전을 구축하여 개체명을 인식하고 의미를 파악하는 것은 어렵다. 따라서 개체명 인식에 대한 연구가 더욱 필요하다는 것을 알 수 있다.

개체명 인식에 관한 연구는 영어권에서 먼저 발전하였다[2-6]. 영어권에서는 개체명 인식을 위해 대문자 자질 등 영어에서 나타나는 언어 특징을 이용하여 높은 개체명 인식 성능을 보였다. 한국어에서도 개체명 인식에 대한 다양한 연구가 있었다[7-11]. 하지만 한국어는 영어에서 나타나는 대문자와 같은 특정 자질(feature)의 부재로 개체명을 인식하기 어려운 점이 있다.

한편 자연어 처리 분야에서 word embedding 자질을 이용하는 연구가 진행되고 있다[12,13]. 최근 새롭게 제안된 word embedding 방법인 CBOW(Continuous Bag-of-Words) 언어 모델은 기존의 word embedding 방법보다 높은 성능을 보인다[14].

본 논문에서는 한국어 개체명 인식에서 자질 부족 문제를 보완하기 위해 word

embedding 자질을 사용하는 방법을 제안한다. CBOW 모델과 K-means를 이용하여 각각 형태소 단위의 word vector와 군집 정보를 생성하고, 이를 개체명 인식을 위한 자질로 사용하였다. 실험 결과 word embedding 자질을 개체명 인식에 사용할 경우 의미 있는 성능 향상이 있었다.

본 논문의 구성은 다음과 같다. II장에서는 관련 연구로서 영어권과 한국어에서의 이전 개체명 인식 방법과 언어 모델 및 word embedding에 대하여 기술한다. III장에서는 word embedding 자질을 한국어 개체명 인식에 적용한 방법에 대하여 설명한다. IV장에서는 제안된 방법을 이용한 다양한 실험에 대해 기술하고 실험결과에 대한 분석을 한다. 마지막으로 V장에서 결론을 도출하고 향후 과제를 기술한다.

제 Ⅱ 장

관련 연구

본 장에서는 영어권에서의 개체명 인식 시스템과 한국어에서의 개체명 인식 시스템 관련 연구, 그리고 언어 모델에 대한 관련 연구를 살펴본다.

1. 영어권에서의 개체명 인식

개체명 인식에 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 부착 하였다[2]. 이 연구에서는 대문자나 호칭 기호 등 영어에서 나타나는 문자의 특징을 자질로 사용하여 93%의 높은 성능을 보였다. 또한 HMM 외에 다양한 지도 학습 방법이 개체명 인식에 사용되었다[3,4]. [3]과 [4]는 벵골어에서 각각 CRFs(Conditional Random Fields)와 SVMs(Supports Vector Machines)을 이용하여 개체명 인식을 실험하였다. 인명, 지역, 단체, 숫자, 비개체명으로 총 5개의 범주에 대하여 개체명을 부착하였다. 사용한 자질로는 주변 단어 정보, 형태소 분석 결과, 접미사, 접두사, 단어 길이, 단어의 첫 글자 등을 사용하였다. 개체명 인식 실험 결과 [3]은 90.7%의 성능을 보였고, [4]는 91.8%의 성능을 보였다.

최근에는 트위터 글을 분석하여 개체명을 인식하는 실험이 있었다[5,6]. 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. [5]는 트위터 글을 학습하여 개체명 인식 실험을 수행하여 83.6%의 개체명 인식 성능을 보였다.

2. 한국어에서의 개체명 인식

한국어에서의 개체명 인식에 대해서는 다음과 같은 연구가 있었다. 개체명 인식을 위한 학습 중 반지도 학습인 Co-Training 기법을 변형한 규칙 기반의 방식이 있었다[7]. 그리고 지도학습 방법으로 CRFs(Conditional Random Fields)와 최대 엔트로피 모델(Maximum Entropy Model)을 이용하는 방법이 있었다[8]. 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[9]. 이 방법은 CRFs를 이용한 방법[8]보다 높은 성능을 유지하면서 학습 시간은 4% 줄일 수 있었다.

다른 방법으로 개체명 인식을 위해 개체명 사전을 이용하는 방법이 있다[10]. 개체명 인식 성능 향상을 위해 위키피디아를 이용하여 개체명 사전을 구축하고 확장하는 방법이다.

최근에는 딥 러닝을 이용한 개체명 인식[11] 또한 연구 되었는데 영어에 비해 자질이 부족한 한국어에 자질 튜닝 작업에 들어가는 시간과 노력을 줄이면서 기존의 개체명 인식기 성능과 큰 차이가 없음을 보였다.

하지만 앞서 언급된 한국어 개체명 인식에 대한 방법들 모두 영어권에 비해 낮은 성능을 보인다. 이는 영어에서 나타나는 대문자 자질 등의 부재 때문이다. 따라서 본 연구에서는 한국어 개체명 인식의 자질 부족 문제를 보완하기 위해 word embedding 자질을 한국어 개체명 인식에 이용하는 방법을 제안한다.

3. 언어 모델(Language Model)

언어 모델(Language Model)은 문장을 이루는 단어들의 확률분포로서 음성 인식, 기계 번역, 형태소 분석 등의 분야에서 매우 중요한 정보로 사용되고 있다.

Word embedding 이란 언어 모델의 하나로서 문장 속의 단어들 사이의 관계를 비지도 학습(Unsupervised learning)방식으로 분석하여 특징화 하는 것이다. 최근에는 다양한 word embedding 방법을 이용하여 영어의 chunking 과 개체명 인식을 수행하고 각각의 성능을 비교하는 연구가 있었다[12]. 또한 인공 신경망을 이용하는 NNLM(Neural Network Language Model)은 뛰어난 성능을 나타내어 많은 많은 연구에 참고되었다[13].

최근 새로운 word embedding 방법으로 CBOW(Continuous bag-of-Words) 모델이 제안되었다[14]. CBOW 모델은 현재 word 의 문맥을 이루는 vector 들의 합으로 그 word 의 vector 를 결정하는 모델이다. NNLM 의 구조를 변경해 은닉층(Hidden Layer) 대신 투영층(Projection Layer)을 사용함으로써 학습시간을 100 배 이상 단축시켰다. 또한 NNLM 보다 의미 정확도는 1%, 구문 정확도는 11% 높은 성능을 보였다[14].

제 Ⅲ 장

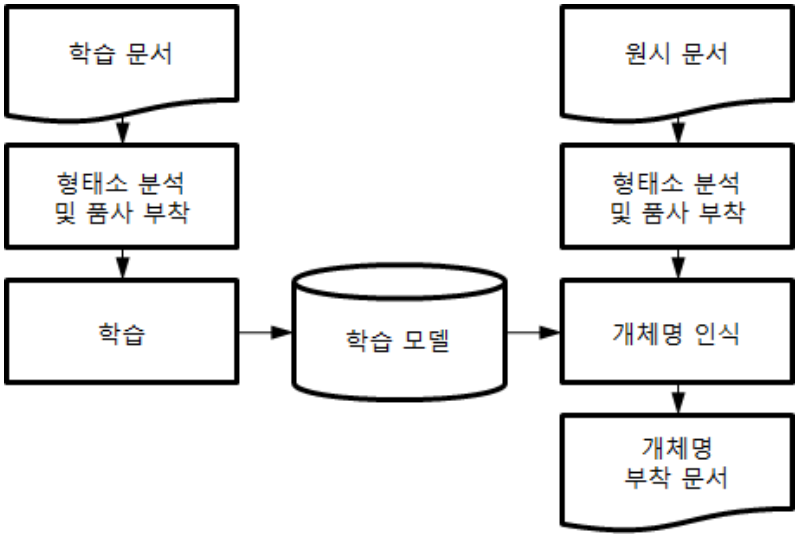
Word Embedding 자질을 이용한 한국어 개체명 인식

한국어 개체명 인식은 영어에서 나타나는 대문자 자질 등의 부재로 개체명을 인식하는데 어려운 점이 있다. 본 연구에서는 word embedding 자질(feature)을 한국어 개체명 인식에 사용하여, 자질 부족 문제를 보완하고 성능이 향상된 한국어 개체명 인식 시스템을 만드는 것을 목표로 한다.

본 장에서는 전체적인 개체명 인식 시스템에 대한 설명과 word embedding 자질을 생성하는 방법에 대하여 설명하고, 학습 및 모델 생성 과정 그리고 개체명 인식 과정에 대하여 설명한다.

1. 개체명 인식 시스템

본 논문에서 제안하는 개체명 인식 시스템의 전체 구조는 <그림 III-1> 과 같다. 시스템은 크게 학습 과정과 개체명 인식 과정으로 나뉜다. 학습 과정은 학습 문서로부터 학습 모델을 생성하는 과정이다. 개체명 인식 과정은 개체명이 부착되지 않은 원시 문서에서 학습 모델을 이용해 개체명을 인식하고 개체명을 부착한 문서를 출력하는 과정이다.



<그림 III-1> 개체명 인식 시스템의 전체 구조도

본 논문에서 제안한 시스템에서 학습 및 학습 모델 생성을 위해, 통계적 기계학습 방법 중 하나인 CRFs(Conditional Random Fields)를 이용한다. CRFs는 조건부 확률을 최대로 하는 비방향성 그래프 모델이다[15,16]. CRFs는 HMMs(Hidden Markov Models)

에 비하여 변수 독립성 조건이 필요 없으며, MEMMs(Maximum Entropy Markov Models)에 비하여 label bias 문제가 없는 장점이 있다[15,16].

개체명 인식 시스템은 사용자가 인식하고자 하는 범주를 결정한다. 본 시스템에서는 <표 III-1>과 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 사용한다.

<표 III-1> 개체명 범주 및 정의

	개체명 범주	태그	정 의
1	PERSON	PER	실존 일물과 가상의 인물(캐릭터, 신화 속 인물)
2	FIELD	FLD	학문 분야 및 이론, 법칙, 기술 등
3	ARTIFACTS_WORKS	AFW	인공물로 사람에 의해 창조된 대상물
4	ORGANIZATION	ORG	기관 및 단체와 회의/회담을 모두 포함
5	LOCATION	LOC	지역명칭과 행정구역 명칭 등
6	CIVILIZATION	CVL	문명 및 문화에 관련된 용어
7	DATE	DAT	날짜
8	TIME	TIM	시간
9	NUMBER	NUM	숫자
10	EVENT	EVT	특정 사건 및 사고의 명칭과 행사 등
11	ANIMAL	ANM	동물
12	PLANT	PLT	식물
13	MATERIAL	MAT	금속, 암석, 화학물질 등
14	TERM	TRM	의학 용어, IT 관련 용어 등의 일반 용어를 총칭

2. 형태소 분석 및 품사 부착

본 연구에서는 형태소 단위로 개체명을 인식한다. 따라서 학습 과정과 개체명 인식 과정을 위해 문서의 형태소 분석 및 품사 부착 과정이 필요하다. 형태소란 언어학에서 의미를 가지는 가장 작은 말의 단위를 나타낸다. 형태소 분석이란 문장을 최소한의 형태소 단위로 분리하는 것을 말하며, 형태소 품사 부착이란 형태소 분석 결과에 형태소가 지니는 구문 기능에 따라 일정한 주석을 부착하는 작업이다. <그림 III-2>는 형태소 분석 및 품사 부착의 예이다.

문장: 최윤수는 2014년 3월 3일 창원대학교에 입학했다.	
최윤수는	최윤수/NNP + 는/JX
2014년	2014/SN + 년/NNB
3월	3/SN + 월/NNB
3일	3/SN + 일/NNB
창원대학교에	창원대학교/NNP + 에/JKB
입학했다.	입학/NNG + 하/XSV + 았/EP + 다/EF + ./SF

<그림 III-2> 형태소 분석 및 품사 부착 예

<그림 III-2>에서 ‘NNP’는 고유명사, ‘NNG’는 일반명사, ‘NNB’는 의존명사를 의미한다. ‘JX’는 보조사, ‘JKB’는 부사격 조사를 의미한다. ‘XSV’는 동사파생 접미사, ‘EP’는 선어말 어미, ‘EF’는 종결 어미, ‘SF’는 종결 기호를 의미한다. 형태소 분석 및 품

사 부착 결과로부터 어휘 정보, 품사 정보, 형태소 길이 정보 등 개체명 인식을 위한 특징 정보를 얻을 수 있다.

형태소 단위로 개체명을 인식할 경우 개체명의 형태소 경계를 구분해야 하는 문제가 발생한다. 본 시스템에서는 형태소의 경계를 표현하기 위해 <표 III-1>의 개체명 범주 태그에 B/I/O 형태를 결합한 개체명 태그를 사용하였다. B/I/O 형태는 개체명의 시작(Begin), 개체명의 중간 혹은 마지막(Inside), 개체명이 아닌 것(Outside)로 구성된다. <표 III-2>에서 형태소 분석 및 품사 부착이 된 문장에 B/I/O 형태의 개체명 태그 부착 예를 보여준다. ‘PER_B’, ‘DAT_B’, ‘ORG_B’는 각각 인명, 날짜, 기관 개체명의 시작 형태소를 의미한다. ‘DAT_I’는 날짜 개체명의 중간 또는 끝 형태소 의미하며, ‘O’는 개체명이 아닌 형태소를 의미한다.

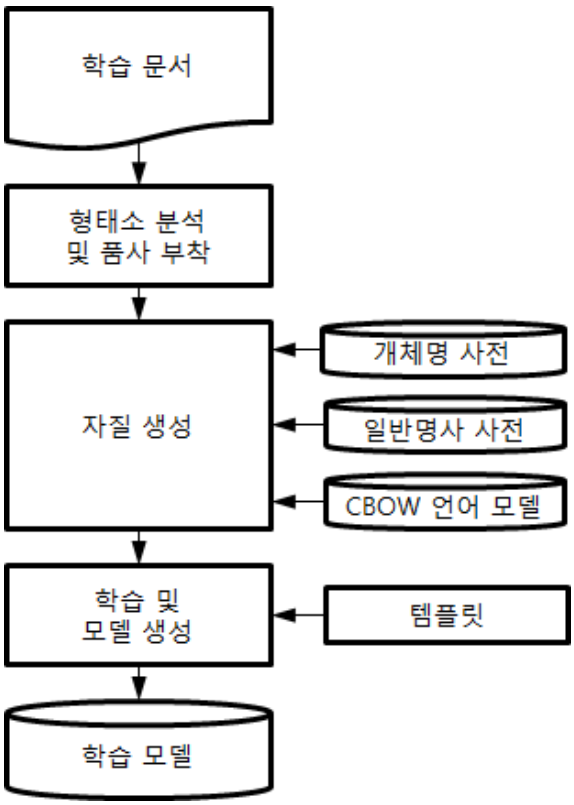
<표 III-2> B/I/O 형태의 개체명 태그 부착 예

형태소	형태소 품사 태그	개체명 태그
최윤수	NNP	PER_B
는	JX	O
2014	SN	DAT_B
년	NNB	DAT_I
3	SN	DAT_I
월	NNB	DAT_I
3	SN	DAT_I
일	NNB	DAT_I
창원대학교	NNP	ORG_B
에	JKB	O
입학	NNG	O
하	XSV	O
았	EP	O
다	EF	O
.	SF	O

3. 학습 과정

<그림 III-3>는 제안 시스템의 학습 과정 구조도이다. 형태소 분석 및 품사 부착이 완료된 학습 문서에서 형태소 정보와 개체명 사전, 일반명사 사전 그리고 CBOW 언어 모델로부터 자질(feature)을 생성한다.

자질 생성이 끝나면 템플릿으로부터 원하는 자질을 선택하여 CRFs로 학습을 수행한다. CRFs의 학습이 끝나면 학습 모델이 생성된다.



<그림 III-3> 제안 시스템의 학습 과정 구조도

3.1 개체명 사전과 일반명사 사전

학습 문서의 형태소 분석 및 품사 부착 작업이 끝나면 형태소의 개체명 사전 존재 유무에 대해 검색한다. 개체명 사전은 <표 III-1>의 14 개 범주에 각각 해당하는 개체명을 모아 생성한 사전이다. 해당 범주의 개체명 사전에 존재 여부만으로도 개체명을 인식하는데 큰 도움이 된다.

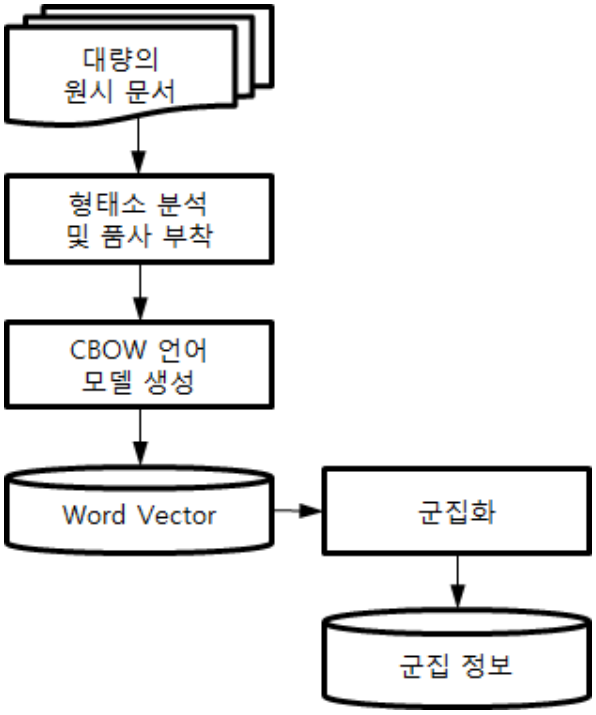
하지만 개체명은 계속해서 새로운 개체명이 생성되고 기존 개체명이 삭제되기 때문에 개체명 사전만으로는 개체명을 인식하는데 어려움이 있다. 본 시스템에서는 이를 보완하기 위해 일반명사 사전을 이용한다. 일반명사 사전이란 우리가 가지고 있는 명사를 모두 모으고, 이 중 개체명이 되는 명사를 제외하고 남은 명사로 생성한 사전이다. 일반명사 사전에 존재한다는 것은 개체명이 아닐 가능성이 크다는 뜻이며, 일반명사 사전에 존재하지 않는다는 것은 개체명일 가능성이 높다라고 해석할 수 있으므로 개체명을 인식하는데 큰 도움이 된다. 실제 본 연구자의 이전 실험에서 일반명사 사전을 개체명 인식에 사용할 경우 약 0.5%의 성능 향상이 있었다.

3.2 CBOW 언어 모델

본 연구에서는 CBOW 언어 모델을 이용하여 word embedding 을 수행하였다. 영어에서는 가공되지 않은 대량의 원시 문서를 그대로 입력하여, word 단위의 word

embedding 을 수행하고 word vector 를 생성한다. 본 연구에서는 한국어에 맞춰 대량의 원시 문서를 형태소 분석 및 품사 부착 단계를 거치고, 이를 입력으로 하여 형태소 및 품사 단위의 word embedding 을 수행하였다.

<그림 III-4>는 CBOW 언어 모델 생성 과정이다. 우선 가공되지 않은 대량의 원시 문서에 형태소 분석 및 품사 부착 과정을 거친다. 형태소 분석 및 품사 부착이 끝나면 <그림 III-5>와 같이 형태소 분석 및 품사 부착된 문서를 형태소 단위로 분리하여 순서대로 나열한다.

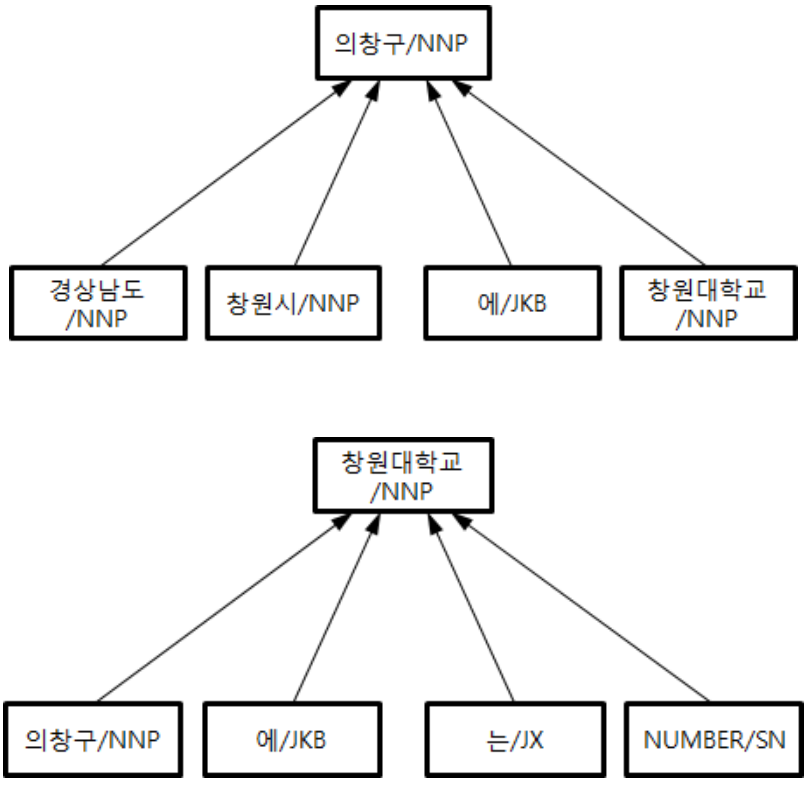


<그림 III-4> CBOW 언어 모델의 생성 과정

원시 문서: 경상남도 창원시 의창구에 창원대학교는 1969년에 개교하였다. …
경상남도/NNP 창원시/NNP 의창구/NNP 에/JKB 창원대학교/NNP 는/JX 1969/SN 년/NNB 에/JKB 개교/NNG 하/XSV 았/EP 다/EF ./SF …

<그림 III-5> 원시 문서의 형태소 분석 및 품사 부착

형태소를 나열하고 난 뒤 첫 번째 형태소부터 차례로 word embedding을 수행한다. 형태소 단위로 word embedding을 수행하므로 <그림 III-6>와 같이 앞, 뒤 주변 형태소로부터 자기 자신의 word vector 값을 결정한다.



<그림 III-6> CBOW 언어 모델을 이용한 형태소 단위의 word embedding

CBOW 언어 모델을 이용하여 word embedding을 수행하면 형태소 단위로 실수 값으로 이루어진 word vector가 생성된다. 본 시스템에서는 word vector를 50차원의 실수로 생성하고, 이 실수 값을 학습과 개체명 인식 과정을 위한 word embedding 자질로 사용한다. <표 III-3>은 형태소 단위로 생성된 word vector의 예이다.

<표 III-3> 형태소 단위로 생성된 word vector의 예

형태소	Word Vector
경상남도/NNP	0.054519, -0.039076, ..., 0.049407
창원시/NNP	0.076326, -0.113740, ..., 0.066791
의창구/NNP	-0.097192, -0.085535, ..., 0.291513
에/JKB	0.214002, -0.124554, ..., 0.146882
창원대학교/NNP	-0.039561, 0.006097, ..., -0.145819
는/JX	0.175349, -0.134337, ..., 0.111820
NUMBER/SN	0.271558, -0.068706, ..., 0.254129
...	...

Word vector 생성이 완료되면 이 vector 값을 이용하여 군집화(Clustering)를 수행할 수 있다. 실수로 이루어진 word vector 값을 입력으로 하고, 군집할 단위 개수 K 값을 정한 후 K-means 알고리즘을 이용하여 군집화를 수행한다. K-means 는 중심을 선택하고 군집화를 수행한 후 군집화가 정상적으로 이루어졌는지 검정하고 이상적인 군집화가 이루어지거나 일정 횟수에 도달할 때까지 중심을 갱신하고 군집화를 수행하는 알고리즘이다.

군집화를 수행하면 형태소 단위의 군집 정보(Cluster Symbol)를 생성할 수 있다. 이 군집 정보를 이전에 생성한 word vector 와 함께 개체명 인식을 위한 word embedding 자질로써 사용한다. <표 III-4>는 형태소 단위의 word vector 를 200 개, 300 개, 400 개,

500 개로 군집화한 군집 정보의 예이다. <표 III-4>를 보면 이상적인 군집화를 수행하는 과정에서 군집 개수에 따라 군집 정보가 달라지는 것을 알 수 있다.

<표 III-4> 형태소 word vector의 군집 정보의 예

형태소	Word vector 군집 정보(Cluster Symbol)			
	200 개	300 개	400 개	500 개
경상남도/NNP	25	155	376	355
창원시/NNP	3	130	101	479
의창구/NNP	90	155	261	355
에/JKB	189	167	383	92
창원대학교/NNP	79	132	242	480
는/JX	93	32	383	92
NUMBER/SN	113	13	29	183
...	...	...	...	...

3.3 자질 생성(Feature Generation)

자질은 문장을 이루는 형태소 단위로 생성한다. 형태소 분석 및 품사 부착 정보로부터 기본자질을 생성하고 개체명 사전, 일반명사 사전 그리고 CBOW 언어 모델을 이용하여 자질을 추가한다.

<표 III-5>은 자질 생성 예제이다. 자질 1 은 형태소의 어휘, 2 는 형태소의 품사 태그, 3 은 형태소의 길이이다. 자질 4 는 형태소의 어절 내 위치로 '0'은 어절 내 첫 번째 형태소, '1'은 어절 내 중간 형태소, '2'는 어절 내 마지막 형태소로 총 3 가지

값을 가진다. 자질 5 는 현재 어절의 마지막 형태소가 조사일 경우 그 조사의 형태소와 품사 태그를 사용하고, 조사가 아닐 경우 '-'를 사용한다. 자질 6 은 <표 III-1>의 개체명 범주 순서로, 14 개 개체명 사전 내에 형태소의 존재 여부이다. 자질 7 은 일반명사 사전 내에 존재 여부이다. 자질 6 과 자질 7 은 존재할 경우 '1', 존재하지 않을 경우 '0'의 값을 가진다. 자질 8 은 3.2 장에서 설명한 것과 같이 대량의 원시 문서로 생성한 CBOW 언어 모델에서, 현재 형태소의 word vector 또는 군집 정보를 word embedding(W/E) 자질로 사용한다.

<표 III-5> 자질 생성 예제

자질번호 / 형태소	1	2	3	4	5	6	7	8
1	최윤수	NNP	3	0	는/JX	00000000000000	0	W/E
2	는	JX	1	2	는/JX	00000000000000	1	W/E
3	2014	SN	4	0	-	00000000000000	1	W/E
4	년	NNB	1	2	-	00000000000000	1	W/E
5	3	SN	1	0	-	00000000000000	1	W/E
6	월	NNB	1	2	-	00000000000000	1	W/E
7	3	SN	1	0	-	00000000000000	1	W/E
8	일	NNB	1	2	-	00000000000000	1	W/E
9	**창원대학교**	**NNP**	**5**	**0**	**에/JKB**	**0001100000000**	**0**	**W/E**
10	에	JKB	1	2	에/JKB	00000000000000	1	W/E
11	입학	NNG	2	0	-	00000000000000	1	W/E
12	하	XSV	1	1	-	00000000000000	1	W/E
13	았	EP	1	1	-	00000000000000	1	W/E
14	다	EF	1	1	-	00000000000000	1	W/E
15	.	SF	1	2	-	00000000000000	1	W/E

<표 III-5>에서 아홉 번째 형태소인 ‘창원대학교’의 경우 어휘는 ‘창원대학교’이고, 품사는 ‘NNP’이고, 형태소의 길이는 5 이다. 어절 내 첫 번째 형태소로 어절 내 위치는 ‘0’이고, 어절의 마지막 형태소가 조사이므로 ‘에/JKB’이고, 개체명 사전에는 ‘기관 개체명 사전’과 ‘지역 개체명 사전’에 존재 하므로 ‘00011000000000’, 일반명사 사전에 존재 하지 않음으로 ‘0’이 자질로써 생성된다. 그리고 CBOW 언어 모델로부터 현재 형태소의 50 차원의 실수로 이루어진 word vector 값 또는 군집 정보가 word embedding 자질로 생성된다.

3.4 학습 및 모델 생성

자질을 모두 생성하면 학습 및 모델 생성을 수행한다. 학습에 사용할 자질은 템플릿으로 결정한다. <표 III-6>은 템플릿 사용 자질 예제이다. 대괄호 안에서 앞의 숫자는 현재 입력 형태소로부터의 거리를 뜻하며, ‘-’는 이전 형태소, ‘+’는 다음 형태소를 의미한다. 대괄호 안에서 뒤의 숫자는 <표 III-5>의 자질 번호를 뜻한다. ‘&’ 기호는 두 개 이상의 자질을 합친 조합 자질을 의미한다. 템플릿 번호 1 번은 현재 형태소의 첫 번째 자질인 형태소 어휘를 뜻한다. 2 번은 현재 형태소의 두 번째 자질인 형태소 품사를 뜻하고, 3 번은 현재 형태소의 여섯번째 자질인 개체명 사전 존재 여부를 뜻한다. 4 번은 이전 형태소의 어휘와 현재 형태소의 품사를 조합한 자질을 의미한다. 5 번은 다음 형태소의 품사와 현재 형태소의 어휘를 조합한 자질을 을 의미한다.

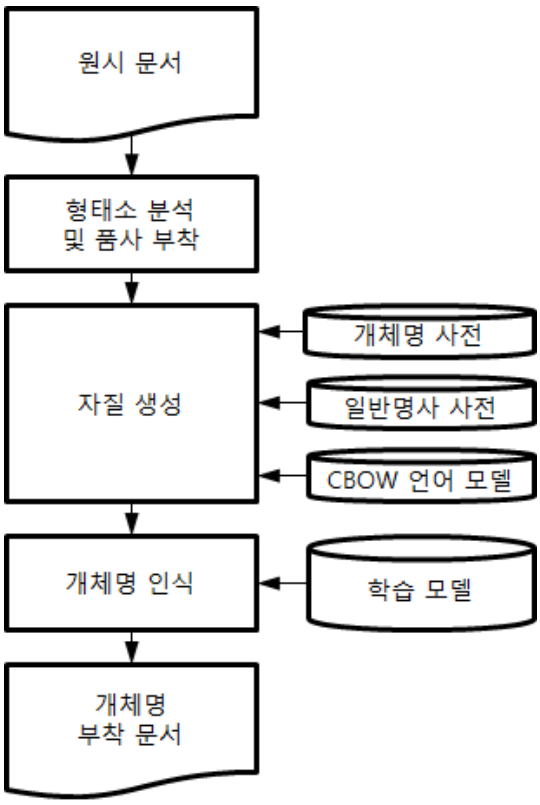
<표 III-6> 템플릿 사용 자질 예제

템플릿 번호	선택 자질
1	[0, 1]
2	[0, 2]
3	[0, 6]
4	[-1, 1] & [0, 2]
5	[+1, 2] & [0, 1]

예를 들어 <표 III-5>에서 아홉 번째 형태소인 ‘창원대학교’는 템플릿 번호 1번에 의해 현재 형태소의 어휘인 ‘창원대학교’ 자질, 2번에 의해 현재 형태소의 품사인 ‘NNP’ 자질이 사용된다. 그리고 템플릿 번호 3번에 의해 개체명 사전 존재 여부로써 ‘0001100000000’ 자질이 사용된다. 또한 템플릿 번호 4번에 의해 이전 형태소의 어휘인 ‘일’과 현재 형태소의 품사인 ‘NNP’를 조합한 조합 자질을 사용하고, 5번에 의해 다음 형태소의 품사인 ‘JKB’와 현재 형태소의 어휘인 ‘창원대학교’를 조합한 조합 자질을 사용한다.

4. 개체명 인식 과정

<그림 III-7>은 개체명 인식 과정 구조도이다. 개체명 인식 과정은 개체명이 부착되지 않은 원시 문서가 입력되면 형태소 분석 및 품사 부착을 수행하고, 학습 과정과 동일하게 자질을 생성한다. 생성된 자질과 학습 과정에서 생성된 모델을 이용하여 개체명을 인식하고 개체명을 부착한 후 결과 문서를 출력한다.



<그림 III-7> 개체명 인식 과정 구조도

아래는 원시 문서에 대한 개체명 인식 결과의 예제이다. 예제에서 보는 것과 같이 ‘박태호’는 인명(PER)으로, ‘창원대학교’는 기관명(ORG)으로, ‘학생’은 문명 및 문화 관련 명칭(CVL)으로 개체명이 인식이 된 것을 알 수 있다.

- <**박태호**:PER>는 <**창원대학교**:ORG>에 다니는 <**학생**:CVL>이다.

제 IV 장

실험 및 토의

1. 실험 환경

본 논문에서 제안된 방법의 효용성을 보이기 위해 다양한 실험을 진행하였다. 한국어 개체명 인식 시스템의 성능을 측정하기 위해서 TV 도메인과 스포츠 도메인, IT 도메인 문서를 사용하였다. 도메인의 문서는 각각 개체명 모델을 생성하기 위한 학습 데이터와 성능 평가를 위한 평가 데이터로 나누어 사용하였다. TV 도메인에서는 104,759문장을 학습 데이터로 사용하고 3,896문장을 평가 데이터로 사용하였다. 스포츠 도메인에서는 42,809문장을 학습 데이터로 사용하고 4,000문장을 평가 데이터로

사용하였다. 마지막으로 IT 도메인에서는 14,075문장을 학습 데이터로 사용하고 1,000문장을 평가 데이터로 사용하였다.

Word embedding 자질을 사용하였을 때 개체명 인식 성능을 알아보기 위해 word embedding 자질을 사용하지 않은 시스템 성능과 word vector 자질과 군집 정보 자질을 추가로 사용한 개체명 인식 성능을 비교 분석하였다. 그리고 기본 시스템에 군집 정보 자질과 word vector 자질을 모두 사용하였을 때의 성능을 비교 분석 하였다.

CBOW 언어 모델을 생성하기 위해 대량의 원시 문서 내에 약 2억 8천만 개의 형태소를 사용하여 word embedding을 수행하고, 50차원의 실수로 이루어진 569,589개의 형태소 단위 word vector를 생성하였다.

개체명 인식을 위한 형태소 분석과 word embedding을 위한 형태소 분석은 창원대학교 적응지능연구실에서 공개한 Espresso[17]를 사용하여 수행하였다. 또한 CRFs를 이용한 학습 및 평가를 위해 CRF++을 사용하였다.

제안한 시스템의 성능 평가를 위해 정밀도와 재현율을 결합한 $F_1 - measure$를 사용하였다. 평가 척도는 식 (1)과 같다.

$$\text{정밀도}(Precision, P) = \frac{\text{실제 정답의 수}}{\text{시스템이 출력한 수}}$$

$$\text{재현율}(Recall, R) = \frac{\text{실제 정답의 수}}{\text{정답문서의 모든 정답의 수}} \qquad (1)$$

$$F_1 - measure = \frac{2 \cdot P \cdot R}{P + R}$$

2. 실험 결과

2.1 기본 시스템 성능

Word embeddgin 자질을 사용하였을 때의 성능 변화를 알기 위해, word embedding 자질을 제외하고 기본 자질만을 사용하여 실험을 수행하였다. <표 IV-1>는 Word embedding 자질을 사용하지 않은 기본 시스템 성능이다. TV 도메인에서는 88.51%, Sports 도메인에서는 89.45%, IT 도메인에서는 80.50%의 성능을 보였다.

<표 IV-1> 한국어 개체명 인식 기본 시스템 성능

도메인	$Precision(\%)$	$Recall(\%)$	$F_1 - measure(\%)$
TV 도메인	89.15	87.88	88.51
Sports 도메인	90.42	88.51	89.45
IT 도메인	82.78	78.34	80.50

2.2 Word Vector 자질을 사용한 성능 평가

다음으로 기본 시스템에 word vector 자질[1]을 추가로 사용하고 실험을 수행하였다. <표 IV-2>는 세 가지 도메인에서 word vector 자질을 추가로 사용하였을 때의 성능이다. <표 IV-1>의 기본 시스템의 성능과 비교 하였을 때, TV 도메인에서는 88.91%로 0.4% 향상된 성능을 보였다. Sports 도메인은 89.92%로 0.47% 향상된 성능을 보였다. IT 도메인에서는 80.51%로 0.01% 향상된 성능을 보였다.

<표 IV-2> Word vector 자질을 사용하였을 때 개체명 인식 성능

도메인	$Precision(\%)$	$Recall(\%)$	$F_1 - measure(\%)$	성능 변화(%)
TV 도메인	89.24	88.59	88.91	+0.4
Sports 도메인	91.00	88.85	89.92	+0.47
IT 도메인	82.86	78.30	80.51	+0.01

이 실험을 통해 word vector 자질을 추가로 사용할 경우 형태소 정보로부터 얻은 자질과 사전으로부터 얻은 자질의 부족한 부분을 보완하여 개체명 인식 성능을 향상시킬 수 있음을 알 수 있다. 하지만 IT 도메인에서는 0.01% 성능 향상에 그쳐 word vector 자질 외에 추가로 다른 word embedding 자질을 사용할 필요성이 있음을 알 수 있다.

[1] Word vector는 실수 값으로 소수점 여섯째 자리에서 반올림 하여 소수점 다섯째 자리로 사용하였다.

2.3 군집 정보 자질을 사용한 성능 평가

두 번째로 기본 시스템에 군집 정보 자질을 추가로 사용하여 실험을 수행하였다. 군집 정보 자질은 앞서 생성한 word vector를 K-means 알고리즘을 이용하여 군집화 하였다. K-means 알고리즘은 군집화 하고자 하는 군집 개수를 지정한다. 본 실험에서는 200, 300, 400, 500개로 군집화 하여 각각 군집 정보 자질을 생성하였다. <표 IV-3>, <표 IV-4>, <표 IV-5>는 각각 TV 도메인과 Sports 도메인, IT 도메인의 군집 개수 별 성능을 보여준다. 모든 도메인에서 기본 시스템에 군집 정보 자질을 추가로 사용하였을 때 그 성능이 향상 되었다. 군집 정보 자질을 생성할 때 사용한 군집 개수 별 성능에서 TV 도메인에서는 300개, Sports 도메인은 200개, IT 도메인은 400개에서 성능이 가장 높았다. <표 IV-1>의 기본 시스템의 성능과 비교 하였을 때, TV 도메인에서는 88.74%로 0.23% 성능이 향상되었고, Sports 도메인에서는 89.93%로 0.48% 성능이 향상 되었다. IT 도메인에서는 81.32%로 0.82% 성능이 향상되었다.

<표 IV-3> 군집 정보 자질을 용하였을 때 개체명 인식 성능(TV 도메인)

군집 개수	$Precision(\%)$	$Recall(\%)$	$F_1 - measure(\%)$	성능 변화(%)
200개	89.08	88.01	88.54	+0.03
300개	89.19	88.29	**88.74**	+0.23
400개	89.24	87.86	88.54	+0.03
500개	89.19	88.16	88.67	+0.16

<표 IV-4> 군집 정보 자질을 사용하였을 때 개체명 인식 성능(Sports 도메인)

군집 개수	$Precision(\%)$	$Recall(\%)$	$F_1 - measure(\%)$	성능 변화(%)
200개	90.97	88.92	**89.93**	+0.48
300개	90.96	88.85	89.90	+0.45
400개	90.81	88.77	89.78	+0.33
500개	90.54	88.56	89.54	+0.09

<표 IV-5> 군집 정보 자질을 사용하였을 때 개체명 인식 성능(IT 도메인)

군집 개수	$Precision(\%)$	$Recall(\%)$	$F_1 - measure(\%)$	성능 변화(%)
200개	83.33	79.16	81.19	+0.69
300개	83.18	78.79	80.92	+0.42
400개	83.39	79.36	**81.32**	+0.82
500개	83.48	79.18	81.27	+0.77

<표 IV-3>, <표 IV-4>, <표 IV-5>에서 군집 개수에 따른 성능 변화의 추세가 일정하지 않고, 도메인에 따라 가장 좋은 성능을 보이는 군집 정보가 다르다는 것을 알 수 있다. 이는 도메인에 관계없이 대량의 원시 문서를 사용하여 word embedding을 수행하고 CBOW 언어 모델을 생성하였지만, 각 도메인에서 나타나는 형태소가 조금씩 다르기 때문이다. 또한 군집 개수에 따라 군집 정보가 달라지면서, 같은 개체명 범주를 가지는 형태소들이 군집 될 수도 있고 군집되지 않을 수도 있기 때문이다. 예를 들어 <표 III-4>에서 '경상남도/NNP'와 '의창구/NNP'는 모두 지명을 의미하는 개체명

이다. 그런데 300개와 500개로 군집화 하였을 때는 같은 군집 정보를 가지지만 200개와 400개로 군집화 하였을 때는 다른 군집 정보를 가지게 된다. 따라서 도메인에 따라 가장 높은 성능을 보이는 군집 개수를 찾기 위해, 다양한 실험을 수행할 필요성이 있다는 것을 알 수 있다.

2.4 Word Embedding 자질을 모두 사용한 성능 평가

마지막으로 word vector 자질과 군집 정보 자질을 모두 사용하여 실험을 수행하였다. 각 도메인에서 군집 정보 자질을 사용했을 때 가장 성능이 좋은 군집 개수 자질을 word vector 자질과 함께 사용하였다. TV 도메인은 300개, Sports 도메인은 200개, IT 도메인은 400개로 군집화하여 생성한 군집 정보 자질을 사용하였다. <표 IV-6>은 word embedding 자질을 모두 사용하였을 때 개체명 인식 성능이다. <표 IV-1>의 기본 시스템의 성능과 비교 하였을 때, TV 도메인 에서는 89.03%로 0.52% 성능이 향상 되었다. Sports 도메인에서는 89.98%로 0.53% 성능이 향상되었으며, IT 도메인은 80.69%로 0.19% 성능이 향상되었다.

<표 IV-6> Word embedding 자질을 모두 사용하였을 때 개체명 인식 성능

도메인	$Precision(\%)$	$Recall(\%)$	$F_1 - measure(\%)$	성능 변화(%)
TV 도메인	89.33	88.73	89.03	+0.52
Sports 도메인	91.10	88.89	89.98	+0.53
IT 도메인	82.91	78.58	80.69	+0.19

하지만 모든 자질을 사용하는 것이 가장 높은 성능을 보이는 것은 아니다. <표 IV-7>, <표 IV-8>, <표 IV-9>은 각 도메인에서 사용 자질 별 시스템 성능을 비교한 것이다. <표 IV-7>와 <표 IV-8>의 TV 도메인과 Sports 도메인에서는 word vector 자질과 군집 정보 자질을 모두 사용하는 것이 가장 성능이 높았다. 하지만 <표 IV-9>의 IT

도메인에서는 word vector 자질을 사용하지 않고, 400개로 군집화한 군집 정보 자질만을 사용하였을 때 가장 성능이 높았다. 생성한 자질들을 모두 사용하는 것보다 선택적으로 사용하는 것이 개체명 인식 성능에 더 유리할 수도 있다는 것을 보여준다.

<표 IV-7> 자질별 시스템 성능 비교 (TV 도메인)

사용 자질	$F_1 - measure(\%)$
기본 시스템	88.51
기본 시스템 + Word Vector	88.91
기본 시스템 + 군집 정보(200 개)	88.54
기본 시스템 + 군집 정보(300 개)	88.74
기본 시스템 + 군집 정보(400 개)	88.54
기본 시스템 + 군집 정보(500 개)	88.67
기본 시스템 + Word Vector + 군집 정보(300 개)	**89.03**

<표 IV-8> 자질별 시스템 성능 비교 (Sports 도메인)

사용 자질	$F_1 - measure(\%)$
기본 시스템	89.45
기본 시스템 + Word Vector	89.92
기본 시스템 + 군집 정보(200 개)	89.93
기본 시스템 + 군집 정보(300 개)	89.90
기본 시스템 + 군집 정보(400 개)	89.78
기본 시스템 + 군집 정보(500 개)	89.54
기본 시스템 + Word Vector + 군집 정보(200 개)	**89.98**

<표 IV-9> 자질별 시스템 성능 비교 (IT 도메인)

사용 자질	$F_1 - measure(\%)$
기본 시스템	80.50
기본 시스템 + Word Vector	80.51
기본 시스템 + 군집 정보(200 개)	81.19
기본 시스템 + 군집 정보(300 개)	80.92
기본 시스템 + 군집 정보(400 개)	**81.32**
기본 시스템 + 군집 정보(500 개)	81.27
기본 시스템 + Word Vector + 군집 정보(400 개)	80.69

<표 IV-10>은 TV 도메인에서 기존 한국어 개체명 인식 시스템과의 성능 비교표이다. Structural SVM을 사용한 방법과 제안 방법은 동일한 성능을 보였으며 FFNN과 CNN을 사용한 방법보다는 더 우수한 성능을 보였다. 제안 방법이 최신 한국어 개체명 인식 시스템의 성능과 큰 차이가 없음을 입증한다.

<표 IV-10> 한국어 개체명 인식 시스템 성능 비교 (TV 도메인)

시스템	$F_1 - measure(\%)$
Structural SVM	89.03
FFNN	87.74
CNN	88.57
제안방법	89.03

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 유형으로 나눌 수 있다. 첫 번째로 애매성에 의해 잘못된 개체명 범주가 부착된 경우이다. 예를 들어 '창원대학교'는 지명(LOC)과 기관명(ORG) 두 가지 범주에 모두 속하는 애매성을 가지고 있다. 이때 주변 정보를 이용하여 애매성을 해결하고 하나의 범주를 선택해야 한다. 하지만 주변 정보가 부족하면 애매성을 해결하지 못하고 잘못된 개체명 범주가 부착되는 오류가 발생한다. 잘못된 개체명 범주가 부착되는

두 번째 경우는 잘못된 분석에 의한 오류로 개체명이 아닌 것에도 개체명 범주를 부착하는 것이다.

다른 하나의 개체명 오류 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. <그림 IV-1>을 보면 인명(PER)으로 분류되는 개체명인 '박지성'은 형태소 분석 및 품사 부착 단계에서 '박지성/NNP'로 분석되어야 한다. 하지만 '박지/NNG+성/XSN'으로 잘못 분석될 경우 개체명으로 인식하지 못하는 오류를 발생시킬 수 있다.

박지성	박지/NNG + 성/XSN[2]

<그림 IV-1> 형태소 분석 및 품사 부착 오류 예

개체명을 인식하지 못한 두 번째 경우는 주변 정보가 부족하거나 학습이 부족하여 개체명을 인식하지 못한 오류이다. 특히 2어절 이상의 개체명의 경우 어절 중 일부를 개체명으로 인식하지 못한 오류가 있었다. 예를 들어 '웨인 루니'나 '크리스티아누 호날두' 같은 개체명에서 '웨인', '크리스티아누' 등을 개체명으로 인식하지 못한 오류이다.

<표 IV-11>, <표 IV-12>, <표 IV-13>는 각 도메인에서 가장 성능이 좋은 자질의 Confusion Matrix이다. 표에서 열은 시스템에서 부착한 개체명 출력 결과이고, 행은 실제 정답 개체명이다. 표에서 'None'은 개체명이 아닌 것을 의미한다. 각 Confusion

[2] XSN은 명사파생접미사를 의미한다.

Matrix를 살펴보면 개체명이 아닌데 개체명을 부착한 오류와 개체명을 인식하지 못한 오류가 많음을 알 수 있다. 각 표를 살펴보면 특히 문명/문화 관련 명칭 범주의 'CVL'에서 개체명을 인식하지 못한 오류가 많음을 알 수 있다. 이는 'CVL'에 2어절 이상의 개체명에서 나타나는 오류가 많이 포함되어 있기 때문이다. 예를 들어 '항공 전문가', 'IT 전문가'의 개체명에서 '전문가'만을 인식하거나, '주미 일본대사'에서 '일본대사'만을 인식하는 오류 등이 있다. 그리고 이벤트 범주의 'EVT'와 용어 범주의 'TRM' 또한 2어절 이상의 개체명을 많이 포함하고 있기 때문에 개체명 미 인식 오류가 많이 나타난다.

인명 범주의 'PER'과 기관 범주의 'ORG'에서도 개체명 미 인식 오류가 많이 나타난다. 두 범주에서는 2어절 이상의 개체명에서 나타나는 오류뿐만 아니라 개체명의 특성에서 발생하는 오류가 있다. 인명과 기관명의 경우 고유성이라는 특성을 가진다. 고유성은 데이터에서 출현 빈도가 낮다는 것을 의미하고, 평가 데이터의 개체명이 학습 데이터에서 나타나지 않는 경우를 발생시킨다. 이런 경우 주변 정보를 이용하여 개체명을 인식해야 하지만, 주변 정보가 부족할 경우 개체명을 인식하지 못하는 오류로 이어진다.

<표 IV-11> 기본 시스템+Word Vector+군집 정보(300개) (TV 도메인)

Sys / 정	PER	FLD	AFW	ORG	LOC	CVL	DAT	TIM	NUM	EVT	ANM	PLT	MAT	TRM	None	총합
PER	770	1	2	-	1	2	-	-	-	-	-	3	-	2	57	838
FLD	1	366	2	2	2	6	-	-	-	-	-	-	2	4	69	454
AFW	4	2	273	1	2	4	2	-	3	2	1	-	2	2	66	364
ORG	2	3	2	151	11	1	-	-	1	-	-	-	1	1	33	206
LOC	4	4	4	2	723	6	-	-	-	2	1	-	-	-	61	807
CVL	4	2	4	-	8	1,531	-	-	4	-	1	2	1	3	206	1,766
DAT	1	3	-	-	4	1	1,258	-	3	1	-	-	-	-	42	1,313
TIM	-	-	-	-	-	-	-	217	4	-	-	-	-	1	16	238
NUM	1	1	-	1	1	-	23	-	2,451	1	2	-	-	-	109	2,590
EVT	-	-	-	-	1	-	-	-	-	23	-	-	-	-	5	29
ANM	-	-	4	-	1	1	-	-	-	-	787	-	3	2	103	901
PLT	-	-	1	-	-	2	-	-	-	-	2	112	-	-	14	131
MAT	-	-	4	-	2	2	-	-	-	-	-	-	165	2	43	218
TRM	2	6	3	-	1	8	-	-	-	-	4	-	-	635	150	809
None	50	38	59	21	27	170	76	21	209	7	74	11	23	116	-	902
총합	839	426	358	178	784	1,734	1,359	238	2,675	36	872	128	197	768	974	

<표 IV-12> 기본 시스템+Word Vector+군집 정보(200개) (Sports 도메인)

Sys / 정	PER	FLD	AFW	ORG	LOC	CVL	DAT	TIM	NUM	EVT	ANM	PLT	MAT	TRM	None	총합
PER	2,836	-	-	3	4	2	-	-	1	-	1	-	-	1	151	2,999
FLD	-	12	-	10	-	2	-	-	-	1	-	-	-	-	20	45
AFW	2	-	116	6	3	-	-	-	1	4	-	-	-	-	22	154
ORG	12	-	1	2,427	25	12	1	-	3	14	-	-	-	-	130	2,625
LOC	6	-	3	21	690	8	-	-	-	14	-	-	-	-	54	796
CVL	10	-	-	7	5	2,704	3	-	15	7	-	-	-	6	429	3,186
DAT	2	-	-	-	-	-	1,098	-	46	1	-	-	-	-	34	1,181
TIM	-	-	-	-	-	-	1	160	1	-	-	-	-	-	6	168
NUM	2	-	-	1	-	5	-	2	4,048	3	-	-	-	2	121	4,184
EVT	2	1	-	42	23	8	1	-	8	580	-	-	-	4	172	841
ANM	-	-	-	-	-	1	-	-	-	-	199	-	-	6	39	245
PLT	-	-	-	-	-	-	-	-	-	-	-	3	-	-	7	10
MAT	-	-	-	-	-	1	-	-	-	-	-	-	1	-	11	13
TRM	9	-	2	10	1	8	-	-	10	4	4	-	-	750	332	1,130
None	99	1	7	87	31	276	38	9	223	146	11	-	-	174	-	1,102
총합	2,980	14	129	2,614	782	3,027	1,142	171	4,356	774	215	3	1	943	1,528	

<표 IV-13> 기본 시스템+군집 정보(400개) (IT 도메인)

Sys / 정	PER	FLD	AFW	ORG	LOC	CVL	DAT	TIM	NUM	EVT	ANM	PLT	MAT	TRM	None	총합
PER	373	-	-	5	2	2	-	-	-	-	-	-	-	1	63	446
FLD	1	271	-	4	-	4	-	-	1	-	-	-	-	6	61	348
AFW	1	2	62	7	2	3	-	-	-	1	-	-	-	2	38	118
ORG	3	15	9	945	12	5	1	-	-	-	-	-	-	8	76	1,074
LOC	-	-	3	9	435	1	-	-	-	2	-	-	-	4	44	498
CVL	2	13	-	14	2	597	-	-	3	2	-	-	1	10	116	760
DAT	-	-	-	-	-	-	384	-	-	-	-	-	-	-	85	469
TIM	-	-	-	-	-	-	-	15	-	-	-	-	-	-	3	18
NUM	-	1	-	-	-	-	1	-	413	-	-	-	-	1	85	501
EVT	-	6	1	4	4	2	-	-	-	25	-	-	-	4	33	79
ANM	1	-	-	-	-	-	-	-	-	-	13	-	-	-	12	26
PLT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	3
MAT	-	1	-	-	1	-	-	-	-	-	-	-	1	-	-	3
TRM	2	27	2	7	3	3	-	-	-	1	-	-	-	346	155	546
None	39	55	16	70	27	101	77	2	48	18	1	-	-	84	-	538
총합	422	391	93	1,065	488	718	463	17	465	49	14	-	2	466	774	

제 Ⅴ 장

결론 및 향후 연구

본 논문에서는 한국어 개체명 인식에서 영어에 비해 부족한 자질문제를 보완하고, 더 높은 성능의 한국어 개체명 인식 시스템을 만들기 위해 word embedding 자질을 이용하는 방법을 제안하였다. Word embedding 을 수행하기 위하여 CBOW 언어 모델을 이용하였다. CBOW 언어 모델은 현재 word 의 문맥을 이루는 word 들이 가지고 있는 vector 값의 합으로 현재 word 의 vector 값을 결정하는 모델이다. CBOW 언어 모델을 이용하여 형태소 단위의 word vector 를 생성하고, 이 vector 값을 K-means 알고리즘으로 군집화하여 군집 정보를 생성하였다.

Word embedding 자질을 이용하는 방법으로 형태소 단위의 word vector 와 군집 정보를 CRFs 의 자질로 사용하였다. TV 도메인과 Sports 도메인, IT 도메인으로 총 세가지 도메인에서 실험을 수행하였다. 실험을 수행한 결과 최고 성능이 기본 시스템보다 TV 도메인에서는 0.52%, Sports 도메인에서는 0.53%, IT 도메인에서는 0.82%로 각각 성능이 향상되어 그 효용성을 입증했다. 또한 최신의 한국어 개체명 인식 시스템과도 큰 성능 차이가 없었다. 하지만 어떤 word embedding 자질이 개체명 인식에서 가장 효용성이 있는지 알기 위해 다양한 실험이 필요하다.

향후에는 word embedding 을 형태소 단위가 아닌 개체명 단위로 수행하고, 그 word vector 를 자질로 사용하여 개체명 인식을 수행하는 실험을 수행할 것이다. 그리고 K-means 알고리즘이 아닌 다른 알고리즘을 이용하여 군집화를 수행하고, 생성된 군집 정보를 자질로 사용하는 실험을 수행할 것이다. 또한 아직까지 영어권 개체명 인식 시스템 성능에 비해 떨어지는 한국어 개체명 인식 시스템 성능을 향상시키기 위한 방법을 연구할 것이다.

참고문헌

1. 이경희, 이주호, 최명석, 김길창, "*한국어 문서에서 개체명 인식에 관한 연구*", 제 12 회 한글 및 한국어 정보처리 학술대회, pp. 292-299, 2000.
2. Daniel M. Bikel, Scott Miller, Richard Schwartz, Ralph Weischedel, "*Nymble: a High-Performance Learning Name-finder*", Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 194-201, 1997.
3. A. Ekbal, R. Haque, and S. Bandyopadhyay, "*Named Entity Recognition in Bengali: A Conditional Random Field Approach*", Proceedings of 3rd International Joint Conference Natural Language Processing (IJCNLP-08), pp. 589-594, 2008.
4. A. Ekbal and S. Bandyopadhyay, "*Bengali Named Entity Recognition using Support Vector Machine*", Proceedings of Workshop on NER for South and South East Asian Languages, 3rd International Joint Conference on Natural Language Processing (IJCNLP), (India), pp. 51-58, 2008.
5. Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, Xiangyang Zhou, "*Joint Inference of Named Entity Recognition and Normalization for Tweets*", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 526-535, 2012.

6. Ritter, Alan, Sam Clark, and Oren Etzioni, "*Named entity recognition in tweets: an experimental study*", Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1524-1534, 2001.
7. 정의석, 이현숙, 황이규, 윤보현, "*한국어 개체명 인식을 위한 CoTraining 기법*", 한국정보과학회 인간과 컴퓨터 상호작용 연구회 학술 대회 발표 논문집 (HCI) 제 2 호, pp. 525-529, 2003.
8. 이창기, 황이규, 오효정, 임수종, 허정, 이충희, 김현지, 왕지현, 장명길, "*Conditional Random Fields 를 이용한 세부 분류 개체명 인식*", 제 18 회 한글 및 한국어 정보처리 학술대회, pp. 268-272, 2006.
9. 이창기, 장명길, "*Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식*", 인지과학 제 21 권 제 4 호, pp. 655-667, 2010.
10. 송영길, 정석원, 김학수, "*위키피디아를 이용한 정보검색 기반 개체명 사전 구축 방법*", 한국정보과학회 학술발표논문집, pp. 648-650, 2015.
11. 이창기, 김준석, 김정희, 김현기, "*딥 러닝을 이용한 개체명 인식*", 한국정보과학회 학술발표논문집, pp. 423-425, 2014.
12. J. Turian, L. Ratinov and Y. Bengio, "*Word representations: A simple and general method for semi-supervised learning*", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 384-394, 2010.
13. Y.Bengio, R.Ducharme, P.Vincent. "*A neural probabilistic language model*", Journal of Machine Learning research, Vol.3, pp. 1137-1155, 2003.

14. T. Mikolov, K. Chen, G. Corrado and J. Dean, "*Efficient Estimation of Word Representations in Vector Space*", ICLR Workshop, 2013.
15. J. Lafferty, A. McCallum, F. Pereira, "*Conditional random fields: Probabilistic models for segmenting and labeling sequence data*", Proceedings. 18th International Conference on Machine Learning, pp. 282-289, 2001.
16. 이호석, "*조건부 랜덤 필드와 응용에 대한 고찰*", 한국정보과학회 가을 학술발표논문집 제 36 권 제 2 호, pp. 184-187, 2009
17. 홍진표, 차정원, "*어절패턴 사전을 이용한 새로운 한국어 형태소 분석기*", 한국정보과학회 종합학술대회 논문집 제 35 권 제 1 호, pp. 279-274, 2008

ABSTRACT

Korean Named Entity Recognition
Using Word Embedding Features

Choi Yunsu
Dept. of Eco-Friendly Offshore Plant FEED Engineering
Graduate School, Changwon National University
Changwon, Korea

Named Entity Recognition (NER) is the task to recognize and classify named entities such as person name, location, and organization. There were various studies on Korean Named Entity Recognition, but those have some problems, for example lacking features as compared to English NER. In this paper, we propose a method that uses word embedding as features for Korean NER. We generate word vector using Continuous-Bag-of-Words(CBOW) model from POS tagged corpus, and word cluster symbol using K-means algorithm from word vector. We use word vector and word cluster symbol as word embedding features in Conditional Random Fields(CRFs). From the result of experiment, performance improves 0.52%, 0.53% and 0.82% respectively in TV domain, Sports domain and IT domain over the baseline system. Showing better performance than other NER systems, we demonstrate effectiveness and efficiency of the proposed method.

KEYWORDS

NLP(Natural Language Processing), NE(Named Entity), NER(Named Entity Recognition), Word Embedding, CBOW(Continuous-Bag-of-Words), Machine Learning, CRFs(Conditional Random Fields)

부록 A. 품사 집합

TAG	POS	TAG	POS
NNG	일반명사	IC	감탄사
NNB	의존명사	VCP	긍정지정사
NNP	고유명사	VCN	부정지정사
NP	대명사	VV	동사
NR	수사	VA	형용사
JKS	주격조사	VX	보조용언
JKC	보격조사	EF	종결어미
JKO	목적격조사	EC	연결어미
JKG	관형격조사	ETN	명사형 전성어미
JKB	부사격조사	ETM	관형형 전성어미
JKV	호격조사	EP	선어말어미
JKQ	인용격조사	SF	마침표, 물음표, 느낌표
JC	접속조사	SP	쉼표, 가운뎃점, 콜론, 빗금
JX	보조사	SS	따옴표, 괄호표, 줄표
XPN	명사접두사	SE	줄임표
XSN	명사파생접미사	SO	붙임표(물결, 숨김, 빠짐)
XSB	부사파생접미사	SL	외국어
XSV	동사파생접미사	SH	한자
XSA	형용사파생접미사	SN	숫자
XR	어근	NF	명사추정범주
MM	관형사	NV	용언추정범주
MAG	일반부사	SW	기타기호
MAJ	접속부사	NA	분석불능범주

이 력 서

성　　명: 최 윤 수
생년월일: 1988년 05월 01일
출 생 지: 부산광역시 해운대구
주　　소: 부산광역시 해운대구 반여1동 918-7번지 20통 1반

학 력

2007-20014: 창원대학교 공과대학 정보통신공학과(B.S.)
20014-2016: 창원대학교 대학원 친환경해양플랜트FEED공학과정
(컴퓨터 · 정보통신공학)(M.S.)

발표논문

1. 최윤수, 정진욱, 황민태, 진교홍, “스마트 교육을 위한 전자칠판시스템용 판서 소프트웨어 개발”, 2014 한국정보처리학회 춘계학술발표대회 논문집 제21권 제1호, pp. 1043-1046, 2014
2. 최윤수, 정진욱, 황민태, 진교홍, “사용자 동작 인식 기능을 지원하는 판서 소프트웨어 개발”, 한국정보통신학회논문지 제19권 제5호, pp. 1213-1220, 2015.
3. 김중한, 최윤수, 박태호, 차정원, “개체명 부착 말뭉치에서 자동 오류 수정”, 2015 한국컴퓨터종합학술대회(KCC2015) 논문집, pp. 669-671, 2015.
4. 최윤수, 황민태, “태블릿 기기와 전자칠판 시스템 간의 연동 기술 연구”, 한국정보통신학회논문지 제19권 제7호, pp. 1719-1727, 2015.
5. 최윤수, 차정원, “Word Embeddings 자질을 이용한 한국어 개체명 인식 및 분류”, 2015 한국정보과학회 동계학술발표회, pp. 546-548, 2015.