KIPS Tr. Software and Data Eng.
Vol.3, No.7 pp.285~292 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.7.285

Named Entity Recognition and Dictionary Construction for Korean Title: Books, Movies, Music and TV Programs

Yongmin Park† · Jae Sung Lee††

ABSTRACT

A named entity recognition method is used to improve the performance of information retrieval systems, question answering systems, machine translation systems and so on. The targets of the named entity recognition are usually PLOs (persons, locations and organizations). They are usually proper nouns or unregistered words, and traditional named entity recognizers use these characteristics to find out named entity candidates. The titles of books, movies and TV programs have different characteristics than PLO entities. They are sometimes multiple phrases, one sentence, or special characters. This makes it difficult to find the named entity candidates. In this paper we propose a method to quickly extract title named entities from news articles and automatically build a named entity dictionary for the titles. For the candidates identification, the word phrases enclosed with special symbols in a sentence are firstly extracted, and then verified by the SVM with using feature words and their distances. For the classification of the extracted title candidates, SVM is used with the mutual information of word contexts.

Keywords : Named Entity Recognition, Title Named Entity, Dictionary Construction, SVM

한국어 제목 개체명 인식 및 사전 구축: 도서, 영화, 음악, TV프로그램

박 용 민† · 이 재 성††

요 약

개체명 인식은 정보검색 시스템, 질의응답 시스템, 기계번역 시스템 등의 성능을 향상시키기 위하여 사용된다. 개체명 인식은 일반적으로 PLOs(인명, 지명, 기관명)을 대상으로 하며, 주로 미등록어와 고유명사로 이루어져 있기 때문에 고유명사나 미등록어는 중요한 개체명 후보로 쓰일 수 있다. 하지만 도서명, 영화명, 음악명, TV프로그램명과 같은 제목 개체명은 PLO와는 달리 단어부터 문장까지 매우 다양한 형태를 지니고 있어서 개체명 인식이 쉽지 않다. 본 논문에서는 뉴스 기사문을 이용하여 제목 개체명을 빠르게 인식하고 자동으로 사전을 구축하는 방법을 제안한다. 먼저 특수기호로 묶인 어절을 추출하고, 주변 문맥 단어 및 단어 거리를 이용하여 SVM으로 제목 후보들을 추출하였다. 이렇게 추출된 제목 후보들은 상호 정보량을 가중치로 SVM을 이용해 제목 유형을 분류하였다.

키워드 : 개체명 인식, 제목 개체명, 사전 구축, SVM

1. 서 론

개체명(Named Entity)이란 인명, 지명, 기관명, 날짜, 시간 등 문장에서 핵심적인 의미를 지닌 고유명사나 미등록어 등을 말하며, 개체명 인식(NER : Named Entity Recognition)이란 텍스트에 나타난 이러한 개체명을 찾아 그 범위와 종류를 결정하는 작업을 말한다[1].

개체명 인식을 활용하면 기계번역 시스템에서 중의성을 가지는 단어들로 인한 번역 오류를 줄일 수 있고, 질의응답 시스템에서는 질의문에 포함된 개체명 정보를 활용함으로써 질의 의도에 부합하는 정답을 찾아낼 수 있다. 이렇듯 개체명 인식은 자연어처리에 있어서 핵심적인 기능이고, 정보검색, 질의응답, 기계번역 등 각 시스템의 성능 향상에 큰 역할을 한다[2].

개체명 인식은 개체명이 될 수 있는 후보를 찾는 개체명

※ 본 연구는 미래 창조 과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음[10044577, (1세부) 휴먼 지식증강 서비스를 위한 지능진화형 Wise QA 플래폼 기술개발].
† 준 회 원 : 충북대학교 디지털정보융합학과 석사
†† 종신회원 : 충북대학교 소프트웨어학과 교수
Manuscript Received : March 17, 2014
First Revision : May 26, 2014
Accepted : May 27, 2014
* Corresponding Author : Jae Sung Lee(jasonlee@cbnu.ac.kr)

경계 인식과 찾은 개체명 후보를 인명, 지명, 조직명 등으로 분류하는 개체명 유형 분류로 나뉜다. 개체명은 주로 미등록어와 고유명사로 이루어져 있기 때문에 개체명 후보 추출에 있어서 고유명사나 미등록어는 중요한 개체명 후보로 쓰일 수 있다.

하지만 도서명, 영화명, 음악명, TV프로그램명 등과 같은 제목 개체명(Title Named Entity)을 인식하는 것은 기존의 인명, 지명, 조직명과 같은 형식의 개체명에 비해 상대적으로 어렵다[3]. 그 이유는 첫째, 제목 개체명은 단어부터 문장까지 매우 다양한 형태를 지니고 있다. 인명, 지명, 조직명이 보통 하나의 단어나 두세 개 정도의 어절로 구성되어 있는 것과 비교할 때, 제목 개체명은 단어부터 명사구, 또는 문장에 이르기까지 형태가 매우 다양하다. 둘째, 제목 개체명은 다른 종류의 개체명이 되는 경우도 있다. 예를 들어, '스티브 잡스'는 사람 이름이면서 도서명이기도 하고, '베를린'은 지역명이면서 영화명이기도 하다. 셋째, 제목 개체명은 개체명 인식에 사용할 수 있는 특징적인 내부 자질이 존재하지 않는다. 일반적으로 '특별시'나 '광역시', '국립공원' 등과 같은 접미사는 해당 단어를 지명으로 인식할 수 있는 결정적인 단서가 될 수 있다. 하지만 제목 개체명은 다양한 형식으로 인하여 내부에 제목 개체명만의 특징을 지니지 않는다.

이렇게 제목과 같이 형태의 다양성으로 인하여 개체명 경계 인식이 불분명할 경우에는 사전을 구축하여 사전 매칭으로 해결하는 것이 효과적이다[4,5]. 따라서 본 논문에서는 도서, 영화, 음악, TV프로그램의 제목 개체명에 대한 특징을 살펴보고, 뉴스 기사문을 이용하여 새롭게 생성되는 제목 개체명을 실시간으로 인식하며, 사전으로 구성하는 방법을 제안한다.

논문의 구성은 다음과 같다. 2장에서는 기존의 개체명 인식 방법 및 제목 개체명 인식에 관한 연구를 살펴본다. 3장에서는 뉴스 기사문에서 제목 개체명이 가지는 특징에 대하여 살펴보고, 문맥 단어를 이용한 개체명 인식 방법 및 사전 구축 방법을 제안한다. 4장에서는 뉴스 기사문에서 추출한 제목의 식별과 분류 성능을 평가하며, 5장에서는 본 논문의 결론과 향후 연구 방향을 제시한다.

2. 관련 연구

개체명 인식에 관한 연구는 1990년대 영어권에서 시작되었다. 미국 방위 고등 연구 계획국(Defence Advanced Research Projects Agency)은 새롭고 보다 나은 정보 추출(Information Extraction)방법을 개발하기 위하여 MUC(Message Understanding Conference)를 주최하였으며, MUC-6, MUC-7에서 개체명 인식에 관한 연구를 본격화하였다[6,7]. 또한 MET(Multilingual Entity Task)나 일본의 IREX(Information Retrieval and Extraction Exercise) 등을 통하여 비영어권 국가에서도 개체명 인식에 관한 연구들이 진행되어 왔으며[8,9], 한국어 개체명 인식에 관한 연구도 다양하게 진행되어 왔다[1,2,10,11,12,13,14].

개체명 인식 방법은 크게 '규칙 기반 개체명 인식'과 '통계 기반 개체명 인식'으로 나눌 수 있다. 따라서 기존에 연구된 '규칙 기반 개체명 인식'과 '통계 기반 개체명 인식' 방법에 대하여 살펴보고, '제목 개체명 인식'의 특징을 살펴보았다.

2.1 규칙 기반 개체명 인식

규칙 기반 개체명 인식은 패턴이나 규칙을 수동이나 반자동으로 작성하고 개체명 사전을 확장시켜서 규칙에 적용되는 개체명을 새로운 문서에서 직접 추출하는 방법이다[10,12].

[10]의 연구에서는 사전 정보(개체명 사전, 개체명과 함께 사용되는 단어들의 사전)와 함께 개체명 인식을 위한 4단계 규칙을 순차적으로 적용시켜 단계별로 부여된 가중치에 따라 인명, 지명, 조직명에 대한 개체명 범주를 결정하였다.

[12]에서는 수작업으로 작성된 규칙을 이용한 한국어 문서에서의 지명 인식 방법을 제안하였다. 우편번호부에 등록되어 있는 일부 지역을 대상으로 지명 사전을 검색하여 사전에 등재되어 있을 경우, 지명 개체명으로 인식하였으며, 사전에 등재되어 있지 않을 경우, 지명 후보가 되는 단어의 앞이나 뒤에 나타나는 단어의 품사를 고려하여 문맥 규칙을 적용시켜서 지명에 대한 개체명을 추출하였다.

규칙 기반 개체명 인식의 성능은 사전의 품질에 영향을 많이 받는다. 또한 규칙이나 패턴을 수작업으로 작성할 경우 비용과 시간이 많이 든다는 단점이 있다. 이러한 단점을 해결하기 위하여 연구되어 온 방법이 통계 기반 개체명 인식 방법이다.

2.2 통계 기반 개체명 인식

통계 기반 개체명 인식은 학습 방법에 따라 은닉 마르코프 모델(HMM : Hidden Markov Model)에 기반한 방법[13]과 최대 엔트로피 모델(MEM : Maximum Entropy Model)을 이용한 방법[1], CRFs(Conditional Random Fields)를 이용한 방법[14], SVM(Support Vector Machine)을 이용한 방법[2] 등으로 나눌 수 있다.

[13]에서는 HMM에 기반한 복합 명사 구성 원리를 이용한 한국어 개체명 인식 방법을 제안하였다. 단어들을 개체명 독립 단어, 구성 단어, 인접 단어, 개체명과 관련 없는 단어의 네 가지 범주로 분류하였다. 개체명 관련 단어 유형과 품사를 이용한 HMM으로 개체명 경계를 인식하였으며, 가변길이의 개체명을 인식하기 위해 트라이그램 모델을 이용하였다.

[1]에서는 개체명 인식을 경계 인식과 유형 분류로 구분하고, 2단계 최대 엔트로피 모델을 적용한 개체명 인식 방법을 제안하였다. 1단계 경계 인식에서는 개체명의 시작과 중간, 끝, 단일어 등의 레이블이 표기된 학습 말뭉치를 이용하여 경계 인식 ME(Maximum Entropy) 학습을 하였으며, 2단계 유형 분류에서는 경계 인식에서 태깅된 개체명 후보들의 조사 정보, 용언 정보, 헤드 명사 정보 등을 이용하여 개체명 의미 범주를 결정하였다.

[14]의 연구에서는 CRFs와 MEM을 혼합하여 질의응답 시스템을 위한 세부 분류 개체명 인식 방법을 제안하였다.

먼저 질의응답 시스템의 정답이 될 수 있는 15개 대분류와 이에 속하는 147개 세부 분류된 개체명 카테고리를 정의하였다. 개체명 경계는 개체명의 시작, 중간, 끝을 나타내는 태그와 형태소 위치 정보 등을 이용하여 CRFs로 인식하였으며, 개체명 클래스는 개체명 클래스 정보와 개체명 경계 정보가 합쳐진 클래스를 학습 말뭉치로 MEM을 이용해 분류하였다.

[2]에서는 Structural SVMs와 수정된 Pegasos 알고리즘을 이용하여 한국어 개체명 인식 시스템을 개발하였다. 특히, 기존의 CRFs를 이용한 한국어 개체명 인식 시스템[13]을 Structural SVMs을 이용하여 성능을 향상시키고, 수정된 Pegasos 알고리즘으로 기존의 높은 성능을 유지하면서 CRFs에 비하여 학습 시간을 단축하였다.

이상에서 살펴본 규칙 기반 개체명 인식과 통계 기반 개체명 인식은 대부분 인명, 지명, 기관명 등 일부 개체명 유형만을 대상으로 하고 있으며, 주로 패턴이라든지 주변 단어들 간의 관계를 살펴서 개체명 경계를 인식한다. 하지만 제목 개체명과 같이 형태가 다양한 경우, 개체명 경계 인식에 매우 취약하다. 따라서 사전을 구성하여 활용하는 경우가 많다. 다음으로 제목 개체명 인식에 관한 연구에 대하여 살펴보았다.

2.3 제목 개체명 인식

제목 개체명 인식(Title Named Entity Recognition)은 주로 영화명이라든지 도서명, 드라마명 등의 제목에 대한 개체명을 인식하는 것으로, 다음과 같은 연구가 진행되었다[3,15,16].

[15]는 e-mail의 내용을 분석하여 영화의 제목, 날짜, 시간을 추출하는 방법을 제안하였다. 개인 e-mail을 수작업으로 영화와 관련된 문서와 그렇지 않은 문서로 분류하고, 이것을 나이브 베이지안 분류기(Naive Bayes classifier)의 학습 데이터로 사용하였다. 영화와 관련된 문서는 다시 최대 엔트로피 마르코프 모델(MEMM : Maximum Entropy Markov Model)을 이용하여 영화 제목, 장소, 날짜, 시간 등의 개체명을 인식하였다. 하지만 영화 관련 문서의 분류 성능은 F_1-score 75.34%이었지만, MEMM을 이용한 제목 개체명 추출은 F_1-score 13.33%로 매우 낮은 성능을 보였다. 따라서 논문에서는 영화 관련 데이터베이스를 이용하거나 자주 상영되는 영화명을 학습시킬 것을 권장하고 있다.

[3]에서는 원시 말뭉치로부터 영화명, 도서명, 노래명의 주변 문맥에 등장하는 문맥 패턴과 개체명 사전을 자동으로 구축하고, 이를 이용해 제목 개체명을 인식하는 방법을 제안하였다. 제목 개체명의 앞, 뒤 단어를 문맥 패턴 후보로 추출하며, 인용부호나 MUC형식 개체명의 다양성을 일반화하기 위해 사람, 날짜, 시간, 금액의 경우 특정 기호로 일괄 변경하였다. 패턴의 신뢰도가 임계값 이상인 것만을 패턴 저장소에 등록하였으며, 새로운 패턴이 구축되면 사전 확장 단계에서 말뭉치 내 해당 패턴과 일치하는 개체명 후보를 사전에 추가하였다. 이러한 패턴 구축 및 사전 확장단계를 반복하여 패턴과 사전의 크기를 점진적으로 증가시켰다. 그러나 NY Times 문서를 대상으로 하고 있기 때문에 한국어 문서에 직접 적용하기는 힘들다. 또한 사전과 패턴을 모두 사용하였을 경우 정확률(Precision) 58.38%, 재현율(Recall) 52.22%, F_1-score 55.16%로, 사전을 구성함에 있어 정확률이 부족한 것을 알 수 있다.

[16]의 연구에서는 위키백과를 이용한 제목 개체명 인식 모델을 제안하였다. 위키백과에 포함되어 있는 분류 태그 중 '영화', '드라마' 등 제목 개체명이라 판단되는 총 10가지 분류체계를 선정하고 해당 분류체계인 문서의 제목을 제목 개체명 사전으로 구축하였다. 또한 제목에 대한 약어 생성 규칙을 통해 제목의 약어 후보를 생성하고, 웹 검색을 통한 검증 과정을 거쳐서 약어 사전에 등록하였다. 이렇게 구축한 제목 및 제목 약어 사전을 자질로 CRFs를 이용하여 제목 개체명을 인식하였으며, 제목 개체명과 약어를 포함하여 F_1-score 82.1%의 성능을 보였다. 하지만 위키백과는 사용자가 직접 작성하는 백과사전으로, 새로운 도서나 드라마 등에 대한 정보가 위키백과에 등록되어 있지 않으면 제목 개체명 사전을 구축하는 데 제약이 따른다. 따라서 실시간 이슈가 되는 소셜 데이터의 분석에 있어서 위키백과의 활용은 한계가 있다.

본 논문에서는 일반 문서에 비해 정규화된 형식으로 이루어진 뉴스 기사문에서 제목 개체명이 가지는 특성을 분석하였으며, 이를 이용해 제목 개체명 추출과 제목 개체명 사전을 생성하기 위한 방법을 제안한다.

3. 문맥 단어를 이용한 제목 개체명 인식

도서명, 영화명, 음악명, TV프로그램명과 같은 제목은 다양한 형식으로 표현된다. 제목은 특정한 형태를 가지지 않기 때문에 개체명 인식의 방해요소로 작용하기도 한다. 이러한 제목 개체명의 형태는 표 1과 같이 분류할 수 있다.

Table 1. Various forms of titles

Form	Example
Noun	변호인, 숨바꼭질, 베를린, 관상
Modifier+Noun	주군의 태양, 불의 여신 정이, 꽃보다 할배
Adverb	은밀하게 위대하게
Number	28, 300
Sentence	천 번을 흔들려야 어른이 된다
Foreign language	굿 닥터, 다큐프라임, 더 테러 라이브
Compound word	PD수첩, 강연100℃, Let 美人
Original title +series(season)	우리 결혼했어요 시즌4, 슈퍼스타K 시즌5
Original title +series(number)	아이언맨3, 나의 문화유산답사기7
Original title + Subtitle	일밤:진짜 사나이, 나인:아홉 번의 시간여행
Etc.	다시 아이를 키운다면, 부모라면 유대인처럼

제목은 단순히 하나의 단어로 구성된 것부터 3어절 이상의 문장형으로 이루어진 것까지 매우 다양한 형태를 가진다. 이와 같은 제목의 다양성으로 인하여 제목 개체명의 경계 인식은 쉽지 않다.

본 논문에서는 제목 개체명의 경계 인식 문제를 해결하기 위해 정규화된 형식을 갖춘 뉴스 기사문을 활용하였다. 우선, 제목 개체명을 포함하는 뉴스 기사문[1]을 살펴보면 표 2와 같다.

Table 2. The examples of title named entity in news articles

Type	Example
Movie	**할리우드 블록버스터** ***'그래비티'***가 주말 **예매 점유율**에서 1위를 차지했다.
Music	**밴드 버스커버스커**가 2012년 발표했던 **노래** ***'벚꽃엔딩'***이 다시 **차트**에 재진입해 눈길을 끈다.

일반적으로 뉴스 기사문에서 제목을 나타내는 개체명은 특수기호 사이에 존재하는 경우가 많다. 또한 제목 개체명의 주변에 제목을 설명하기 위한 단어들이 분포함을 알 수 있다.

본 논문에서는 뉴스 기사문에서 특수기호 사이에 존재하는 어절을 대상으로 제목을 식별하고, 주변 문맥 단어로 제목의 유형(도서명, 영화명, 음악명, TV프로그램명)을 분류하는 방법을 제안한다. 전체적인 과정은 그림 1과 같다.

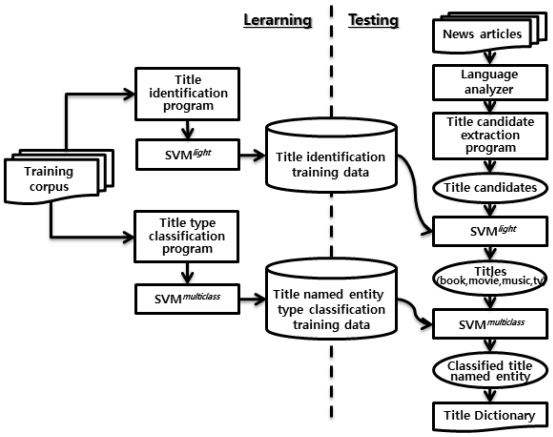

Fig. 1. The conceptual diagram of title named entity recognition and dictionary construction

뉴스 기사문에서 특수기호 사이에 존재하는 어절들을 추출하고, 이것을 제목과 비제목으로 구분하였다. 이때, 미리 학습된 제목 식별 학습 데이터를 사용하였으며, 이후에 제목으로 판단되는 어절을 대상으로 도서명, 영화명, 음악명, TV프로그램명으로 각각 분류하였고, 이를 이용해 제목 개체명 사전을 구축하였다.

학습과 실험에는 기계학습 방법 중 하나인 SVM(Support Vector Machine)을 사용하였다. SVM은 이원 패턴 인식 문제를 해결하기 위하여 1995년 Vapnik에 의해 제안된 방법이다[17]. 기존의 개체명 인식에서 주로 사용하던 HMM, MEMM, CRFs는 단어 자질뿐만 아니라 단어 간의 순서 정보도 필요로 하지만 SVM은 단어 간의 순서 정보를 필요로 하지 않고 단순히 문서 분류를 위한 자질만을 이용하기 때문에 비교적 간단하게 연산할 수 있다. 또한 나이브 베이지안 분류기(Naive Bayes Classifier)나 최근접 이웃 분류기(Nearest Neighborhood Classifier), 신경망 모형(Neural Networks Models) 등 문서 범주화와 관련된 학습 방법 중 SVM의 분류 성능이 가장 우수한 것으로 알려져 있다[18].

본 논문에서는 Cornell University, Dept. of Computer Science에서 Support Vector Machine[17]을 구현한 이진 분류기인 SVM^{light}를 이용해 제목과 비제목을 구분하였으며, Multi-class SVMs[19]을 구현한 다중 분류기인 $SVM^{multiclass}$를 이용해 제목 개체명의 유형을 분류하였다.

3.1 문맥 단어 추출

본 논문에서는 제목 추출 및 제목 유형 분류를 위하여 문맥 단어를 이용하였다. 문맥 단어는 제목 개체명 좌/우에 공기하는 단어들을 의미하며, 가중치 부여를 위하여 제목 개체명으로부터 떨어진 거리정보까지 포함하였다. 문맥 단어는 제목과 비제목을 구분하기 위한 학습자료 생성, 제목 개체명의 분류를 위한 학습자료 생성, 그리고 평가를 위한 실험에 사용된다. 문맥 단어 추출을 위한 알고리즘은 다음과 같다.

1단계 : 뉴스 기사문 내에서 특수기호(' ', < >,' ')로 묶인 어절을 포함하는 문장을 추출한다.

2단계 : 1단계에서 추출한 문장에서 특수기호로 묶인 어절의 좌/우에 공기하는 명사를 추출한다. 이때, 가까운 거리 순으로 사용자가 지정한 개수만큼 추출하며, 지정한 개수보다 명사가 적을 경우 가능한 만큼만 추출한다.

예를 들어, 표 3과 같은 문장에서 제목 개체명의 좌/우 문맥 단어를 추출한다고 하자.

Table 3. The example of context word extraction

지난 21일 오후 방송된 MBC 예능프로그램 '**무한도전**'이 토요일 전체 예능 프로그램 시청률 1위를 차지했다.

	L_5	L_4	L_3	L_2	L_1	Main word
Noun	오후	방송	MBC	예능	프로그램	**무한도전**

	Main word	R_1	R_2	R_3	R_4	R_5
Noun	**무한도전**	토요일	전체	예능	프로그램	시청

1) http://goo.gl/ul0UJu, http://goo.gl/wSjeGh

우선, 특수기호(' ', < >, ' ')로 묶인 어절을 찾는다. 특수기호 사이의 어절인 '무한도전'을 중심어로 간주하고, 이를 기준으로 사용자 지정 개수만큼 좌/우에 존재하는 명사를 추출한다.

3.2 제목 식별

학습 말뭉치에서 문맥 단어 추출 알고리즘을 통해 추출된 데이터 중, 특수기호 사이의 중심어가 도서, 영화, 음악, TV프로그램 개체명으로 태그된 것을 제목 학습 데이터로, 나머지를 비제목 학습 데이터로 사용하였다. 특히, 말뭉치에 포함된 뉴스 기사문 카테고리 중, 도서는 '출판', 영화는 '영화', 음악은 '음반', TV프로그램은 '예능', '드라마', '연예가뉴스', '방송/TV속으로' 카테고리에 속하는 것만을 학습 데이터로 사용하였다. 이는 뉴스 기사문의 주된 내용(주제)이 해당 제목 개체명에 관한 내용일 때 주변 문맥 단어를 제대로 활용할 수 있기 때문이다.

제목 개체명 인식에는 기계학습 도구인 SVM^{light}를 사용하여 제목과 비제목 구분을 위한 문맥 단어 학습 및 분류를 수행하였으며, 문맥 단어들을 이용하여 학습할 때 위치에 따른 가중치와 빈도에 따른 가중치를 사용하였다. 가중치에 대한 수식은 식 1과 같다.

$$\begin{aligned} &W^l = \{w_1^l, w_2^l, \ldots, w_n^l\},\ \ W^r = \{w_1^r, w_2^r, \ldots, w_m^r\} \\ &V(w_i) = F(w_i) \times D(w_i),\ \ F(w_i^l) = \frac{f(w_i^l)}{|W^l|} \\ &D(w_i) = C - d(w_i) + 1\ \ where\ \ w_i \in W^l \text{ or } w_i \in W^r \end{aligned} \tag{1}$$

- w : 단어 ($w \in W^l$ or $w \in W^r$)
- W^l, W^r : 중심어 좌측, 우측 단어 집합
- $V(w)$: w의 가중치, $F(w)$: w의 빈도 가중치
- $f(w)$: w의 빈도, $D(w)$: w의 위치 가중치
- $d(w)$: w와 제목 개체명 사이의 거리($1 \le d(w) \le C$)
- C : 좌/우 문맥 단어 개수(윈도우 크기)

제목 개체명 좌측에는 주로 제목을 설명하기 위한 단어들이 분포하고, 우측에는 제목과 관련된 내용을 언급하는 경우가 많다. 따라서 본 실험에서는 좌측과 우측의 단어를 구분하였다. 또한 일반적으로 특수기호 사이의 중심어에 가까울수록 중심어에 대한 문맥 단어의 영향력이 크기 때문에 특수기호 사이의 중심어에 대하여 위치값에 따른 가중치를 부여하였으며, 좌/우 각각 빈도에 따른 가중치도 적용하였다.

3.3 제목 개체명 유형 분류

제목 개체명을 이용하여 제목의 유형에 따라 4가지로 각각 분류하였다. 우선, 특수기호 사이의 중심어가 도서명, 영화명, 음악명, TV프로그램명으로 태그된 것을 각각 분류하고, 중심어 주변의 문맥 단어를 추출하였다. 이때, 문맥 단어들은 제목 개체명 유형별 상호정보량(MI : Mutual Information)을 이용하여 가중치를 부여하였다. 상호정보량이란 두 독립사건 X, Y 사이의 의존관계를 수치화하여 나타낸 값으로 식 2와 같다.

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

즉, X와 Y의 연관성이 높을수록 상호정보량은 높은 값을 가진다. 상호정보량은 두 변수 X, Y 사이의 독립성에 관한 가정이 없으며, 데이터 변형이나 잡음(noise)에 의한 영향이 적어서 다른 방법들에 비하여 신뢰도가 높은 편이다[20].

본 논문에서는 각각의 문맥 단어와 제목 개체명 유형사이의 상호정보량을 구하여 제목 개체명 분류를 위한 문맥 단어의 가중치로 활용하였으며, 가중치에 대한 수식은 식 3과 같다.

$$\begin{aligned} &W = \{w_1, w_2, \ldots, w_n\},\ \ C = \{C_{book}, C_{movie}, C_{music}, C_{tv}\} \\ &MI(w; C_i) = \frac{N_{11}}{N} log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} log_2 \frac{NN_{01}}{N_{0.}N_{.1}} \\ &\qquad + \frac{N_{10}}{N} log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} log_2 \frac{NN_{00}}{N_{0.}N_{.0}} \end{aligned} \tag{3}$$

- w : 단어 ($w \in W$)
- C : 제목 개체명 유형 집합
- $MI(w; C_i)$: C_i에 대한 단어 w의 상호정보량($C_i \in C$)
- $N = N_{00} + N_{01} + N_{10} + N_{11}$
- $N_{1.} = N_{10} + N_{11}, N_{.1} = N_{01} + N_{11}$
  $N_{0.} = N_{00} + N_{01},\ N_{.0} = N_{00} + N_{10}$
- N_{11} : 단어 w를 포함하고 C_i에 속하는 문서 수
- N_{01} : 단어 w를 포함하지 않고 C_i에 속하는 문서 수
- N_{10} : 단어 w를 포함하고 C_i에 속하지 않는 문서 수
- N_{00} : 단어 w를 포함하지 않고 C_i에 속하지 않는 문서 수

본 논문에서는 뉴스 기사문 중 특수기호 사이에 중심어가 포함된 각각의 문장을 하나의 문서로 보았다. 또한 제목 개체명 유형별 단어들의 상호정보량을 가중치로 사용하고, 다중 분류 기계학습 도구인 $SVM^{multiclass}$를 이용하여 학습 및 실험을 하였다.

특히, 하나의 문서 내에 2회 이상 등장하는 제목 개체명 후보는 동일한 제목 유형만을 가진다고 가정하고 후처리를 진행하였다. 예를 들어, '변호인'이라는 제목 개체명 후보가 한 문서에서 4번 등장하였고, 3개는 '영화'로, 1개는 'TV프로그램'으로 분류되었을 경우 개수가 많은 '영화'로 개체명 태깅을 하였다. 반면, '영화'와 'TV프로그램'으로 각각 2개씩 분류되었을 경우, $SVM^{multiclass}$의 분류 가중치를 활용하여 가중치가 가장 높은 분류를 따르도록 하였다.

4. 실험 및 평가

실험에는 2013년 7월 한 달 동안 수집된 온라인 뉴스 기사문 19,745개를 대상으로 ETRI 언어 분석기를 통해 분석된 언어 분석 결과를 학습 말뭉치로 사용하였다. 말뭉치는 각 기사문에 대한 카테고리와 기사 제목 및 본문을 포함하고 있으며, 기사문에 대한 문장단위 형태소 분석, DP(dependency parser) 결과 등을 포함한다. 전체 말뭉치 중 90%는 학습에 사용하고, 나머지 10% 중 임의로 추출한 200개 문서를 정제하여 실험에 사용하였다.

실험 데이터를 분석한 결과, 200개 문서 중 제목 개체명을 포함하는 문서는 75개였다. 문서 내에 포함된 제목 개체명은 전체 270개(중복 제외 158개)였으며, 제목 개체명 중 특수기호로 묶여 있는 제목이 97.04%(262개), 특수기호로 묶여 있지 않은 제목이 2.96%(8개)이었다. 또한 특수기호를 분석한 결과, 세 가지 유형의 특수기호(' ', < >, ' ')로 묶여 있음을 확인할 수 있었다. 특수기호의 유형 및 비율은 표 4와 같다.

Table 4. The types of special symbol that encloses title named entity

Special symbol	EA.	Ratio
' '	178	73.25%
' '	58	23.87%
< >	7	2.88%
Total	243	100%

본 실험에서는 특수기호 여부와 상관없이 문서에 나타난 모든 제목 개체명을 대상으로 정답집합을 구성하였으며, 평가는 식 4와 같이 정확률과 재현율을 이용한 F_1-score로 계산하였다.

$$\text{정확률}(P)=\frac{n(A\cap B)}{n(B)},\ \text{재현율}(R)=\frac{n(A\cap B)}{n(A)}$$

$$F_1\text{-}score=\frac{2\times R\times P}{R+P} \quad (4)$$

- $n(A)$: 정답집합 A의 원소 개수
- $n(B)$: 결과집합 B의 원소 개수

실험은 제목/비제목 구분에 대한 실험, 제목 개체명 유형 분류에 대한 실험, 두 가지를 합친 전체 실험 순으로 진행하였다.

4.1 제목 식별 실험 (단계 1)

제목 식별을 위하여 특수기호 사이 어절이 제목 개체명인 것과 그렇지 않은 것을 대상으로 SVM^{light}를 이용하여 학습하였다. 또한, 좌/우에 공기하는 문맥 단어의 개수(윈도우 크기)의 변화에 따른 제목 개체명 인식 성능을 비교하였다.

특히, 학습을 위하여 제목/비제목 문장을 추출할 때 제목에 대한 학습 데이터로 모든 학습 말뭉치를 사용한 경우와, 제목과 관련된 카테고리에 속하는 문장만 사용한 경우를 나누어 실험하였으며, 결과는 표 5와 같다.

Table 5. The performance of title identification (F_1-score)

Training corpus \ Window size	1	2	3	4	5
Title related category documents	**84.47**	83.63	80.60	81.55	81.48
All category documents	**81.59**	80.72	78.65	78.77	78.96

제목 좌/우의 문맥 단어를 각 1개씩 추출하였을 때 가장 높은 성능을 나타냈다. 이는 제목의 바로 앞 또는 뒤에 제목과 관련된 핵심 단어가 나타난다는 것을 의미한다. 또한 좌/우 단어를 많이 사용할수록 제목과 관련 없는 단어들이 포함될 가능성이 커져서 오히려 제목 구분에 대한 성능의 저해요소로 작용한다. 더욱이 제목 유형을 대상으로 학습시켰을 때 성능이 향상되었는데, 이는 제목과 관련된 카테고리에 속한 뉴스 기사문에서 좌/우 문맥 단어를 추출했을 때, 제목과 가장 관련 있는 문맥 단어가 추출되었기 때문으로 판단된다.

4.2 제목 개체명 유형 분류 실험 (단계 2)

제목 개체명 유형 분류는 특수기호 사이 어절이 모두 도서, 영화, TV프로그램의 제목 개체명인 것을 가정한 후, 개체명 유형 분류에 대한 성능을 측정하였다.

제목 개체명 유형에 대한 문맥 단어의 상호정보량을 가중치로, $SVM^{multiclass}$를 사용하여 세 가지 제목 개체명 분류를 위한 학습과 실험을 하였으며, 표 6과 같은 성능을 보였다.

Table 6. The performance of title named entity type classification

Window Size	1	2	3	4	5	6	7	8	9	10
Precision	84.88	87.77	86.44	86.44	90.96	92.09	90.96	89.83	87.57	87.57

4.3 제목 개체명 인식(통합 모델) 및 사전 구축 실험

표 5와 표 6에서 살펴본 것과 같이, 제목/비제목 구분에 있어서는 좌/우 문맥 단어 각 1개, 제목 개체명 유형 분류에 있어서는 각 6개를 추출했을 때 가장 높은 성능을 나타냄을 확인하였다. 따라서 실험 1단계에서의 제목/비제목 구분과 실험 2단계에서의 제목 개체명 유형 분류에 대한 실험은 성능이 가장 높았던 좌/우 문맥 개수로 고정하였으며, 학습 시 제목의 유형에 따른 카테고리도 고려하였다. 전체 통합 실험의 결과는 표 7과 같다.

Table 7. Title named entity recognition(integration model)

	Recall	Precision	F_1-score
Performance	72.82%	82.56%	77.38%

표 8은 본 논문의 실험 방법과 기존의 제목 개체명 추출에 대한 방법을 비교한 것이다. 실험 데이터 및 방법의 차이로 인하여 직접적인 비교는 어렵지만, 기존의 방법들과 비교할 때 우수한 성능을 보여주었다. 특히, [1], [15]에 비하여 높은 성능을 보였으며, [16]에 비하여 다소 낮은 성능을 보였으나, 본 논문에서 제안한 방법은 정규화된 뉴스 기사문을 이용하여 제목 개체명의 경계 인식을 간략화하였으며, 실시간 개체명 사전 확장이 가능하다는 장점을 지니고 있다.

Table 8. The comparison of the performances

Author	Experimental methods and types of title named entity	Experimental data	F1-score
Lai, A. (2009)[15]	Naive Bayes classifier - Extract movie-related documents Maximum Entropy Markov Model - Extract title of movie	E-mail	13.33%
Lee, J. Y. et al. (2005)[1]	Using constructed context pattern and title entity dictionary - Extract title of movies, books, music	NY Times	55.16%
Park, Y. M. et al. (2013)[16]	Using CRFs with dictionary of title and title abbreviations - Extract title of movies, music, cartoon, novel etc.	Wikipedia	82.10%
Proposed method	Using SVM with word contexts - Extract title of books, movies, music, TV programs	Korean news articles	77.38%

또한 제목 개체명 사전을 구축했을 경우에 대한 성능도 측정하였다. 제목의 띄어쓰기 일관성을 위해 제목에 포함된 공백을 모두 없앤 후 같은 유형의 제목 개체명에 속하는 동일한 제목은 하나로 합쳐서 사전을 구성하였다. 사전 구성에 대한 성능은 표 9와 같다.

Table 9. The performance of title named entity dictionary construction

	Recall	Precision	F_1-score
Performance	72.90%	79.58%	76.09%

하나의 뉴스 기사문에서 특수기호 사이에 존재하는 제목 후보가 2회 이상 출현한 것만을 대상으로 추가 실험을 하였다. 실험 데이터 200개 문서 중 문서 내에 동일한 제목 개체명을 2개 이상 포함하는 문서는 43개였으며, 전체 200개 문서 내 중복을 제외한 제목 개체명의 개수는 43개였다. 정답집합은 실험과 같은 방법으로 2회 이상 출현한 제목만으로 구성하였으며, 성능은 표 10과 같다.

Table 10. Title named entity recognition considering the number of appearance

Appearance \ Performance		Recall	Precision	F_1-score
More than one times	Integration model	84.00%	91.30%	87.50%
	Dictionary	80.95%	89.47%	85.00%

표 10의 실험은 문서 내에 한 번만 출현하는 제목은 추출할 수 없지만, 두 번 이상 출현하는 제목에 대해서는 정확률이 증가하여 사전으로 구성하여 활용하기에 비교적 양호함을 알 수 있었다.

5. 결론 및 향후연구

개체명 인식은 정보검색 시스템, 질의응답, 기계번역 등에 있어서 필수적인 기능이며, 각 시스템의 성능 향상에 큰 역할을 한다. 특히, 제목 개체명과 같이 문장으로 구성되어 있거나, 3개 이상의 많은 어절로 이루어진 개체명은 사전으로 구성한 후, 사전 매칭을 이용하는 것이 효과적이다.

본 논문에서는 뉴스 기사문에서 도서, 영화, 음악, TV프로그램에 해당하는 새로운 제목 개체명을 추출하여 사전을 구축하는 방법을 제안하였다.

뉴스 기사문의 특성상 특수기호 사이에 등장하는 제목 개체명을 대상으로 좌/우에 공기하는 문맥 단어에 위치 가중치와 빈도 가중치를 적용하여 SVM으로 학습하였으며, 제목과 비제목의 구분은 85.87%(F_1-score)의 성능을 나타내었다.

제목 개체명의 유형 분류는 제목 개체명의 좌/우 문맥 단어와 각 개체명 유형 사이의 상호정보량을 가중치로 SVM을 이용해 학습하였으며, 92.09%의 정확률을 보였다.

본 논문은 정규화된 형식을 갖춘 뉴스 기사문을 이용하여 제목 개체명의 경계 인식을 간략화하였다. 또한 뉴스 기사문을 대상으로 제목 개체명 인식 및 사전을 구축하였기 때문에, 실시간으로 생성되는 뉴스를 활용하여 빠르게 제목 개체명 사전을 확장시킬 수 있다. 하지만 제목과 비제목 구분에 대한 성능이 제목 개체명 유형 분류 성능에 비하여 부족함을 알 수 있다. 이로 인해 전체 성능은 78.67%(F_1-score), 사전 구성 성능은 76.09%(F_1-score)로 사전을 개체명 인식에 사용하기 위해서는 추가적인 정제과정이 필요하다.

따라서 본 논문에서는 사전 구성의 정확률을 높이기 위하여 뉴스 기사문 내에 2회 이상 등장하는 제목 후보들만을 이용하여 제목 개체명 인식 및 사전 구성을 하였다. 전체 제목 개체명으로 이루어진 정답 집합으로 평가하였을 때 사전 구성 성능은 재현율(Recall) 21.94%, 정확률(Precision) 89.47%, F_1-score 35.23%이고, 2회 이상 출현한 제목 개체명으로 이루어진 정답 집합으로 평가하였을 때, 사전 구성 성능은 재현율(Recall) 80.95%, 정확률(Precision) 89.47%, F_1-score 85.00%이었다.

향후 연구에서는 정제된 말뭉치를 이용하여 학습에 활용하고, 제목과 비제목의 구분을 위한 가중치 변경과 학습 데이터 크기의 불균형을 해결한다면 사전 구성의 성능을 높일 수 있을 것으로 보인다.

Reference

[1] Seong-Won Kim, Dong-Yul Ra, “Korean Named Entity Recognition Using Two-level Maximum Entropy Model,”

Proc. of the KIISE Symosium, Vol.2, No.1, pp.81-86, 2008.
[2] Changki Lee, Myung-Gil Jang, "Named Entity Recognition with Structural SVMs and Pegasos algorithm," Proc. of KSCS Congnitive Science, Vol.21, No.4, pp.655-667, 2010.
[3] Joo-Young Lee, Young-In Song, Hae-Chang Rim, "Title Named Entity Recognition based on Automatically Constructed Context Patterns and Entity Dictionary," Proc. of the KIISE Conference, The 16th Annual Conference on Human & Cognitive Language Technology, pp.40-45, 2004.
[4] Black, W., F. Rinaldi and D. Mowatt, "Facile: Description Of The Ne System Used For Muc-7," in Proceedings of the 7th Message Understanding Conference, 1998.
[5] Chen H., Y. Ding, S. Tsai and G. Bian, "Description of the NTU System Used for MET2," in Proceedings of 7th Message Understanding Conference, 1998.
[6] Aberdeen, J., J. D. Burger, D. S. Day, L. Hirschman, P. Robinson and M. B. Vilain, "MITRE : Description Of The Alembic System Used For MUC-6," in Proceedings of 6th Message Understanding Conference, pp.141-155, 1995.
[7] Borthwick, A., J. Sterling, E. Agichtein and R. Grishman, "NYU : Description of the MENE Named Entity System as Used in MUC-7," in Proceedings of 7th Message Understanding Conference, 1998.
[8] Merchant, R. and M. E. Okurowski, "The multilingual entity task (MET) overview," in Proceeding TIPSTER'96 Proceedings of a workshop on held at Vienna, pp.445-447, 1996.
[9] Sekine, S. and Y. Eriguchi, "Japanese named entity extraction evaluation : analysis of results," in Proceeding COLING'00 Proceedings of the 18th conference on Computational linguistics - Vol.2, pp.1106-1110, 2000.
[10] Kyung Hee Lee, Ju Ho Lee, Myung Seok Choi, Gil Chang Kim, "Study on Named Entity Recognition in Korean Text," Proc. of the KIISE Conference, The 12th Annual Conference on Human & Cognitive Language Technology, pp.292-299, 2000.
[11] Yi-Gyu Hwang, Hyun-Sook Lee, Eui-Sok Chung, Bo-Hyun Yun, Sang-Kyu Park, "Korean Named Entity Recognition Based on Supervised Learning Using Named Entity Construction Principles," Proc. of the KIISE Conference, The 14th Annual Conference on Human & Cognitive Language Technology, pp.111-117, 2002.
[12] Hae-Suk Jang, Kyu-Cheol Jung, Jin Kwan Lee, Kihong Park, "Recognition of Korean Place Names on the Internet by Using the Rules of Dictionary Use," Proc. of the KSII Fall Conference, Vol.6, No.1, pp.397-400, 2005.
[13] Yi-Gyu Hwang, Bo-Hyun Yun, "HMM-based Korean Named Entity Recognition," Proc. of the KIPS Transaction Vol.10(B), No.2, pp.229-236, 2003.
[14] Changki Lee, Yi-Gyu Hwang, Hyo-Jung Oh, Soojung Lim, Jeong Heo, Chung-Hee Lee, Hyeon-Jin Kim, Ji-Hyun Wang, Myung-Gil Jang, "Fine-Grained Named Entity Recognition using Conditional Random Fields for Question Answering," Proc. of the KIISE Conference, The 18th Annual Conference on Human & Cognitive Language Technology, pp.268-272, 2006.
[15] Lai, A., "Movie Title Recognition in E-Mail," Stanford University Natural Language Processing, CS224N Final Project, 2009.
[16] Young-Min Park, Sang-woo Kang, Byoung-Kyu Yoo, Jung-Yun Seo, "Title Named Entity Recognition using Wikipedia and Making Acronym," Proc. of the KIISE Korea Computer Congress, pp.637-639, 2013.
[17] Vapnik, V. N., The nature of statistical learning theory, Springer, 1995.
[18] Dumais, S., J. Platt and D. Heckerman, "Inductive Learning Algorithms and Representations for Text Categorization," in Proceeding of ACM-CIKM '98, pp.148-155, 1998.
[19] Crammer, K., Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," Journal of Machine Learning Research 2, pp.265-292, 2001.
[20] Peng H., F. Long and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max- Dependency, Max-Relevance, and Min-Redundancy," Pattern Analysis and Machine Intelligence, IEEE Transactions on Vol.27, Issue 8, pp.1226-1238, 2005.

박 용 민
e-mail : yongmin@cbnu.ac.kr
2011년 충북대학교 컴퓨터교육과(학사)
2014년 충북대학교 디지털정보융합학과 석사
관심분야 : 정보 검색, 인공지능, 자연 언어 처리

이 재 성
e-mail : jasonlee@cbnu.ac.kr
1983년 서울대학교 컴퓨터공학과(학사)
1985년 KAIST 전산학과(석사)
1999년 KAIST 전산학과(박사)
1985년~1988년 큐닉스 컴퓨터 과장
1988년~1993년 미국 및 한국 마이크로소프트 개발부 차장
1988년~1993년 미국 및 한국 마이크로소프트 개발부 차장
1999년~2000년 ETRI 컴퓨터소프트웨어기술연구소 팀장
2005년~2006년 (美)아리조나 대학 방문 교수
2000년~2011년 충북대학교 컴퓨터교육과 교수
2011년~2014년 충북대학교 디지털정보융합학과 교수
2014년~현 재 충북대학교 소프트웨어학과 교수