http://dx.doi.org/10.14400/JDC.2015.13.5.187

언어자원 자동 구축을 위한 위키피디아 콘텐츠 활용 방안 연구

류철중*, 김 용**, 윤보현***
전북대학교 소프트웨어공학과*, 전북대학교 문헌정보학과**, 목원대학교 컴퓨터교육학과***

A Study on Utilization of Wikipedia Contents for Automatic Construction of Linguistic Resources

Cheol-Jung Yoo*, Yong Kim, Bo-Hyun Yun*****
Dept. of Software Engineering, Chonbuk National University*
Dept. of Library & Information Science, Chonbuk National University**
Dept. of Computer Science Education, Mokwon University***

요 약 급변하는 자연언어를 기계가 이해할 수 있도록 하기 위해서는 다양한 언어지식자원(linguistic knowledge resources)의 구축이 필수적으로 수반된다. 본 논문에서는 온라인 콘텐츠의 특성을 활용해 언어지식자원을 자동으로 구축함으로써 지속적으로 확장 가능한 방법을 고안하고자 한다. 특히 언어분석 과정에서 가장 활용도가 높은 개체명(NE: Named Entity) 사전을 자동으로 구축, 확장하는데 주안점을 둔다. 이를 위해 본 논문에서는 개체명 사전 구축 대상문서로 위키피디아(Wikipedia)를 선정, 그 특성을 파악하기 위해 다양한 통계 분석을 수행하였다. 이에 기반하여 위키피디아 콘텐츠가 갖는 구문적 특성과 구조 정보 등의 메타데이터를 활용하여 개체명 사전을 구축, 확장하는 방법을 제안한다.

주제어 : 언어자원 구축, 위키피디아, 개체명 사전, 지식구축, 온라인 콘텐츠 활용

Abstract Various linguistic knowledge resources are required in order that machine can understand diverse variation in natural languages. This paper aims to devise an automatic construction method of linguistic resources by reflecting characteristics of online contents toward continuous expansion. Especially we focused to build NE(Named-Entity) dictionary because the applicability of NEs is very high in linguistic analysis processes. Based on the investigation on Korean Wikipedia, we suggested an efficient construction method of NE dictionary using the syntactic patterns and structural features such as metadatas.

Key Words : Linguistic Resource Construction, Wikipedia, Named-Entity Dictionary, Knowledge Construction, Utilization of online contents

* 본 연구는 2014년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음
Received 17 March 2015, Revised 22 April 2015
Accepted 20 May 2015
Corresponding Author: Bo-Hyun Yun
(Dept. of Computer Science Education, Mokwon University)
Email: ybh@mokwon.ac.kr

ISSN: 1738-1916

1. 서론

21세기에 들어 컴퓨터와 인터넷의 발달과 더불어 스마트 단말의 대량 보급으로 인해 다양한 사용자로부터 매일같이 새로운 형태와 의미를 갖는 신조어를 비롯한 채팅어 등의 다양한 언어 변이(language variation)들이 발생하고 있다. 이같이 급변하는 자연언어(natural languages)를 기계가 이해할 수 있도록 하기 위해서는 전자사전을 비롯한 시소러스(thesaurus)[1], 개념망(concept network)[2], 온톨로지(ontology) 등의 다양한 언어지식자원(linguistic knowledge resources)[3]의 구축이 필수적으로 수반된다.

본 논문에서는 이러한 언어지식자원을 온라인 콘텐츠의 특성을 활용해 자동으로 구축함으로써 지속적으로 확장 가능한 방법을 고안하고자 한다. 특히 언어분석 과정에서 가장 활용도가 높은 개체명(NE: Named Entity) 사전을 자동으로 구축, 확장하는데 주안점을 둔다. 개체명이란 인명, 지명, 기관명, 날자, 시간 등 문장에서 핵심적인 의미를 지닌 고유명사나 미등록어 등을 말하는 것으로[4,5], 개체명 사전은 해당 개체명과 분류 태그(tag)로 구성되어 있다(예: 홍길동:사람).

텍스트로부터 개체명을 자동으로 인식하기 위한 기술(NER: Named Entity Recognition)은 1990년대부터 본격적으로 시작되었다[6,7]. 초창기 연구에서는 단순히 수작업으로 워드넷(WordNet)과 같은 사전을 탐색하거나 정규표현 등을 활용하는 방법[8,9,10]에서 최근에는 주로 통계기반의 기계학습(machine learning) 방법인 은닉 마코프 모델(Hidden Markov Model), 최대 엔트로피 모델(Maximum Entropy Model), 지지벡터머신(Support Vector Machines) 기법[11] 등의 학습 모델을 생성하는 방향으로 진행되고 있다.

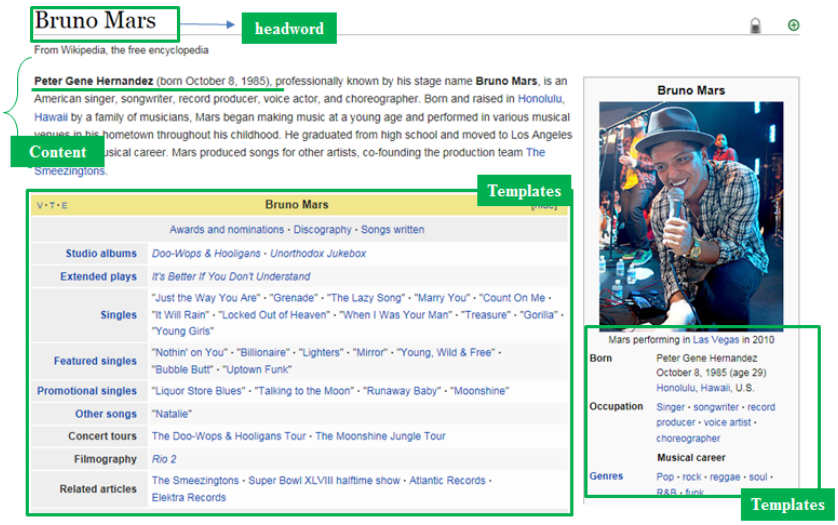
그러나 이와 같은 방법을 적용하여 높은 성능을 내기 위해서는 많은 양의 코퍼스(corpus)를 필요로 하며, 그에 따른 수작업 비용을 요구한다. 뿐만 아니라 많은 양의 코퍼스를 구축하였다 하더라도, 새로운 도메인에 최적화된 개체명 인식기를 개발하기 위해서는 새로운 코퍼스가 필요하기 때문에 이러한 교사기반(supervised) 기계학습 기법은 확장성이 떨어진다.

특히 본 논문에서는 새로운 매체의 등장과 생성되는 많은 양의 데이터에서 새로운 개체명이 꾸준히 만들어지고 있는 현실을 반영하여, 온라인 미디어로부터 집단지성을 통해 구축된 정보를 활용하여 신규 개체명 후보를 수집, 개체명 사전을 확장하는 방안을 모색하고자 한다. 제안된 방법은 온라인상의 방대한 양의 신규 콘텐츠를 통해 스스로 패턴이나 사전을 확장시켜 가면서 발전 가능한 형태의 지속적 언어자원 구축이 가능하게 한다.

이를 위해 본 논문에서는 개체명 사전 구축 대상문서로 위키피디아(Wikipedia)를 선정하였다. 한국어 위키피디아[12]는 사용자가 직접 참여하여 생성, 수정해나가는 오픈 지식 자원으로, 현재 306,190개의 표제어(headword)로 구성되어 있다(2015년 2월 말 현재 기준). 위키피디아를 활용해 개체명 사전을 확장하고자 하는 연구는 일부 시도된 적이 있는데, 첫째는 WordNet을 이용한 방법[13]이다. 위키피디아 표제어의 분류를 위해 WordNet을 활용하여 개체명을 인식한 경우로, 이를 개체명 사전의 엔트리로 활용하기 위해서는 수작업으로 확인하는 일일이 확인, 검수하는 과정이 필요하다. Bunescu[14]과 Kazama[15] 등은 표제어의 대문자나 '.' 문자를 활용하여 약자 및 이형태 축약어 등을 찾는 방법을 고안하였는데, 이는 영어에 국한된 방법으로 한국어에는 이러한 표층적 특성이 부족해 적용하기 어려운 난점이 있다. 본 논문에서는 이러한 단점을 보완하기 위해 한국어 위키피디아가 갖는 구문적 특성과 집단지성을 통해 구축된 구조적 정보를 활용해 개체명 사전을 자동으로 구축, 수작업을 최소화하여 확장하기 위한 방법을 모색하고자 한다.

2. 위키피디아 콘텐츠 특성 분석

지속적인 언어자원 구축을 위해서는 적용할 대상문서 집합의 특성 파악이 필수적이다. [Fig. 1]은 위키피디아 문서 중 "브루노 마스(Bruno Mars)"와 관련된 문서 예제로, 해당 문서는 <Table 1>과 같은 문서 기본 정보(Page Information)와 <Table 2>와 같은 문서 속성(Page Properties)으로 구성되어 있다. 자세히 살펴보면, '브루노 마스'라는 표제어를 설명하는 '본문(contents)'과 표제어별 특성에 따른 정형화된 정보를 '표(table)'형식 (위키피디아에서는 '틀(template)'이라는 용어를 씀)으로 제공한다. 모든 위키피디아 문서는 <Table 1>과 같은 기본 정보를 갖는데, 보여줄 이름, 정렬 키, 문서 길이나 ID와 같은 정보뿐 아니라 '주시하는 사용자 수'와 같은 인기도와 언어정보, 영어문서와 매핑되는 공통 ID를 제공한다.



[Fig. 1] An Example of Wikepedia Contents (headword: Bruno Mars)

〈Table 1〉 Wikipedia Page Information

Tag	Value
Display Title	Bruno Mars
Default sort key	Mars, Bruno
Page length (in bytes)	38,102
Page ID	608480
Page Content Language	Korean (ko)
Page Content Model	wikitext
# of page watchers	275
wikidata item ID	Q1450
Page Image	

〈Table 2〉 Wikipedia Page Properties

Tag	# of types	Value
# of redirects to this page	9	- Peter Hernandez - Bruno Hernandez - Peter Gene Bayot Hernandez - Peter G. Hernandez ...
Hidden categories	8	- Category:Articles with hAudio microformats - Category:Articles with hCards - Category:Wikipedia articles with MusicBrainz identifiers..
Transcluded templates	119	- Template:**Age** - Template:All**music** - Template:Authority control - Template:BillboardChartNum - Template:BillboardEncode ...

〈Table 3〉 Wikipedia Templates

Tag	Value
Autonmy	Peter Gene Hernandez
Born	October 8, 1985 (age 29) Honolulu, Hawaii, U.S.
Occupation	**Singer**· songwriter· record producer· voice artist·...
Genres	Pop· rock· reggae· soul· R&B· funk...
Instruments	Vocals· drums· guitar ..
Years Active	2004 - present
Labels	Universal Motown· Atlantic· Elektra...

기본정보 외에 표제어의 특성에 따른 부가정보가 제공되는데, <Table 2>와 같이 '속성(Properties)'이라는 이름으로 제공된다. '브루노 마스'의 경우, 관련된 이형태 정보로 총 9개를 제공하는데, 외국 인명의 경우 발음의 차에 따른 변형정보를 함께 제공함으로써 검색의 효율을 높인다.

특히 개체명 인식을 위해 중요한 자질과 활용된 정보는 '분류'와 '틀' 정보로, '브로노 마스'의 경우 <Table 3>과 같은 '틀' 정보와 연결되어 있다. 그 밖에도 분류 정보를 통해 '브로노 마스'가 인명이며 그 중에서도 가수 직업군에 해당함을 유추할 수 있다.

<Table 4>는 위키피디아 문서 집합에 대한 통계 정보이다. 위키피디아는 약 30만 건의 문서로 전체 1,500만 개의 문장으로 구성되어 있다. 이는 평균 표제어 당 약 41개의 설명문으로 구성되어 있음을 알 수 있다. 한국어 위키피디아 내의 '틀' 정보 약 10만개로(<Table 5>) 전체 표제어의 절반에 못 미치는 정보이다. 그러나 '틀' 정보를 가지는 표제어의 특성이 주로 인물, 작품명, 지역 이름 등에 국한된 점과 이들 도메인이 개체명 분야에서 차지하는 비율이 크다는 점을 고려하면 '틀' 정보를 활용한 개체명 사전 구축 방법의 효용성이 매우 높음을 유추할 수 있다.

〈Table 4〉 Wikipedia Pages

Item	Quantity
Numver of Pages	306,190
Size of Pages	2.2 GB
Number of Sentence	12,554,790

〈Table 5〉 Wikipedia Templates

Item	Quantity
Number of Templates	106.612
Size of Templates	0.42 GB
Number of Informa	1,153,279

3. 개체명 사전 자동 구축을 위한 위키피디아 콘텐츠 활용

2절에서 분석된 위키피디아 문서 특성을 반영하여 개체명 사전 자동 구축 방법은 크게 해당 문서 내의 정보(inner-info)를 활용하는 방법과 문서와 문서 사이의 링크 정보 등 문서 밖의 정보(outer-info)를 활용하는 방법으로 나뉠 수 있다. 문서 내의 특성을 활용하는 방법은 다시 자연어 문장의 '패턴(pattern)'을 이용하는 방법과 정형화된 '구조정보(structure)'를 참조하는 방법으로 나뉜다.

3.1 Inner-info1: 해당 표제어의 설명 문장의 '패턴' 활용

아래는 문장은 [Fig. 1]의 '브루노 마스' 표제어의 첫 번째 문장으로, 대부분의 위키피디아 문서의 첫 문장은 해당 표제어의 정의를 설명하는 설명문이다.

브루노 마스(Bruno Mars, 1985년 10월 8일 ~)는 미국의 싱어송라이터이자 음악 프로듀서로 본명은 피터 진 에르난데스(Peter Gene Hernandez)이다

위 예제에 나타나듯, 대부분의 인명의 경우 첫 번째 문장이

[인명 (영문, 날짜정보 ~ 날짜정보)]

의 형식으로 구성된다. 이를 통해 첫 번째 문장이 위와 같은 형식으로 구성되어 있다면 해당 표제어는 인명이고, '( )' 내의 영문은 '영어 인명' 정보이며, 영문 뒤의 숫자는 '날짜' 정보에 해당하는 개체명이라고 역으로 유추할 수 있다. 궁극적으로 본 연구의 목표는 이러한 문장 패턴을 학습, 추출된 명사구가 개체명인지 여부와 어떤 개체명 태그에 해당하는지 개체명 분류코드를 유추함으로써 지속적 언어자원 학습이 가능한 모델을 개발하고자 한다.

그 밖에도 책, 영화, 노래 등 작품의 이름은 '〈〉', '《》' 등의 문장부호로 구별되어 있는데, 본문에는 <Table 6>와 같은 쓰임 규칙을 갖는 것으로 분석되었다.

그 외에도 작품 제목을 표기할 때에는 보통 따옴표(' "), 낫표(「 」 『 』), 꺾쇠표(〈 〉 《 》), 부등호(< >

〈Table 6〉 Usages of Punctuation Marks

Characters	Punctuation Marks in Korean	Usage	Examples
〈 〉	angle brackets	Titles of a Thesis	<Good Samaritan laws> offer legal protection to people who give reasonable assistance to those who are injured, ill, in peril, or otherwise incapacitated.
《 》	double angle brackets	Titles of Books or Separated volumes	《Romeo and Juliet>> is a tragedy written by William Shakespeare early in his career.
' '	Single quotation marks	quoted words or phrases	He shared his 'wisdom' with me.
" "	Double quotation marks	quoted of sentences	"To be, or not to be..." is the opening phrase of a dialog in the Nunnery Scene of William Shakespeare's play Hamlet.

<< >>) 등을 사용하는데, 아래와 같은 형식으로 사용되는 특성이 있다.

- **한글 제목에 겹꺾쇠표를 사용하는 작품의 예**
  - 단행본 이상 규모의 문헌 - 《다빈치 코드》
  - 희곡 - 《햄릿》
  - 컴퓨터 게임과 비디오 게임 - 《바람의 나라》
  - 영화 - 《올드보이》
  - 장편시, 서사시 - 《오디세이아》
  - 음반 - 《백아절현》
  - 교향곡, 오페라 등 - 《핀란디아》
  - 텔레비전 시리즈 - 《겨울연가》
  - 시각예술 작품 - 《게르니카》

- **한글 제목에 홑꺾쇠표를 사용하는 작품의 예**
  - 논문, 기사(일회성 보고서 포함) 따위 - 〈시일야방성대곡〉
  - 더 긴 작품에서 따온 한 장(章)
  - 짧은 시 - 〈황조가〉
  - 노래 - 〈가고파〉
  - 단편소설 - 〈메밀꽃 필 무렵〉

- **한자나 외국어 병기시의 예**
  - 《태백산맥》(太白山脈)
  - 《레 미제라블》(Les Misérables)

3.2 Inner-info2: 해당 표제어의 정형화된 '구조정보' 를 활용하는 방안

앞서 2장에서 기술한 바와 같이 위키피디아는 표제어별로 <Table 2>와 <Table 3>과 같은 정형화된 구조정보를 제공한다. 이러한 구조정보는 개체명의 태그를 규명하기 매우 좋은 단서로 활용될 수 있는데, '브루노 마스'의 경우, 아래와 같은 '틀' 정보를 제공한다.

틀: ISO 이름 틀: 나이 틀: SKFW 틀: 음악가 정보 틀: 출생일 틀: 유투브 채널 정보 …

자세히 살펴보면, '틀:나이', '틀:음악가 정보' 등과 같이 '인명'과 관련된, 그 중에서도 '음악가'들의 정보를 요약해 놓은 것임을 알 수 있다. 뿐만 아니라 위키피디아에서는 표제어의 '분류정보'를 제공하는데, '브루노 마스'의 경우 다음과 같은 분류 정보와 연결되어 있다.

분류정보: 1985년 태어남 / 살아있는 사람 / 미국의 남자 가수/ 미국의 R&B 가수 / 미국의 팝 가수 / 미국의 소울 가수 / 미국인 그래미상 수상자…

위 정보를 통해 '브로노 마스'가 인명이며 그 중에서도 '음악가' 이며, 특히 '가수' 직업군에 해당함을 유추할 수 있다. 이와 같이 주어진 표제어 문서가 어떤 '틀' 정보와 '분류' 정보를 가지고 있는지에 파악함으로써, 역으로 해당 표제어의 개체명 여부와 어떤 개체명 태그에 해당하는지 유추가 가능하게 된다.

3.3 Outer-info: 문서 간의 메타데이터 정보를 활용하는 방법

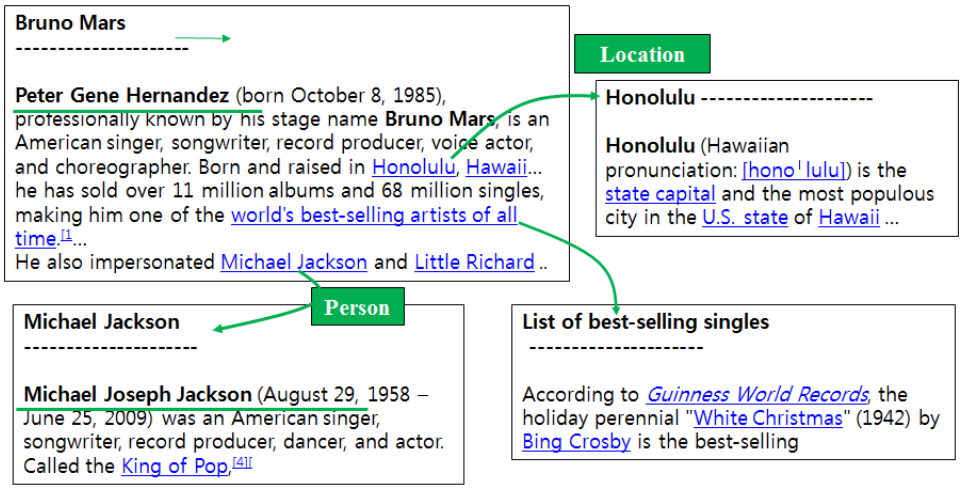
위 방법들은 해당 문서 내의 정보를 활용하는 방법인 반면, 이 방법은 문서와 문서 사이의 링크나 해당 문서와 연결된 '분류'와 '틀'의 종류 등 문서 밖의 메타데이터 정보을 활용하는 방법으로, [Fig. 2]를 통해 개체명 후보 인식 과정을 상세 설명하기로 한다.

[Fig. 2]는 [Fig. 1]의 '브루노 마스' 문서와 연결된 문서를 활용해 개체명 후보들을 추출하는 예이다. 개체명 후보는 문서 내의 세부 설명문에 연결된 링크의 시작점으로 본 논문에서는 앵커 텍스트(anchor text)라 명명하기로 한다. [Fig. 2]의 경우, '호놀롤루', '마이클 잭슨', '베스트 셀링 아티스트' 등이 개체명 후보 대상이다.

세부적으로 살펴보면, '호놀룰루'의 경우 링크를 통해 찾아간 '호놀룰루'표제어에 관한 설명문을 통해 해당 표제어가 '주도'임을 알 수 있다. 뿐만 아니라 '하와이 주의 도시'라는 분류 정보를 참조함으로써 해당 단어가 '도시'의 이름임을 명확히 구별할 수 있다.

'마이클 잭슨'의 경우를 살펴보면, '브루노 마스' 표제어의 설명문과 똑같은 패턴의 설명문으로 시작되는 것을 알 수 있다.

[인명 (영문, 날짜정보 ~ 날짜정보)]
마이클 잭슨(Michael Joseph Jackson, 1958년 8월 29일 ~ 2009년 6월 25일)



[Fig. 2] Named Entity Candidate Extraction based on Anchor Information

뿐만 아니라 '미국의 팝 음악가', '그래미상 수상자'와 같은 분류 정보를 통해 해당 단어가 인명이자 그 중에서도 '음악가'임을 유추할 수 있다.

그러나 모든 앵커 텍스트가 '개체명'인 것은 아니다. [Fig. 2]의 '베스트 셀러 아티스트'의 경우, 해당 표제어의 설명문이 별다른 특징이 없을 뿐 아니라 관련 '분류' 정보나 '틀' 정보가 없는 것으로 볼 때 해당 명사구가 개체명은 아닌 것으로 유추 할 수 있다.

본 논문에서 제안한 방법의 실효성과 효과를 입증하기 위해 자동으로 구축한 개체명 사전을 수작업으로 검수, 그 품질을 측정하는 과정이 필요하다. 본 논문에서는 이 제안된 방법 중 문서 간의 메타데이터 정보를 활용하는 방법의 효과를 입증하기 위해서 위키피디아 문서 중 10%(를 임의로 선정, 표제어와 앵커 텍스트의 개체명 추출 결과를 분석하였다. <Table 7>은 제안된 방법을 통해 선정된 개체명 사전의 정확도이다. 실험 집합 30,245개의 문서에 포함된 전체 앵커 수는 604,823개이다. 이 중에서 '개체명'에 해당하는 대상 앵커 수는 9,753개로 이번 실험의 목적은 제안된 방법을 통해 대상 앵커가 개체명인지 여부 및 개체명 태그를 자동으로 분류하는지 여부이다. 실험에 사용한 개체명 태그는 한국전자통신연구원에서 제안한 개체명 범주 체계 상위 체계를 참조하였다[16].

실험 결과, 전체 9,753개 중 88.9%인 8,666개의 앵커를 '개체명' 후보로 인식하였고, 그 중에서 개체명 태그까지 정답인 경우는 전체 76.4%로 7,452개로 평가되었다.

〈Table 7〉 Accuracy of Automatic Construction Results

Item	Result
Number of Test Pages	30,245
Number of Anchors	604,823
Number of Target Anchors	9,753
Number of extracted NEs	8,666 (88.9%)
Number of corrected NE Tags	7,452 (76.4%)

4. 결론 및 향후 연구 방향

기존의 언어자원 구축을 위해 기계학습 방법을 적용하는 경우, 높은 성능을 내기 위해서는 많은 양의 코퍼스를 필요로 하며, 그에 따른 비용이 요구된다. 또한 많은 양의 코퍼스를 구축하였다 하더라도, 새로운 도메인에 최적화된 개체명 인식기를 개발하기 위해서는 새로운 코퍼스가 필요한 단점이 있었다.

본 연구에서는 신규 언어지식을 지속적으로 학습할 수 있는 적응형 기계학습 방법을 고안하기 위한 것으로, 온라인 미디어로부터 집단지성을 통해 구축된 정보를 활용하여 신규 개체명 후보를 수집, 개체명 사전을 확장하는 방안을 모색하였다. 이를 위해 본 논문에서는 개체명 사전 구축 대상문서로 위키피디아를 선정, 그 특성을 파악하기 위해 다양한 통계 분석을 수행하였다. 그 결과 위키피디아 콘텐츠가 갖는 구문적 특성과 구조 정보 등 메타데이터를 활용해 개체명 사전을 구축, 확장하기 위한

방법을 제안하였다.

제안된 방법은 온라인상의 방대한 양의 신규 콘텐츠를 통해 스스로 패턴이나 사전을 확장시켜 가면서 발전 가능한 형태의 지속적 언어자원 구축이 가능할 것으로 예상되며 향후 연구방향으로는 제안된 방법을 실제 적용하여 그 효과를 입증하고자 한다. 특히 수작업으로 구축한 사전과의 품질 및 비용을 비교함으로써 제안된 방법의 효율성을 검증하고자 한다. 또한 SVM 등의 기계학습 기반의 언어분석 결과[16]에 제안된 방법을 병합하여 향상되는 효과 분석을 수행하고자 한다. 나아가 최근 대두되고 있는 비교사학습 (unsupervised learning) 방법인 딥러닝 (deep-learning) 기법[17]을 설명문의 문장 패턴 분석에 적용하고자 한다.

ACKNOWLEDGMENTS

This paper was supported by research funds of Chonbuk National University in 2014.

REFERENCES

[1] Michael Scriven. Evaluation thesaurus. UK: Sage Press, 1991.

[2] V. Nastase, M. Strube, B, Boerschinger, C. Zim, A. Elghafari, WikiNet: A Very Large Scale Multi-Lingual Concept Network. Proc. of LREC. pp.1015-1022, 2010.

[3] Y. J. Bae, C. Y. Ok, Semantic Analysis of Korean Compound Noun using Lexical Semantic Network(U-WIN). Journal of KIISE: Software and Applications, pp.833-847, 2013.

[4] T. J. Kim, E. Sang, D. M. Fien, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Proc. of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, pp.142-147, 2003.

[5] Y.M. Park, J. S. Lee, Named Entity Recognition and Dictionary Construction for Korean Title: Books, Movies, Music and TV Programs, KIPS Tr. Software and Data Eng. Vol. 3, No. 7, pp.285-292, 2014.

[6] A. Mikheev, C. Grover, M. Moens, Description of the LTG System Used for MUC-7. Proc. of MUC-7. pp.1-8 1998.

[7] S. Brin, Extracting Patterns and Relations from the World Wide Web. Proc. of the International Workshop on The World Wide Web and Databases, pp.172-183, 1998.

[8] M. Negri, B. Magnini, Using wordnet predicates for multilingual named entity recognition. Proc. of The Second Global Wordnet Conference, pp.169 - 174. 2004.

[9 B. Magnini, N. Matteo, R. Prevete, and H. Tanev, A wordnet-based approach to named entities recognition. Proc. of the 2002 workshop on Building and using semantic networks, pp.1-7, 2002.

[10] S. Sekine, R. Grishman, H. Shinnou, A decision tree method for finding and classifying names in Japanese texts. Proc. the Sixth Workshop on Very Large Corpora. 1998.

[11] C. K. Lee, P-M. Ryu, H. K Kim, Named Entity Recognition using a modified Pegasos algorithm. Proc. of the CIKM, pp.655-667. 2010.

[12] Korean Wikipedia, http://ko.wikipedia.org/

[13] Toral, A. R. Munoz, A proposal to automatically build and maintain gazetters for named entity recognition by using Wikipedia, NEW TEXT Wikis and blogs and other dynamic text sources, 2006.

[14] R. Bunescu, M. Pasca, Using encyclopedia knowledge for named entity disambiguation. Proc. of EACL pp.9-16, 2006.

[15] T. Nguyen H. cao, Exploiting Wikipedia and text features for named entity disambiguation. Proc. of the 2nd international conference on intelligent information and database system, pp.101-104, 2010.

[16] C. Lee, Y. Hwang, M. Jang, Fine-Grained Named Entity Recognition and Relation Extraction for Question Answering. Proc. of SIGIR, pp.799-800, 2007.

[17] L. Deng, D. Yu, Deep Learning: Methods and

Applications. Foundations and Trends® in Signal Processing. Vol. 7, No. 3 - 4, pp 197-387, 2014.

유 철 중(Yoo, Chul Jung)

- 1985년 8월 : 전남대학교 대학원 계산통계학과(이학석사)
- 1994년 8월 : 전북대학교 대학원 전산통계학과(이학박사)
- 2012년 1월 ~ 2013년 7월 : University of California, Irvine(UCI) 국외연구교수
- 1997년 1월 ~ 현재 : 전북대학교 소프트웨어공학과 교수
- 관심분야 : 소프트워어품질/메트릭스/테스팅, 임베디드 소프트웨어/테스팅, GIS, 교육공학, 인지과학 etc.
- E-Mail : cjyoo@jbnu.ac.kr

김 용(Kim, Yong)

- 1995년 8월 : University of NorthTexas (MS in Information Science)
- 2000년 2월 : 충남대학교 컴퓨터과학과 (이학석사)
- 2006년 6월 : 연세대학교 문헌정보학과 (문헌정보학 박사)
- 1996년 2월 ~ 2008년 8월 : ㈜ KT 중앙연구소 책임연구원
- 2008년 9월 ~ 현재 : 전북대학교 문헌정보학과 부교수
- 관심분야 : 정보검색, 디지털도서관, 데이터마이닝, e-Learning, 전자기록물, 전자기록관리시스템
- E-Mail : yk9118@jbnu.ac.kr

윤 보 현(Yun, Bo Hyun)

- 1999년 8월 : 고려대학교 컴퓨터학과 이학박사
- 1999년 9월 ~ 2002년 9월 : 한국전자통신연구원 선임연구원(팀장)
- 2003년 3월 ~ 현재 : 목원대학교 컴퓨터교육과 교수
- 관심분야 : 자연어처리, 정보검색, 시맨틱웹, e-Learning
- E-Mail : ybh@mokwon.ac.kr