능동 학습 기법을 활용한 개체명 사전 반자동 구축 도구 개발

윤보현† · 오효정††

요 약

웹 3.0 시대의 도래와 IoT(Internet of Things) 기술을 발달에 따라 생산된 정보의 양 역시 기하급수적으로 늘고 있다. 본 논문에서는 이 중에서 사용자의 관심도가 높은 개체명(NE: Named Entity) 사전을 반자동으로 구축하는 도구를 개발하였다. 제안된 방법은 초기 학습 모델을 통해 인식된 결과로부터 오류 후보를 자동으로 생성하고 사용자로부터 최소한의 보정 작업을 수행하여 이를 재학습한다, 특히 공개지식자원인 위키피디아 내의 다양한 메타데이터의 특성을 활용하여 능동 학습에 필요한 학습 예제 작성을 위한 수작업을 최소화하고자 한다. 도구 활용 효과를 분석한 결과, 능동 학습을 통해 자동 인식 결과의 오류의 약 68.6%가 보정됨을 보였다.

주제어 : 능동 학습, 개체명 사전, 위키피디아

Development of Semi-automatic Construction Tool for Named Entity Dictionary based on Active Learning

Bo-Hyun Yun† · Hyo-Jung Oh††

ABSTRACT

Along with advent of Web 3.0 era and advanced technologies of IoT(Internet of Things), massive amounts of information are generated. Reflecting this trend, this paper developed a semi-automatic construction tool for named entity dictionary based on active learning. Our proposed method chose error candidates to verify among the preliminary results using initial trained model and re-trained the model for correctly labeled data by user. We adopt active learning approach for minimizing human effort utilized metadata features of Wikipedia. Based on experimental results using our tool, we show that 68.6% errors were automatically corrected.

Keywords : Active Learning, Named Entity Dictionary, Wikipedia

† 정 회 원: 목원대학교 컴퓨터교육과 교수 (제1저자)
†† 정 회 원: 전북대학교 대학원 기록관리학과 조교수, 문화융복합 아카이빙 연구소 연구원 (교신저자)
* 논문접수: 2015년 10월 12일, 심사완료: 2015년 11월 10일, 게재확정: 2015년 11월 23일
* 이 논문은 2015년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음
* 본 연구의 일부는 2015년 미래창조과학부 및 정보통신기술진흥센터의 SW컴퓨팅산업원천기술개발사업의 일환으로 수행하였음. [과제번호: 10044577, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]

1. 서 론

오늘날 우리는 웹 2.0 시대에서 웹 3.0 시대로 변화하는 시기에 살고 있다. 사용자가 정보를 공유할 수 있는 수단이 기존의 전통적인 퍼스널 컴퓨터(PC)에서 나아가 스마트 폰, 태블릿 PC 등과 같은 모바일 기기 뿐 아니라 스마트TV와 같은 가전제품, 자동차 네비게이션, 스마트 워치(watch) 등 다양한 전자기기가 IoT(Internet of Things) 환경 하에 연결이 가능해짐에 따라 생산된 정보의 양 역시 기하급수적으로 늘고 있다[1].

이러한 환경에서 사용자의 관심도가 높은 특정 정보를 자동으로 수집하여 효과적으로 관리하는 기술에 대한 수요 역시 폭증하고 있다. 이때 사용자들의 관심이 높은 정보라 함은 주로 특정 개체에 대한 것으로, 본 논문에서는 특히 개체명에 초점을 두기로 한다. 개체명(Named Entity)이란 인명, 지명, 기관명, 날자, 시간 등 문장에서 핵심적인 의미를 지닌 고유명사나 미등록어 등을 말하는 것으로[1], 개체명 사전은 해당 개체명과 분류 태그(tag)로 구성되어 있다(예: 인명-홍길동).

웹 3.0 시대의 또다른 화두는 '사용자 참여'에 의한 '집단지성'이다[2]. 집단지성의 가장 대표적인 예로는 우리가 지금 사용하고 있는 위키피디아(Wikipedia)가 있다. 위키피디아는 2001년 1월, 지미 웨일스(Jimmy Wales)와 래리 싱거(Larry Sanger)의해 시작된 백과사전 구축 프로젝트로[3], 여러 사람이 자유롭게 열람하고 확실하지 않거나 잘못된 정보는 누구나 수정 및 삭제 할 수 있는 형태의 자료열람 사이트이다. 다양한 사람들의 종합적 지식이 한데 모아져 있어 거의 정확한 정보로 수렴되어 갱신된다.

본 논문에서는 위키피디아 문서가 갖는 특성을 활용하여 개체명 사전을 구축하고 확장하기 위한 반자동 도구를 개발하고자 한다. 특히 초기 학습 결과를 활용해 인식된 오류를 재학습, 학습 결과를 보정하는 능동 학습(active learning) 기법을 적용함으로써 수작업을 최소화 하고 효율적으로 학습 데이터를 확장하는 방법에 대해 기술하고자 한다.

논문의 구성은 다음과 같다. 2장에서는 기존의 능동 학습 방법을 활용해 새로운 지식을 구축하고 확장하는 연구 사례와 위키피디아를 대상으로 한 언어 자원 구축 연구 사례들을 살펴보고 본 논문과의 차이점을 밝힌다. 3장과 4장에서는 본 논문에서 제안하고자 하는 능동 학습에 기반한 개체명 사전 구축 방법과 이를 위한 핵심 요소인 위키피디아 메타데이터 특성에 대해 각각 살펴보고 5장에서는 개발된 도구와 실제 적용 결과를 통해 효과를 검증하고 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

2.1 능동 학습 기반 지식 처리

일반적으로 기계 학습(machine learning)을 위해서는 수작업으로 정답이 부여된(labeled) 다수의 학습 데이터가 필요하다. 능동 학습 방법은 학습 예제로 사용할 수 있는 예제의 수가 제한되어 있는 상황에서 학습에 가장 도움이 되는 데이터를 선택하여 전문가의 태깅 혹은 검증에 의해 훈련 집합에 포함시키거나, 초기 훈련 집합이 주어지지 않았을 경우에 전체 데이터 분포를 잘 나타내는 데이터 부분 집합을 선택하여 훈련 집합을 만드는 방법으로[4], 학습에 필요한 예제 생성 비용을 효과적으로 줄이기 위해 다양한 분야에서 적용되고 있다.

능동 학습 방법은 특히 어휘 정보를 활용한 지식 처리 분야에 매우 활발히 적용되고 있는데, 1994년 Lewis 와 Gale[5]이 문서 분류(text classification) 문제에 적용한 것을 시작으로 1995년 이후에는 형태소 분석과 구문 분석, 어휘 중의성 해소 등의 문제에 적용한 연구가 속속 발표되었다[6]. 2000년대에 들어서는 개체명 인식[7]을 비롯해 음성인식, 자동 통역 등 언어자원을 구축, 보강하는 분야로 적용 범위가 확장되고 있으며 최근에는 기존의 능동 학습 방법이 초기 훈련 집합이 주어진 경우를 가정하는 제약에서 탈피하여 학습 데이터가 전혀 없는 경우를 대비하거나, 학습 속도를 줄이는 방향으로의 연구가 활발히 진행되고 있다[8]. 본 논문에서는 초기 학습 결과의 오류를 보정하기 위해 작업자의 수작업을 최소화 하는 방안에 주안점을 둔다.

2.2 위키피디아 기반 개체명 사전 구축

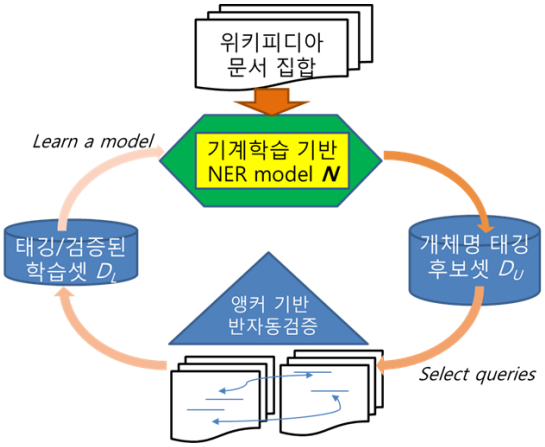
디지털 콘텐츠로부터 개체명 사전을 자동 혹은 반자동으로 구축하기 위한 연구들은 1990년대부터 시작되었다가, 2000년대 들어서 기계학습 기법이 발달함에 따라 가속화되었다. 특히 본 논문에서와 같이 위키피디아를 활용해 개체명 사전을 확장하고자 하는 연구도 다수 시도된 적이 있는데, 주로 위키피디아 표제어의 분류를 위해 WordNet을 활용하여 개체명을 인식한 경우[9]로, 이를 개체명 사전의 엔트리로 활용하기 위해서는 수작업으로 확인하는 일일이 확인, 검수하는 과정이 필요하다는 한계를 보였다.

국내에서도 한국어를 대상으로 개체명 사전 구축을 위한 연구 역시 활발히 진행되고 있다. [10]은 위키피디아 엔트리와 분류정보를 이용하여 개체명 범주에 대한 가상문서를 구축하고, 분류 대상 엔트리의 분류정보를 질의로 하여 정보검색 기법을 적용하는 방법을 제안하였으며 [11]에서는 위키피디아 분류정보를 활용하여 개체명의 중의성을 해소하는 기법을 제안하였다. 그러나 이들 방법들 역시 대부분이 위키피디아의 메타데이터 중 '분류' 정보만을 활용하는 것으로 본 논문에서와 같이 한국어 위키피디아가 갖는 구문적 특성과 다양한 구조적 정보를 활용해 개체명 사전을 자동으로 구축, 수작업을 최소화하여 확장하기 위한 방법과는 그 차이가 있다.

무엇보다도 가장 근본적으로 위와 같은 방법을 적용하여 높은 성능을 내기위해서는 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.

3. 능동 학습 기반 개체명 사전 확장

<그림 1>은 본 논문에서 제안하고자 하는 능동 학습 과정을 도식화한 것으로, 크게 초기 후보셋으로부터 보정 혹은 확장 대상을 선택하는 단계('select queries')와 보강된 학습셋으로부터 재학습을 통해 확장된 기계학습 모델을 견고히 하는 단계('learn a model')로 구성된다.



<그림 1> 능동 학습 기반 개체명 인식

Input: base learner B , labeled training data set D
Output: NER N.

1. Initialize the process by applying base learner B to labeled training data set D_L to obtain NER N.
2. Apply N. to unlabeled data set D_U to obtain D_U'
3. From D_U', select the most informative n instances to learn from, I.
4. Ask the teacher for NER of the instances in I.
5. Move I, with supplied NER, from D_U' to D_L.
6. Re-train using B on D_L to obtain a new NER, N''.
7. Repeat steps 2 through 6, until D_U is empty or until some stopping criterion is met.
8. Output a NER that is trained on D_L

<그림 2> 개체명 인식을 위한 능동 학습 의사 코드

<그림 2>은 세부 알고리즘을 의사코드(psudo-code) 형태로 기술한 것으로, 기본 학습모델(base model B)을 적용해 학습된 개체명인식기 N을 태깅되지 않은 문서셋(DU)에 적용, 식된 결과로부터 보정 대상 후보셋(DU')을 선정하고, 앵커 정보를 활용해 반자동 검증을 수행한 후(DL), 이를 다시 학습시켜 개선된 학습모델 N''을 생성한다.

본 논문에서 확장 대상으로 사용한 개체명 사전은 한국전자통신연구원(이하 ETRI)에서 개발한 개체명 인식기[12]에서 활용하기 위한 것으로, 상위 15개 대분류와 184개의 세부 분류로 구성되어 있으며 학습에 활용한 기계학습 방법은 지지벡터

기계(SVM: Support Vector Machine) 알고리즘을 활용하였다[13].

다음 장에서는 능동 학습에 활용된 위키피디아 문서의 메터데이터 특성에 대해 살펴보기로 한다.

4. 위키피디아 메타데이터 특성 분석

현재 위키피디아는 세계적으로 291개국 언어로 구축되어 활용되고 있다. 가장 구축이 되는 언어는 역시 영어로, 2015년 9월말을 기준으로 총 4,976,099 개 표제어(headword)로 구성되어 있다. 이에 비해 한국어 위키피디아는 329,159개로 영어 위키피디아에 비해 15분의 1 수준으로 작은 편이다. 그러나 최근에는 매월 2~3천 문서씩 꾸준히 증가하고 있는 추세를 보이고 있으며, 이는 한국어로 작성된 공개지식자원으로는 최대 규모이다. 이와 같이 집단지성을 통해 지속적으로 축척된 어떤 지식이 또 다른 지식자원에 반영되어 확장되는 순환 학습 과정에 본 논문의 주안점이 있다.

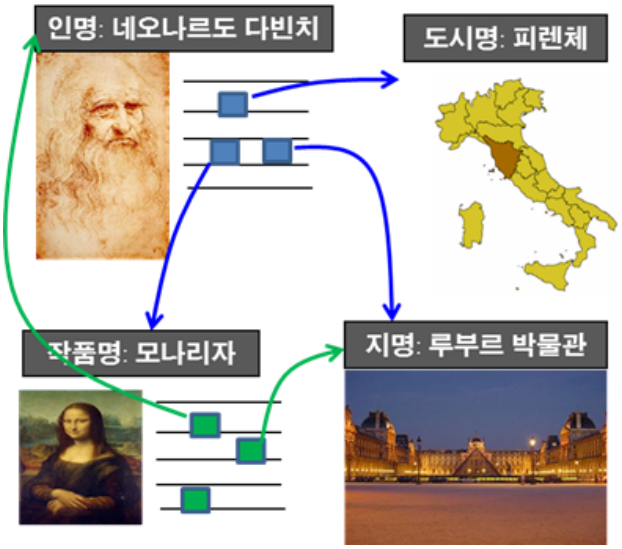
<표 1> 위키피디아 문서 속성 정보

제목	찰스 다윈
가리키는 문서(4)	다빈치 레오나르도 다빈치 다 빈치 레오나르도 디 세르 피에로
숨은 분류 (13개)	BIBSYS 식별자 포함 문서 ISNI 식별자 포함 문서 글로벌세계대백과 인용 문서 이탈리아어 표기를 포함한 문서 …
포함한 틀 (31개)	틀:구텐베르크 저자 틀: 예술가 정보 틀: 위키공용분류 틀: 인물데이터인용 ….
…	…

모든 위키피디아 문서는 기본 정보(Page Information)와 <표 1>과 같은 '속성(Properties)' 정보가 함께 제공된다[14]. 자세히 살펴보면, 특정 표제어를 설명하는 [본문(contents)]과 각 표제어별 특성에 따라 정형화된 정보를 [표(table)] 형식(위키피디아에서는 '틀(template)'이라는 용어를 씀)으로 제공한다. '레오나르도 다 빈치'의 경우, 관련된 이형태 정보가 4개, 관련 '분류' 정보가 13개, 본문 내용에 포함된 '틀'정보가 31개로 구성됨을 알 수 있다. <표 1>의 부가정보 중 개체명 인식을 위해 중요한 자질로 활용된 정보는 '분류'와 '틀' 정보로, '레오나르도 다 빈치'의 경우 '예술가 정보', '인물데이터 인용'과 같은 '틀' 정보와 연결되어 있다. 그 밖에도 아래와 같은 분류 정보를 통해 '레오나르도 다 빈치'가 인명이며 그 중에서도 화가 및 건축가, 조각가 등 예술가 직업군에 해당함을 유추할 수 있다.

> **분류정보:** 1452년 태어남 / 1519년 죽음 / 르네상스 화가 / 이탈리아 화가 / 이탈리아의 건축가 / … / 지폐의 인물…

한편, 위피키디아는 해당 표제어를 설명하기 위해 필요한 관련 문서들을 연결하기 위해 '앵커(anchor)'라는 메타데이터를 제공한다. 앞서 설명한 '속성' 정보가 해당 문서 내의 기술된 메타데이터(inner-information)인 반면, '앵커' 정보는 문서 밖의 메타데이터 정보(outer-information)를 활용한다는 점에서 구별된다.



<그림 3> 앵커 기반 위키피디아 문서 항해

<그림 3>은 <표 1>의 '레오나르도 다 빈치' 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.

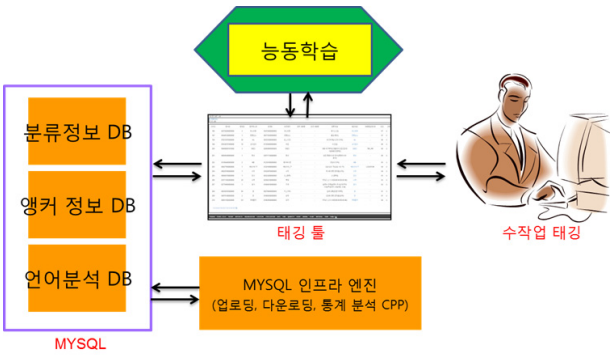
<그림 3>의 경우, '레오나르도 다 빈치'를 설명하는 문장에 나타난 앵커 텍스트, '피렌체', '모나리자', '루부르 박물관' 등이 개체명 후보 대상이다.

세부적으로 살펴보면, '피렌체' 앵커의 경우 링크를 통해 찾아간 '피렌체' 표제어에 관한 설명문과 '틀', '분류' 정보 등을 통해 해당 표제어가 '도시명' 임을 알 수 있다. '모나리자' 표제어의 경우, '레오나르도 다 빈치' 앵커를 통해 다시 '루부르 박물관' 표제어까지 링크로 연결되는데, 이와 같이 앵커를 많이 갖는 문서부터 시작해 문서를 항해하는 것이 개체명 사전으로 구축할 후보와 단서를 많이 접할 수 있다.

5. 개체명 사전 반자동 구축 도구 활용

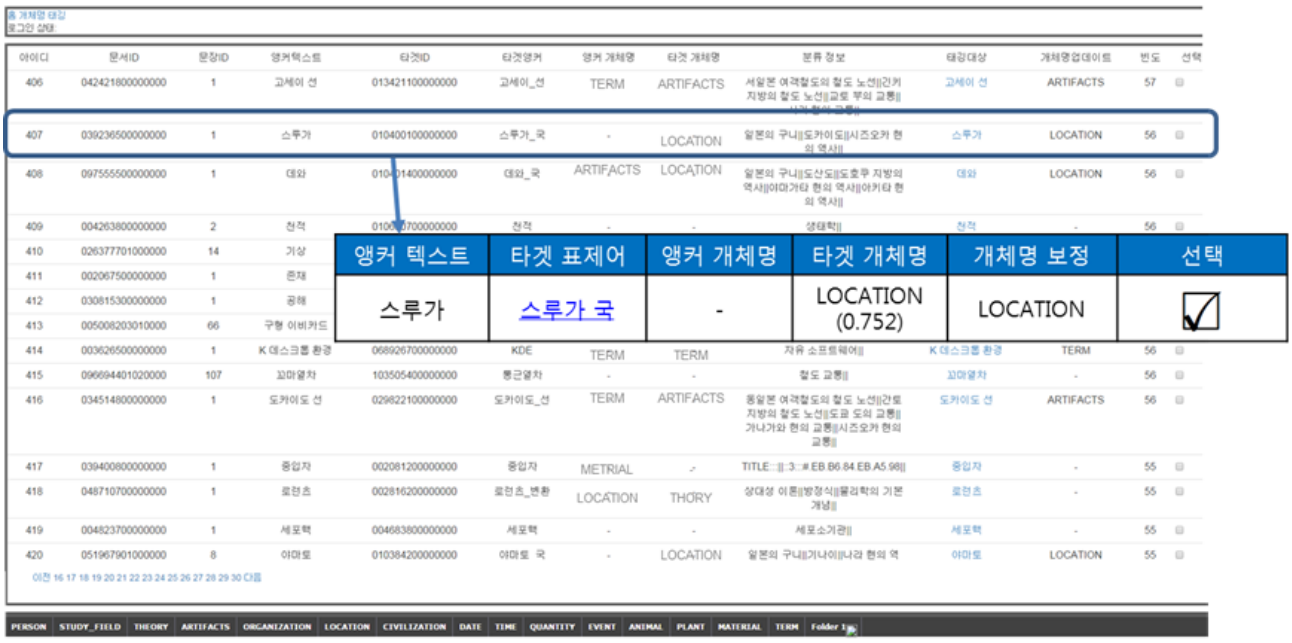
5.1 개체명 사전 태깅 도구

위키피디아 메타데이터 특히 '앵커'와 '틀' 정보를 활용한 태깅 도구는 다수의 앵커 텍스트를 빠르게 검색하고 정렬하는 기능이 우선적으로 고려되어야 한다. 또한 단순히 앵커 텍스트 자체만을 참조하여 정확한 개체명 사전 범주를 결정하기에는 부족한 경우가 있기 때문에 부가 설명이 필요한 앵커 텍스트에 대해 원문 정보와 분류 정보를 함께 보고 판단할 수 있도록 하는 기능도 고려하였다.



<그림 4> 능동 학습 기반 개체명 사전 반자동 구축 도구

<그림 4>은 본 논문에서 개발한 개체명 사전 반자동 구축 도구의 개발환경을 도식화 한 것이다. 개발환경으로는 다수의 사용자가 동시에 그리고 어디서든 태깅 가능한 환경을 제공하기 위해 웹 기반 아파치(apache) 서버와 HTML을 사용하여 사용자 인터페이스를 개발하였으며, 태깅된 결과물은 MySQL과 PHP를 사용하여 데이터베이스와 연동 및 저장이 가능하도록 개발되었다. 또한 태깅 대상 앵커 텍스트와 분류 정보와 같은 기반 정보들은 C++언어를 사용해 데이터베이스와 연동되어 데이터를 업로딩하고, 태깅된 결과물 또한 텍스트(txt) 파일 및 엑셀(틴) 파일로 다운로드가 가능한 환경으로 구성하였다.



<그림 5> 개체명 태깅 화면

<그림 5>는 실제 개체명 사전을 구축하는 태깅 작업 화면이다. 각 열은 하나의 앵커 정보와 대응되는 것으로 앵커 텍스트 관련정보를 확인할 수 있는데, 세부적으로는 해당 앵커가 나타난 문서의 아이디, 문장 번호, 대상 표제어 문서의 아이디, 표제어, 분류 정보 및 빈도 등을 보여준다.

특히 '앵커 개체명'과 '타겟 개체명'은 3장에서 설명한 초기 학습 결과를 통해 자동으로 인식된 개체명 태그 정보로, 양쪽 정보가 다른 경우를 주로 대상으로 수작업 보정을 수행한다. 보정 대상은 체크박스를 선택 후 인터페이스 하단의 개체명 대 분류 중 하나를 선택하면 개체명 태깅 결과가 적용된다.

본 논문에서 개발된 도구를 통해서 보정하고자 하는 대상은 주로 기계학습을 통해 구축된 자동 개체명 인식결과가 서로 다른 경우로, 이 경우 <그림 5>의 단순 정보만으로는 정확한 개체명 범주를 분류하기 어려운 경우도 다수 발생한다. 이를 위해 본 논문에서는 <그림 6>와 같이 앵커가 가리키는 타겟 표제어의 '분류' 정보 및 '틀' 정보를 브라우징할 수 있는 창을 제공함으로써 개체명 태깅 보정의 정확도를 꾀한다.

<그림 6> 앵커 클릭 결과 화면

5.2 도구 활용

이번 장에서는 본 논문에서 개발한 개체명 사전 반자동 구축 도구의 활용 효과를 분석한다.

<표 2>은 개발된 도구를 통해 수집된 초기 학습 오류 양상을 세부적으로 분석한 표이다. 우리가 앵커에 대한 개체명 자동인식 수행 결과로 기대하는 이상적인 모습은 앵커 텍스트 및 대상 표제어 모두 정답 개체명 태그로 같게 분석되는 경우이다. 그러나 실제 분석 결과, <표 2>와 같이 약 24% 정도가 서로 다른 결과를 나타내는 것으로 분석되었다.

<표 2> 초기 개체명 인식 모델 결과 차이 분석

개체명 분류	같은 수	다른 수	비율
수량표현(QT: quantity)	1,935	16,651	89.60%
시간(TI: time)	340	728	68.20%
학술분야(FR: field)	14,591	15,893	52.10%
인공물(AF: artifact)	117,608	98,747	45.60%
이론(TR: thory)	20,823	11,480	35.50%
용어(TM: term)	43,517	22,126	33.70%
물질(MT: material)	11,476	5,103	30.80%
문명/문화(CV: civil)	108,706	46,966	30.20%
사건(EV: event)	37,300	15,870	29.80%
식물(PT: plant)	9,696	3,981	29.10%
동물(AM: animal)	17,669	7,207	29.00%
날짜(DT: date)	358,903	140,290	28.10%
조직명(OG: orginazation)	317,116	106,492	25.10%
지명(LC: location)	659,942	150,950	18.60%
인명(PS: person)	537,109	85,879	13.80%
기타	1	100	99.00%
합	2,256,732	728,463	24.4%

세부적으로 살펴보면, 앵커 텍스트와 대상 표제어의 개체명 인식 결과가 서로 다른 결과를 보인 상위 분류 태그들은 수량(QT)의 경우 89.6%, 시간(TI)의 경우 68,2%, 학술분야(FD)는 52.1%로 분석되었다. 이들 태그들은 사용자의 관심도가 떨어지는 분류인 반면 관심도가 높고 전체 개체명의 다수를 차지하는 분류들인 인명(PS, 13,8%), 지명(LC, 18.6%), 기관(OG, 25.1%)의 경우 앵커와 표제어 자동인식 결과가 다른 비율이 비교적 낮은 것으로 분석되었다. 이는 자동인식 결과가 정답일 확률이 크다는 것을 암시하는 것으로, 나아가 전체 개체명 사전 정확도에 긍정적인 영향을 미치는 것으로 해석될 수 있다.

<표 2>에서 가장 높은 비율로 앵커와 표제어

개체명 인식 결과가 다른 태그는 수량표현(QT)으로, 가장 혼돈을 많이 일으킨 오류는 수량(QT)에 해당하는 개체명을 지역명(LC)으로 인식한 경우로, 거의 절반(45.37%)이 이에 해당한다. 오류의 예는 다음과 같다.

- 제5공화국: 순서표현(QT_order) --> 정답: 국가명(LCP_country)
- 제3영업일: 순서표현(QT_order) --> 정답: 날짜(DT_day)

그 밖에 오류 양상으로는 가격(QT_Price), 나이(QT_Age), 전화번호(QT_phone) 등 같은 수량(QT) 범주 내에서 세부 태그가 다른 경우(8.37%) 혹은 날짜(DT)와 같이 숫자표현에서 오인식된 경우(7.87%) 등으로 분석되었다. 그러나 수량(QT), 시간(TI) 등의 태그들은 개체명 사전으로써의 중요도가 떨어지는 범주로 보정 우선순위에서 후위로 두었다.

한편 관심도가 비교적 높은 인공물(AF)이나 용어(TM), 이벤트(EV) 와 같은 태그 역시 각각 45.6%, 33.7%, 29.8%로 상이한 경우가 다수 발생하였다. 이는 개체명 자동인식 결과의 성능 저하를 의미하므로 해당 태그에 대한 양상 분석 수행을 통해 오류를 보정하고자 한다. 인공물(AF)에 해당하는 앵커의 개체명 결과와 대상 표제어 개체명 결과의 차이를 세부 분석한 결과 다음과 같은 오류가 발생하였다.

- 바흐 BWF No. 37: 도로명(AF_road) --> 정답: 음악작품명(AFWA_music)
- 중앙일보: 작품명(AF_works) --> 정답: 언론사명(OGG_media)

인공물에 해당하는 앵커는 주로 같은 인공물 대분류 내의 세부 범주가 다른 경우가 40,994로 전체 43%를 차지하였으며 그 다음으로는 기관명(OG)으로 분류한 경우가 16.801건, 인명(PS)으로 분류한 경우 7,845건, 용어(TM) 6,474 건 순으로 분석되었다. 이들 태그들은 정보 추출 및 검색 응용에 미치는 영향이 큰 분류들로 제안된 도구를 통해 보정이 필요한 대상 후보들이다.

5.3 활용 효과 분석

제안된 방법의 효과를 입증하기 위해서 먼저 수작업 정답셋을 구축이 필요하다. 본 논문에서는 위키피디아 문서 중 앵커가 20개 이상 포함된 360 문서를 임의로 선정하고 해당 문서셋에 포함된 전체 앵커 5,713개를 대상으로 평가셋을 구축하였다.

<표 3>는 제안된 방법을 적용한 개체명 사전 리스트를 평가한 결과이다. 전체 5,713개 대상 중에서 기존 개체명 인식기로 성공한 앵커가 3,979개이고 개체명으로 인식하지 못한 경우 혹은 인식했으나 세부 분류를 잘못 할당한 오류 케이스가 1,734개였다. 이들 오류 중 앵커 정보를 활용해 1,321개를 추가 인식했다. 이 중에서 다시 대상 표제어 정보를 통해 세부 태그를 보정한 경우가 1,190개로, 최종적으로 제안된 방법을 통해 전체 68.6%를 자동화 할 수 있음을 보였다.

<표 3> 도구 활용 실험 결과

항 목	갯 수
실험대상 문서	360
전체 앵커 수	5,713
초기 개체명 모델 적용 정답 앵커 수	3,979
오류 보정 후보 대상 앵커 수	1,734
도구 활용으로 보정된 앵커 수	1,321
재학습 후 정답으로 인식된 앵커 수	1,190 (68,6%)

6. 결 론

본 논문에서는 집단지성을 통해 새롭게 생성되고 갱신되는 공개지식자원인 위키피디아의 특성을 활용하여 개체명 인식 결과의 오류를 보정하고 사전을 확장하기 위한 반자동 구축 도구를 개발하였다. 특히 초기 학습 모델을 통해 인식된 결과로부터 오류 후보를 자동으로 생성, 사용자로부터 최소화의 작업을 통해 보정하여 이를 재학습시키는 능동 학습 기법을 적용하였다. 실험 결과 제안된 도구를 활용해 재학습한 결과 오류의 70%가 보정됨을 보였다.

이러한 결과는 도구 사용을 통해 작업자가 전체 보정 대상 중 매우 일부만을 보정한 후 이를 기계가 재학습함으로써 반자동 태깅이 가능함을 의미하는 것으로, 본 논문의 주된 목적인 '작업자

수작업의 최소화'를 달성한 것으로 해석될 수 있다. 그러나 여전히 30%의 오류는 해결되지 못하였는데, 이는 개체명 내의 중의적으로 사용되는 용어가 많아서 인 것으로 분석되었다 (예: 베를린-도시명, 영화제목).

차기 연구 방향으로는 수작업으로 보정할 후보를 보다 효율적으로 선택하는 알고리즘을 개발하고 능동 학습에 사용된 기계 학습 기법을 다양화하고자 한다.

참 고 문 헌

[1] Goldman Sachs (2014), The Internet of Things: Making sense of the next mega-trend, IoT Primer, http://www.goldmansachs.com/our-thinking/outlook/internet-of-things/iot-report.pdf

[2] 정유선 (역) (2008), Web 3.0. (Team Weboook)서울: 라이온북스

[3] Wikipedia, history, https://en.wikipedia.org/wiki/Wikipedia:About

[4] Settles, B. (2009). Active learning literature survey: Computer sciences technical report 1648, University of Wisconsin-Madison,

[5] Lewis, D. & Gale, W. (1994). A sequential Algorithm for Training Text Classifiers. The Proceedings of ACM-SIGIR Conference, 3 - 12.

[6] Olsson, Fredrik (2009). A literature survey of active machine learning in the context of natural language processing, SICS Technical Report T2009:06

[7] Vlachos, Andreas (2006). Active annotation. The Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006), 64 - 71.

[8] 우호영, 박정희 (2013). 계층적 군집화를 이용한 능동적 학습. **정보처리학회논문지/소프트웨어 및 데이터 공학**, 2(10), 705-712

[9] Toral A. & Munoz, R. (2006). A proposal to automatically buid and maintain gazetters for named entity recognition by using Wikipedia", The Proceedings of EACL, 56-61

[10] 송영길, 정석원, 김학수 (2015). 위키피디아를 이용한 정보검색 기반 개체명 사전 구축 방법. 2015년 한국컴퓨터종합학술대회 논문집, 648-659

[11] 김태현, 이창수, 황재원, 고영중 (2015). 위키피디아를 이용한 개체명 부착 코퍼스 자동구축 및 중의성 해소, 2015년 한국컴퓨터종합학술대회 논문집, 745-747

[12] 류법모, 김현진, 김현기, 박상규 (2012). 심층 언어분석 기반 소셜미디어 이슈 탐지 분석 기술, **정보과학회지, 30**(6), 57-68

[13] Lee, C., Hwang, Y. & Jang, M. (2007). Fine-Grained Named Entity Recognition and Relation Extraction for Question Answering, The Proceedings of the ACM-SIGIR conference, 799-800

[14] 유철중, 김용, 윤보현 (2015). 언어자원 자동구축을 위한 위키피디아 콘텐츠 활용 방안 연구, **디지털융복합연구, 13**(5), 187-194

윤 보 현

1999 고려대학교 컴퓨터과학과 이학박사
1999～2002 한국전자통신연구원 선임연구원(팀장)
2003～현재 목원대학교 컴퓨터교육과 교수
관심분야: 컴퓨터교육, 자연어처리, 정보검색, 시맨틱웹
E-Mail: ybh@mokwon.ac.kr

오 효 정

2000 충남대학교 컴퓨터과학과 (이학석사)
2008 한국과학기술원 컴퓨터공학과 (공학박사)
2000～2015.05 한국전자통신연구원 지식마이닝 연구실 책임연구원
2015.05～현재 전북대학교 기록관리학과 조교수
관심분야: 정보검색, 질의응답, 빅데이터정보처리
E-Mail: ohj@jbnu.ac.kr