과학기술 핵심개체 인식기술 통합에 관한 연구

A Study on the Integration of Recognition Technology for Scientific Core Entities

최윤수(Yun-Soo Choi)*
정창후(Chang-Hoo Jeong)**
조현양(Hyun-Yang Cho)***

초 록

대용량 문서에서 정보를 추출하는 작업은 정보검색 분야뿐 아니라 질의응답과 요약 분야에서 매우 유용하다. 정보추출은 비정형 데이터로부터 정형화된 정보를 자동으로 추출하는 작업으로서 개체명 인식, 전문용어 인식, 대용어 참조해소, 관계 추출 작업 등으로 구성된다. 이들 각각의 기술들은 지금까지 독립적으로 연구되어왔기 때문에, 구조적으로 상이한 입출력 방식을 가지며, 하부모듈인 언어처리 엔진들은 특성에 따라 개발 환경이 매우 다양하여 통합 활용이 어렵다. 과학기술문헌의 경우 개체명과 전문용어가 혼재되어 있는 형태로 구성된 문서가 많으므로, 기존의 연구결과를 이용하여 접근한다면 결과물 통합과정의 불편함과 처리속도에 많은 제약이 따른다. 본 연구에서는 과학기술문헌을 분석하여 개체명과 전문용어를 통합 추출할 수 있는 기반 프레임워크를 개발한다. 이를 위하여, 문장자동분리, 품사태깅, 기저구인식 등과 같은 기반 언어 분석 모듈은 물론 이를 활용한 개체명 인식기, 전문용어 인식기를 개발하고 이들을 하나의 플랫폼으로 통합한 과학기술 핵심개체 인식 체계를 제안한다.

ABSTRACT

Large-scaled information extraction plays an important role in advanced information retrieval as well as question answering and summarization. Information extraction can be defined as a process of converting unstructured documents into formalized, tabular information, which consists of named-entity recognition, terminology extraction, coreference resolution and relation extraction. Since all the elementary technologies have been studied independently so far, it is not trivial to integrate all the necessary processes of information extraction due to the diversity of their input/output formation approaches and operating environments. As a result, it is difficult to handle scientific documents to extract both named-entities and technical terms at once. In order to extract these entities automatically from scientific documents at once, we developed a framework for scientific core entity extraction which embraces all the pivotal language processors, named-entity recognizer and terminology extractor.

키워드: 정보추출, 개체명 인식, 전문용어 인식
information extraction, named entity recognition, terminology extraction

---

* 한국과학기술정보연구원 정보기술연구실 선임연구원(armian@kisti.re.kr) (제1저자)
** 한국과학기술정보연구원 정보기술연구실 선임연구원(chjeong@kisti.re.kr) (공동저자)
*** 경기대학교 문헌정보학과 교수(hycho@kyonggi.ac.kr) (교신저자)

■ 논문접수일자 : 2011년 2월 17일 ■ 최초심사일자 : 2011년 2월 23일 ■ 게재확정일자 : 2011년 3월 10일
■ 정보관리학회지, 28(1): 89-104, 2011. 〔DOI:10.3743/KOSIM.2011.28.1.089〕

1. 서 론

인터넷의 발달과 더불어 대용량 데이터를 실시간으로 처리하여 필요한 지식을 발견하기 위한 정보추출 기술들이 핵심적인 분야로 인식되고 있다. 정보추출을 위한 가장 기본적인 요소기술은 인명, 지명, 기관명 등의 개체명을 인식하는 작업과 문서내의 전문용어를 인식하는 작업이다.

개체명 인식은 문서내의 원소를 찾아서 기 정의된 범주로 분류하는 작업으로, 특정 도메인에 따라 개체명 인식기(Named Entity Recognition System)가 개발된다. 본 연구에서는 일반적인 범주인 인명, 지명, 기관명에 대한 부분만 개체명으로 분류하고, 전문용어 인식을 위한 분야는 환경/에너지 분야를 선택하였다.

현재까지 개체명 인식 분야는 신문기사 및 방송기사 등을 중심으로 연구되었고, 전문용어 인식 분야는 생의학 분야에서 유전자, 단백질과 관련된 용어들에 대하여 주로 연구가 행해졌다. 개체명 인식 기술과 전문용어 인식 기술이 서로 독립적인 영역에서 연구되었고, 전문용어 인식에서도 그 범위는 생의학분야 정도로 매우 제한적이다.

과학기술문헌 중 특히 환경/에너지 분야는 국가, 지역 같은 개체명들과 환경/에너지 분야의 전문용어들이 밀접한 연관이 있어, 개체명과 전문용어가 혼재되어 있는 형태로 구성되므로, 기존의 연구결과물을 이용하여 접근하기 위해서는 개체명 인식과 전문용어 인식의 2단계로 작업을 수행해야 하기 때문에, 그 결과물을 통합하는 과정의 불편함이 있고, 처리속도에서도 많은 제약이 따른다.

본 연구에서는 개체명과 전문용어를 통합하여 핵심개체로 정의하고, 한국과학기술정보연구원이 보유하고 있는 해외학술지 데이터베이스 중에서 환경/에너지 분야 문헌들을 대상으로 개체명과 전문용어를 동시에 인식하는 통합된 플랫폼을 구현하고, 이를 통하여 기존 연구에서 수행하던, 개체명간의 관계추출과 전문용어간의 관계추출뿐만 아니라, 개체명과 전문용어간의 관계추출을 위한 기반시스템으로 활용하고자 한다.

2. 관련 연구

본 연구에서는 과학기술문헌에서 중요하게 인식되는 인명, 지명, 기관명의 일반적인 개체명과 분야에 특화된 전문용어를 통합하여 핵심개체로 지정하고 인식한다.

현재까지 개체명과 전문용어 인식에 관한 연구는 독립적으로 수행되어 왔기 때문에 핵심개체의 구성요소인 개체명과 전문용어에 대한 관련연구를 각각 조사한다.

2.1 개체명 인식 기술

개체명은 고유명사(복합명사 포함)와 시간 등을 나타내는 수식 표현을 말한다. 아래는 개체명의 예이며 PER은 인명을 나타내고 LOC는 지명을, ORG는 기관명을 나타낸다.

[PER Wolff], currently a journalist in
[LOC Argentina], played with
[PER Del Bosque] in the final years of
the seventies in
[ORG Real Madrid].

개체명 인식은 정보 추출(Information Extraction)의 한 분야로서 문서내에서 개체명을 추출하고 기정의된 분야(인명, 지명, 기관명 등)로 분류하는 작업을 말한다. 개체명 인식에 관한 연구는 MUC-6에서 유래되어 최근에는 개체명을 약 200여 개로 나누어 질의응답 시스템의 문장 분석에 널리 사용되고 있다(MUC-6: Message Understanding Conference, Extended Named Entity Hierarchy). MUC-6에 참가한 많은 시스템들은 특정 언어에 제한된 규칙과 자신만의 입출력 방법을 사용하여 다른 언어나 다른 영역에 쉽게 적용할 수 없었다.

MUC-6 이후 개체명에 대한 연구가 꾸준히 진행되어 CoNLL 2002와 2003을 통해서 많은 발전이 있었다. 이 대회에 참가한 대부분의 시스템은 기계학습 방법을 이용하였으며 영어의 경우에 약 89%의 정확률을 보인다(CoNLL: Conference on Computational Natural Language Learning). 기계학습 방법에서는 주로 BIO 태그(B: 개체명의 시작, I: 개체명의 중간, O: 관계없음)를 이용하는데 〈표 1〉은 CoNLL에서 사용한 예이다.

Black와 Vasilakopolos(2002)는 90년대 초반 Eric Brill이 처음 소개한 변환기반 학습(Transformation-based learning)을 이용하였다. 변환기반 학습은 형태소 부착, 영어 전치사구 접속문제 해결, 기저구 인식, 철자 수정 등 자연어처리의 다양한 분야에 사용된 기계학습 방법으로 첫 번째 단계에서 개체명을 인식하는 작업을 수행하고, 두 번째 단계에서 인식된 개체명 후보들에 대한 적합한 분류를 제공하는 방법이다.

Carreras, Marques와 Padro(2002)는 CoNLL 2002에서 개체명 인식을 위해 잘 알려진 BIO 태그 외에, 개체명의 시작과 종료 경계를 인식하기 위한 Open-Close&I와 Global Open-Close 방식을 도입하였고, 개체명 분류를 위해 AdaBoost 이진 분류 알고리즘을 사용하였다.

Collins(2002)는 개체명 경계인식을 위해 최대 엔트로피(Maximum Entropy, ME)태거를 이용하여 추천된 상위 N개의 후보에 대해 voted perceptron 알고리즘을 이용하여 가중치를 재부여하는 방법을 제안하였다. 학습 및 실험은 웹 데이터를 이용하였고 ME 태거의 85.3%의 F1-척도에 비해 87.9%의 F1-척도를 보여주었다.

Watanabe, Asahara와 Matsumoto(2007)는 HTML로 구성된 위키피디아 문서들을 그래프 구조로 표현하였다. HTML에서 하이퍼링크를 개체명으로 취급하고 이를 그래프상의 노드로 표현하였고, CRFs(Conditional Random Fields)를 도입하여 그래프 구조상의 노드들을 분류하였다.

〈표 1〉 CoNLL에서 사용된 BIO 태그

토큰	BIO태그	비고	토큰	BIO태그	비고
U.N.	I-ORG	기관명으로 인식	for	O	
official	O		Baghdad	I-LOC	지명으로 인식
Ekeus	I-PER	인명으로 인식	.	O	
heads	O				

이창기 등(2006)은 개체명 경계 인식을 위해 Conditional Random Fields를 이용하였고, 경계 인식된 개체명의 클래스 분류를 위해 Maximum Entropy를 이용하여, F1척도 83.4의 성능을 보여주었다.

2.2 전문용어 인식 기술

전문용어 자동 인식 연구는 크게 규칙 기반 연구, 통계 기반 연구, 그리고 앞의 두 연구방법을 병행하는 혼합형 연구로 구분한다. 규칙 기반 연구는 사전이나 규칙을 사용하는 방법으로 수작업을 통한 규칙의 정확성과 사전의 크기가 인식의 정확률을 결정한다. 통계 기반 연구는 지도 학습과 비지도 학습으로 나눠진다. 지도 학습은 사람의 판단을 통해 만들어진 대량의 말뭉치가 준비되어 있을 때 사용하기 좋은 방법이고, 비지도식 학습은 소량의 말뭉치를 대상으로 초기 규칙을 학습・인식의 과정을 반복해 성능을 향상시키는 방법이다. 일반적으로 비지도 학습보다 지도 학습이 좋은 인식 결과를 보인다. 학습에 사용되는 통계 모델은, 은닉 마코프 모델(hidden Markov model), 신경망(neural network), SVM (support vector machine), 최대 엔트로피 모델(maximum entropy model) 등이 있다.

Tanabe와 Wilbur(2002)는 규칙기반 접근 방법을 이용한 AbGene 시스템을 개발하였다. AbGene은 Brill 품사태거를 확장하여 유전자명과 단백질명에 관한 태그를 추가하였고, 생의학 분야 문헌으로부터 수집된 7,000개의 학습데이터를 이용해 학습하였고, 정확률 85.7%와 재현율 66.7%를 보여준다.

Chang, Schutze와 Altman(2004)은 문장내에서 출현하는 전문용어후보들에 후보의 빈도수, 형태소 분석결과, 문맥 등을 고려하여 가중치를 할당하는 GAPSCORE 시스템을 개발하였다. 더 높은 점수를 획득한 후보는 유전자명, 단백질명이 될 확률이 높다. GAPSCORE 시스템은 Yapex 말뭉치로 학습되었고 74%의 정확률, 81%의 재현율을 보인다.

Zhou, Zhang와 Su(2004)는 은닉마코프 모델(HMM: Hidden Markov Model)을 이용하였다. 대문자 시작 정보, 접두사와 접미사 정보, 품사정보, 시작 단어 등을 전문용어후보 추출을 위한 자질로 선택하였고, GENIA 말뭉치 2.1을 사용하여 학습 및 실험을 한 결과, 66.5%의 정확률과 66.6%의 재현율을 보였다.

오종훈과 최기선(2006)은 생의학분야 전문용어의 경계 인식을 위해 지지벡터기반의 기계학습모델을 사용하였고, 경계인식측면에서 78~86%의 정확률과 87%~90%의 재현율을 보였다.

현재까지 개발된 전문용어 인식 기술을 살펴보면, 주로 생의학분야에 집중되어 있고, 생의학분야의 전문용어들인 유전자, 단백질 등의 특징을 자질로 이용한 기계학습모델이 사용되었으나, 생의학 분야와 달리 이러한 특징을 지니고 있지 않은 과학기술문헌의 다른 분야에서는 기계학습모델에서 이용할 수 있는 자질이 부족하여 기계학습모델을 이용하기 매우 어렵다. 본 연구의 대상문헌 또한 환경/데이터분야로 선정하였기 때문에, 사전기반의 전문용어 인식시스템을 사용하도록 한다.

3. 과학기술 핵심개체 신규탐지 시스템

본 연구에서는 기존에 독립적으로 연구되던 개체명인식 분야와 전문용어 인식분야를 통합하여 과학기술문헌을 대상으로 핵심개체를 탐지하기 위한 플랫폼을 제안한다. 과학기술문헌 중 대상 분야로 환경/에너지 분야를 선정하고, 추출할 핵심개체를 정의하고, 핵심개체 추출을 위하여 개체명 인식을 위한 기계학습모델을, 전문용어 인식을 위해서 사전검색 검색 모듈을 제안한다. 그리고 이를 통합한 과학기술 핵심개체 신규탐지 시스템을 구현한다.

KISTI가 보유하고 있는 해외저널 중에서, 환경/에너지 분야인 193종 282,442건을 대상데이터로 선정하였다.

개체명은 자신을 나타내는 중요한 자질들을 많이 포함하고, 공개된 학습집합들이 풍부하여 기계학습모델을 적용한다. 반면, 전문용어의 경우에는 기계학습 모델을 적용하기 힘든 "환경/에너지 분야"를 처리하기 위하여 분야사전을 기반으로 하는 사전기반 방법을 사용한다. 이러한 이유로 기계학습모델을 적용한 개체명인식에 대해서만 성능을 평가한다.

3.1 핵심개체 탐지 시스템 전체 구성도

본 연구에서 대상 데이터를 환경/에너지 분야 문헌으로 선정하고, 개체명들과 전문용어들을 통합한 핵심개체들을 〈표 2〉와 같이 4개의 분류로 정의한다.

환경/에너지 분야의 과학기술문헌은 국가, 지역 등과 같은 지명, 환경과 관련된 단체 등과 밀접하게 연관된 주제를 취급하기 때문에, 일반개체명은 개체명 인식에서 주로 사용되는 인명(Person), 지명(Location), 기관명(Organization)으로 분류하고, 환경/에너지 분야의 주요 개체인 전문용어는 기술용어로 분류한다.

〈표 2〉 핵심개체 분류표

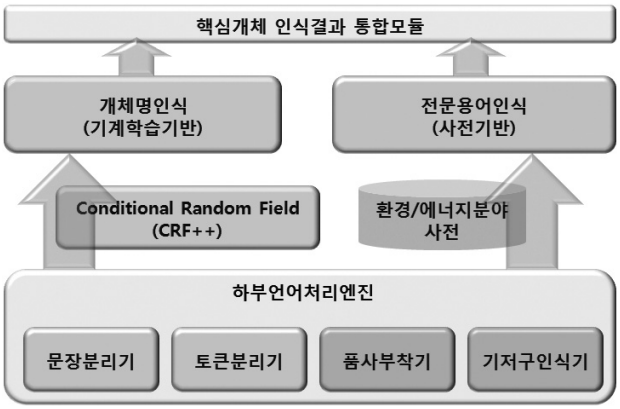
상위분류	상세분류(범주)	비고
일반개체명	Person	인명
	Location	지명
	Organization	기관명
기술명	TechTerm	기술용어

전문용어에 대한 분류의 경우 환경/에너지 분야에 특화된 분류들로 세분화할 수 있지만, 문헌내에서 인식되는 전문용어들의 분야별 통계를 분석한 후, 향후연구에서 세분화할 계획으로, 본 연구에서는 모든 환경/에너지 관련 전문용어들을 기술용어로 분류한다.

과학기술문헌에서 개체명과 전문용어를 인식하는 전체적인 시스템은 〈그림 1〉과 같다. 하부시스템인 언어처리 엔진(Language Processor)은 문장분리기, 토큰 분리기, 품사부착기, 기저구 인식기의 4가지 모듈로 구성되어 입력된 문서를 가공하고, 필요한 자질들을 추출하는 역할을 담당한다. 개체명 인식기, 전문용어 인식기는 하부의 언어처리과정에서 추출된 각각의 자질들 또는 통합된 자질들을 사용한다.

전체적인 작업은 언어처리 단계, 개체명 및 전문용어인식 단계로 진행된다.

언어처리 단계에서는 입력된 문서를 문장단위로 분리하고, 의미 있는 단위로 토큰을 인식하고, 토큰에 대한 적절한 품사를 부착하고, 기저명사구를 인식한다.



〈그림 1〉 과학기술 핵심개체 인식 시스템

하부 언어처리 엔진으로부터 추출된 정보를 이용하여 개체명 인식기에서 인명, 지명, 기관명을 추출하고, 전문용어 인식기에서 기술명을 추출한 뒤, 그 결과를 하나의 문서로 통합한다.

〈그림 2〉는 본 연구에서 제안한 통합된 핵심개체 인식 시스템을 환경/에너지 분야의 한 문서에 적용한 결과의 일부를 보여준다. 문서내에서 인식된 핵심개체들은 entity 엘레멘트를 이용하여 표현된다. 〈그림 2〉의 문서는 asbestos(석면)과 이와 관련된 질병 mesothelioma(중피종), 그리고 발생하는 지역(United Kingdom, South Africa)들을 포함하고 있다. 여기서 asbestos와 mesothelioma는 전문용어 인식 모듈에서, United Kingdom과 South Africa는 개체명 인식모듈에서 각각 인식되어 통합되고, 이후 개체명과 전문용어간의 관계추출에 활용된다.

```
〈?xml version="1.0" encoding="UTF-8"?〉
〈document id="d133"〉
〈sentence id="d133.s8" text="The causal relationship between asbestos exposure and lung cancer was confirmed in 1955 in asbestos textile workers in the United Kingdom, and later, in 1960, in South Africa, mesothelioma was attributed to asbestos exposure to even relatively low airborne concentrations of crocidolite."〉
    〈entity charOffset="32-39" id="d133.s8.e0" type="TechTerm" nn="asbestos"/〉
    〈entity charOffset="54-64" id="d133.s8.e1" type="TechTerm" nn="lung cancer"/〉
    〈entity charOffset="123-136" id="d133.s8.e3" type="Location" nn="United Kingdom"/〉
    〈entity charOffset="162-173" id="d133.s8.e4" type="Location" nn="South Africa"/〉
    〈entity charOffset="176-187" id="d133.s8.e5" type="TechTerm" nn="mesothelioma"/〉
    〈entity charOffset="275-285" id="d133.s8.e7" type="TechTerm" nn="crocidolite"/〉
〈/sentence〉
〈/document〉
```

〈그림 2〉 통합된 핵심개체로 인식된 문헌

3.2 언어처리엔진

언어처리 엔진은 문장분리기, 토큰분리기, 품사부착기, 기저구인식기 4개의 모듈로 구성된다. 언어처리 엔진을 구현하기 위해 규칙기반 방식과 기계학습기반 방식을 모두 사용할 수 있다. 규칙기반방식은 언어처리 엔진의 출력에 대한 오류를 쉽게 수정할 수 있는 장점이 있는 반면, 처리해야 할 규칙의 양이 많은 경우 적용하기 힘들다. 기계학습 방식은 학습데이터만 충분하다면 적절한 자질들을 적용하여 좋은 성능을 얻을 수 있지만, 결과에 대한 오류 수정이 어려운 단점이 있다.

본 연구에서는 하부언어처리 엔진 중 복잡하지 않은 규칙을 사용하여 구현할 수 있는 문장분리기와 토큰분리기는 규칙기반 방식으로 구현하고, 품사부착기와 기저구인식기는 기계학습모델을 사용한다. 지금까지 연구되어온 여러 기계학습모델 중에서 속도, 정확성 등을 고려하여 연속적인 데이터의 태깅이나 파싱을 위해 사용되는 기계학습모델의 한 종류인 CRF 모델을 선정한다(CRF: Conditional Random Fields).

3.2.1 문장분리기

문장분리기는 OAK 시스템을 기반으로 환경/에너지 분야에 적합하도록 수정한다. 정의된 규칙을 기반으로 현재 단어가 문장의 끝인지를 구분하고 다음 단어와의 분리여부를 결정짓는 방식으로, 분리여부를 결정하기 위해 현재단어의 전방단어들과 후방단어들의 정보를 이용하는 알고리즘을 사용하여 구현한다(Sekine 2010).

3.2.2 토큰분리기

토큰분리기는 OAK시스템을 기반으로 환경/에너지 분야의 데이터를 대상으로 발생하는 오류들을 수정한다. 입력된 문자열에 대한 토큰분리를 위해 〈표 3〉과 같이 문자열과 동일한 크기의 배열을 선언한다. flag 값들은 모두 0으로 초기화되고, 분리에 사용된 구분자들에 의해 flag는 1 또는 2의 값을 갖는다. 1은 구분자의 끝을 의미하고 2는 구분자의 연속을 나타낸다(Sekine 2010).

토큰분리에 사용되는 구분자는 “(”, “)”, “ , ” 처럼 1문자부터 , “...”(말줄임표) 같이 여러 문자로 구성된다. 〈표 3〉에서와 같이 “(”과 “)” 구분자인 경우 flag의 값이 1로 지정되고, “...” 구분자인 경우 시작하는 2 문자에 대한 flag값은 2로, 마지막 3번째 문자에 대한 flag값은 1로 지정된다.

3.2.3 품사부착기

품사부착기는 기계학습방법으로 CRF++을 사용하고, Penn Treebank의 Brown 말뭉치를 9:1의 비율(학습: 991,410토큰/평가: 108,800토

〈표 3〉 입력문자열과 토큰분리 flag

src[ ]=	s	o	i	l	(	i	)	.	.	.	w	a	t	e	r
falg[ ]=	0	0	0	0	1	0	1	2	2	1	0	0	0	0	0

〈표 4〉 Penn Treebank의 구성

말뭉치	내용
WSJ	Wall Street Journal articles
Brown	The Brown Corpus
Switchboard	Telephone Conversations
ATIS	Air Travel Information System transcripts

〈표 5〉 언어처리엔진 정확률

	문장분리기	토큰분리기	품사부착기	기저구인식기
사용된 말뭉치	Brown, WSJ	Brown	Brown	Brown
정확률	99.98%	99.98%	96.19%	96.85

큰)로 분리하여 학습 및 실험을 수행한다(Penn Treebank).

품사부착을 위해 선택된 자질은 이전 단어의 품사자질, 현재 단어의 어휘자질, 그리고 현재 단어의 접사자질을 사용한다. 기계학습의 특성상 학습말뭉치에 존재하지 않은 단어, 미등록어에 대해서 높은 성능을 보이지 않는 문제점을 해결하기 위해, 본 연구에서는 접사자질을 추가하여 96.2%의 정확률을 나타낸다.

3.2.4 기저구인식기

〈표 4〉는 언어처리엔진의 기계학습모델을 학습 및 테스트하기 위해 사용한 말뭉치이다. Penn Treebank는 자연어에 대한 언어학적 구조를 분석하기 위한 목적을 위해 진행된 프로젝트의 산출물로 문장의 구문구조와 의미론적 정보, 품사정보를 담고 있다.

기저구인식기는 품사부착기에서 사용한 자질 외에 POS(품사정보), InitCap(첫글자가 대문자) 자질을 선택하여 이전 2개 단어, 현재, 다음 2개 단어와 이전 2개 품사, 현재품사, 다음 2개 품사에 자질들을 적용한다.

하부언어처리 모듈에 대한 성능을 평가하기 위해 기계학습기반을 사용한 품사부착기와 기저구인식기에 대해 말뭉치의 1/10을 학습을 위해 사용하고, 9/10를 테스트를 위해 사용한다. 〈표 5〉는 본 연구에서 개발한 언어처리엔진에서 사용한 말뭉치 종류와 정확률을 보여준다. 4개의 모듈 모두 높은 정확률을 나타낸다.

3.3 개체명 인식기

MUC-6와 CoNLL 2002, 2003에서 제안된 개체명 인식 시스템들은 〈표 1〉에서 언급한 BIO 태그를 이용한 기계학습 방법을 이용한다. 기계학습 방법의 경우 자질의 수와 정확률이 항상 비례하지 않고, 자질의 조합에 따라 성능이 달라지는 경우가 많으므로 모든 자질들을 동시에 사용하지 않는다. 본 연구에서는 정보이득(information gain)을 이용하여 필요한 자질을 선택한다. 〈표 6〉은 선택된 자질 집합을 보여준다.

〈표 6〉에서 보는 바와 같이 사전(Dic)자질과 WordNet 자질을 제외하고 다른 자질들은 3.2절에서 설명한 언어처리엔진에서 사용하는 자질

〈표 6〉 개체명 인식의 자질 집합

자질 이름	비 고	적용 범위
Word	토큰 그대로의 자질	w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, w_{i-1}/w_i, w_i/w_{i+1}
POS	토큰들의 품사	p_{i-2}, p_{i-1}, p_i, p_{i+1}, p_{i+2}, p_{i-1}/p_i, p_i/p_{i+1}, $p_{i-2}/p_{i-1}/p_i$, $p_{i+1}/p_i/p_{i+1}$, $p_i/p_{i+1}/p_{i+2}$
Word/POS	토큰과 품사의 조합	w_i/p_i
BNP	기저 명사구	n_{i-2}, n_{i-1}, n_i, n_{i+1}, n_{i+2}, n_{i-1}/n_i, n_i/n_{i+1}, $n_{i-2}/n_{i-1}/n_i$, $n_{i+1}/n_i/n_{i+1}$, $n_i/n_{i+1}/n_{i+2}$
Suffix3	토큰의 접미문자 3개	w_i
Prefix2	토큰의 접두문자 2개	
InitCap	첫 문자가 대문자 인지	
Dic	개체명 사전에 존재여부	
WordNet	WordNet 의미분류에 존재여부	

들과 동일하다. 개체명 인식을 위해 사전을 사용하는 경우, 시스템 전체가 기계학습 기반이므로 사전에 존재유무 자체를 하나의 자질로 인식하여 시스템을 구현하여야 한다. 하지만 학습 말뭉치의 경우 형태소 단위로 만들어져 있고, 개체명은 구 단위로 이루어져 있어서 바로 적용할 수 없다. 본 연구에서는 "BIPLO" 템플릿을 제작하여 이를 해결하도록 한다. "BI"는 개체명의 처음과 중간을 나타내고, "PLO"는 인명, 지명, 기관명을 구분하는 태그이다.

일반적으로 기계학습 기반 엔진들은 학습데이터가 충분할 경우 좋은 성능을 보여준다. 개체명 인식 성능 향상을 위해 〈표 7〉에서 보는 바와 같이, OntoNote 말뭉치, MUC 말뭉치 등 다양한 언어자원을 사용하여 개체명 사전을 구축한다.

〈표 7〉에서 사용한 외부 말뭉치들은 개체명 정보를 포함하지만, 본 연구의 개체명 인식기에서 사용한 자질들을 위한 문장분리 정보, 토큰분리 정보, 품사 정보가 결여되어 있다. 개체명 인식기의 성능을 높이기 위해, 언어처리모

〈표 7〉 구축된 개체명 사전의 표제어 수

말뭉치		PER	LOC	ORG
OntoNote	ACE	147	123	46
	CNN	656	384	370
	MNB	88	36	34
	NBC	92	99	55
	PRI	307	217	176
	VOA	402	293	230
	WSJ	2,449	1,040	2,983
MUC		2,888	1,157	2,759
ACE		7,579	976	4,592
Wikipedia		18,424	16,010	14,111
총계		33,032	20,335	25,356

듈을 이용하여 문장분리 정보, 품사 정보 등을 외부 말뭉치에 추가하여 본 연구에서 사용할 수 있도록 확장한다.

〈표 8〉은 WordNet 말뭉치에서 추출된 기본 의미 집합이다. WordNet은 많은 의미 정보를 포함하고 있으며, 의미 정보는 개체명 인식에 많은 도움을 준다(Muller 1995). 본 연구에서는 WordNet의 모든 단어(147,816단어)의 의미를 28개의 기본 의미 집합으로 구분하였다. 기본 의미집합을 결정한 기준은 WordNet의 각 노드에 포함된 단어 수를 기준으로 결정하였으며 추출된 노드에 포함된 하위 단어의 수가 균등하게 분포되도록 한다.

WordNet의 모든 단어들을 이 기본의미집합에 적용한다. 하나의 단어는 여러 개의 의미를 동시에 포함할 수 있으므로, 28개의 의미 중 자신이 포함되는 영역들의 자질을 '+'로 할당한다.

개체명 인식기의 성능평가는 〈표 9〉에서 사용

〈표 8〉 WordNet 말뭉치에서 추출된 기본 의미 집합

구 분	분 류 명
3단계 계층 6개	thing
	object, physical object
	causal agent, cause, causal agency
	substance, matter
	process, physical process
	abstraction
4단계 계층 중 가장 많이 나오는 22개	change
	freshener
	horror
	jimdandy, jimhickey, crackerjack
	security blanket
	stiriker
	whacker, whopper
	living thing, animate thing
	phychological feature
	whole, unit
	group, grouping
	attribute
	communication
	location
	measure, quantity, amount
	part, piece
	relation
	agent
	material, stuff
	food, nutrient
	compound, chemical compound
	solid

〈표 9〉 개체명 인식기 성능 평가

구 분		기본 시스템	개체명 사전 적용	WordNet 사전 적용
실제 NE의 수(개)		3,986		
시스템이 낸 NE의 수(개)		3,556	3,729	3,817
맞은 NE의 수(개)		3,229	3,418	3,527
재현율(%)		81.00	85.75	88.48
정확률(%)		88.32	91.66	92.40
F1-점수(%)		84.50	88.60	90.40
기본 시스템에 대한 개선율(%)	F1-점수	-	4.10	5.90
	오류율	-	26.45	38.06

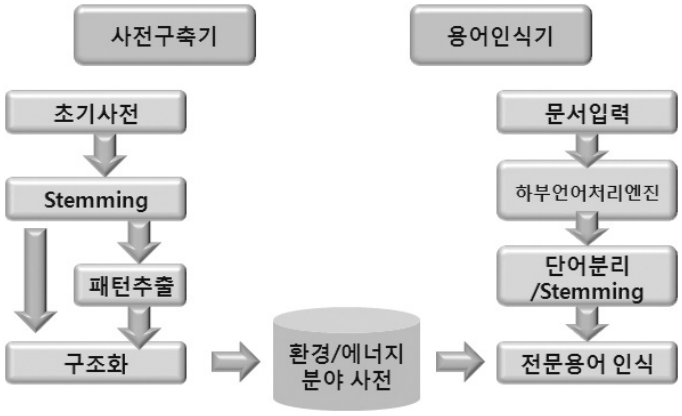
한 기계학습용 자질 중에서 기본적인 자질만을 사용, 개체명 사전 추가, WordNet, 자질 추가한 3가지의 경우에 대해서 각각 수행한다. 〈표 7〉의 말뭉치 중 OntoNote 2(WSJ)를 학습/실험을 위한 말뭉치로 선택한다. 이 말뭉치를 2:1의 비율로 분리하여 학습과 실험 말뭉치로 사용한다.

〈표 9〉는 개체명인식의 실험결과이다. 기본 자질만을 사용한 경우 84.50의 F1점수를 보여주었고, 개체명 사전을 적용하였을 때 기본 시스템에 비해 약 2%, WordNet, 자질을 추가한 경우 약 6%의 성능개선을 나타낸다.

3.4 전문용어 인식기

전문용어는 핵심개체 중 기술명을 의미한다. 대상 데이터를 환경/에너지 분야로 선정하였기 때문에, 환경/에너지 분야의 전문용어들을 수집하여 사용한다.

전문용어 인식기는 사전모델(Automata)학습 부분인 사전구축기와 인식시스템으로 구성한다. 〈그림 3〉은 전문용어 인식 시스템 구성도이다. 학습시스템의 경우 초기 사전을 정제하여 패턴을 추출한 뒤, 오토마타를 이용하여 문자열



〈그림 3〉 전문용어 인식 시스템

값의 일부만을 사용하여 분기하는 트리형태의 자료구조인 트라이(TRIE) 형태로 구조화 한다(DARTS: Double-ARray Trie System). 인식 시스템은 하부 언어처리엔진을 이용하여 추출된 용어후보들을 들을 정규화하고 학습 시스템에서 생성된 사전 오토마타를 이용하여 전문용어로 인식한다.

3.4.1 사전구축 모듈

사전구축시스템은 전문용어의 인식율을 높이기 위하여, 초기 전문용어집합에 대한 정규화 작업을 수행한다. 단어의 변형과 단어의 구성에 대한 2가지 정규화 작업이 있다. 첫 번째는 전문용어의 단/복수, 동명사, 과거분사 등의 여러 변형을 스테머를 이용한 정규화 작업을 거쳐 단일형태로 변환한다. 두 번째는 전문용어의 이형태를 인식할 수 있도록 정규화하는 작업이다. 예를 들어 "P12"라는 단백질 용어가 초기전문용어 집합에 포함되어 있다면, 포함되어 있지 않은 "P13", "P14" 등의 전문용어도 인식하도록 정규화하는 작업이다.

첫 번째 정규화 작업을 위해 전문용어에 적합한 스테머를 제작한다. 스테머는 포터-2 알고리즘을 기반으로 제작하고, 접미사 목록에 본 연구의 대상 데이터인 환경/에너지 분야와 관련한 접미사(〈표 10〉 참조)를 추가한다.

두 번째 정규화작업을 위해 정규표현을 이용한 규칙을 제작하여 이용한다. 초기 사전을 분석하여 생성된 규칙은 다음과 같다.

1) 숫자와 로마숫자는 자리수를 맞추어 #으로 치환한다.
   예) P13 => P##
2) 특수기호는 "-"으로 치환한다. 연속되는 특수기호는 하나로 통합한다.
   예) a$b => a-b
3) 특별한 의미를 지닌 용어는 @로 치환한다.
   예) GeneAlpha => Gene@
4) 단어사이의 모든 공백은 "_"로 치환한다.
   예) animal waste => animal_waste

위와 같이 정제된 단어목록을 오름차순으로 정렬한 뒤 트라이 구조로 저장한다.

〈표 10〉 추출된 접미사

접미사	빈도수	접미사	빈도수
aceae	171	ator	59
virus	130	osis	57
ine	127	mycin	57
ium	102	amide	57
itis	81	ates	56
ase	63	ceae	54
viridae	62	transferase	50
idae	61	ography	50
amine	61	...	...

〈표 11〉 스택을 이용한 검색 알고리즘의 예제

D	i	s	o	r	d	e	r	i	n	g		B	r	a	i	n		G	i	v	e	s		C	l
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25

u	e	s		T	o		B	r	a	i	n		D	i	s	o	r	d	e	r	i	n	g
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49		

3.4.2 전문용어 인식 모듈

전문용어 인식모듈은 입력된 문서에서 구축된 사전을 이용하여 단어를 인식하는 모듈이다. 전문용어 인식은 실시간으로 이루어지기 때문에, 인식속도가 중요한 요소가 된다. 본 연구에서는 인식속도를 향상시키기 위하여 스택을 이용한 병렬 검색 방법을 제안한다.

병렬 검색방법은 공백이 발생할 때마다 스택에 새로운 용어 인식 후보를 추가하여, 스택에 쌓여 있는 모든 단어들을 병렬로 사전과 비교하는 방법이다. 사전과 일치하지 않는 단어는 스택에서 제거되고, 다음 단어로 포인터가 한 번 움직일 때마다 스택 내부에 존재하는 모든 후보들을 검증하여 사전에 없는 경우에는 제거하고, 사전에 있는 경우에는 전문용어로 인식한다.

예를 들어, 사전에 "Brain", "Brain Disordering"가 등록되어 있는 경우, 〈표 11〉에서 Brain(13-17)은 스택에 저장된 후 Gives(19-22)에 대한 처리를 할 때, 더 이상 일치하는 단어가 없으므로 스택에서 제거된다. Brain(33-37)은 스택에 저장된 후 연속되는 단어인 Disordering(39-49)와 복합된 단어로 사전에서 검색되므로 최장일치 단어인 Brain Disordering이 전문용어로 인식된다.

4. 결 론

본 연구에서는 독립적으로 개발되어오던 개체명 인식 기술과 전문용어 인식 기술을 하나의 작업으로 통합하여, 과학기술문헌에 포함된 핵심개체인 개체명(named entity)과 전문용어(terminology)를 통합 인식하는 시스템을 개발하였다.

대상 데이터의 분야는 한국과학기술정보연구원이 보유하고 있는 과학기술데이터 중 환경/에너지 분야로 선정하였고, 일반적인 개체명과 분야 종속적인 전문용어들에 대한 분류를 핵심개체로 정의하였다.

개체명 인식, 전문용어인식에 대한 세계적인 연구방향과 수준을 파악하고, 개체명 인식을 위해 기계학습방법 중 Conditional Random Field(CRF)를 선정하였고, 전문용어 인식은 사전기반 방법을 이용하였다. 문장분리기, 토큰분리기, 품사부착기, 기저구 인식기 등의 하부언어처리 엔진을 개발하여 통합플랫폼에 공통으로 사용하였다.

개체명 인식기는 기계학습을 위한 기본적인 자질 외에, 개체명 사전에의 존재여부 자질과, WordNet의 의미정보 자질을 추가하여 개체명 인식 성능을 향상하였다.

전문용어 인식기는 기계학습기법을 사용하기 힘든 환경/에너지 분야의 과학기술 문헌을 처리하기 위해 사전기반 인식방법으로 구현하였다. 사전기반 검색의 성능을 높이기 위해 기초사전에 대한 정규화 작업을 수행하여 사전을 구축하였고, 인식속도 향상을 위해 스택을 이용한 병렬 검색 알고리즘을 구현하였다.

과학기술핵심개체 자동인식 통합 플랫폼은 개체명 인식부분에서 비교적 높은 성능을 보여주었고, 환경/에너지 분야 데이터에 적용하여 개체명과 전문용어를 통합한 핵심개체를 추출하여, 개체명간의 관계, 전문용어간의 관계뿐만 아니라, 기존 연구에서는 해결하기 어려웠던 개체명과 전문용어간의 관계를 추출하기 위한 관계추출 시스템의 기반데이터로 활용하였고, 향후 과학기술 신규용어 탐지, 과학기술사전 자동구축, 의미 기반 정보검색, 관계 추출 기반 질의응답 등의 다양한 정보서비스 분야에 활용될 것으로 기대된다.

참 고 문 헌

국립국어원. 2007.『전문용어 연구』. 경기: 태학사.

김형철, 서형원, 김재훈, 최윤수. 2009. CRF를 이용한 대명사 참조해서 시스템.『한글 및 한국어 정보처리 학술대회 발표논문집』, 21(1): 87-91.

오종훈, 최기선. 2006. 기계학습에 기반한 생의학 분야 전문용어의 자동인식.『정보과학회 논문지: 소프트웨어 및 응용』, 33(8): 718-729.

이창기, 황이규, 오효정 등. 2006. Conditional Random Fields를 이용한 세부 분류 개체명 인식.『한국정보과학회 언어공학연구회 학술발표논문집』, 18: 268-272.

Ananiadou, S. and G. Nenadic. 2006. "Automatic terminology management in biomedicine." *Text Mining for Biology and Biomedicine*, 67-97.

Black, W. J. and A. Vasilakopoulos. 2002. "Language-independent named entity classification by modified transformation- based learning and by decision tree induction." *Proceedings of CoNLL* 2002: 159-162.

Carreras, X., L. Màrques, and L. Padró. 2002. "Named entity extraction using AdaBoost." *Proceedings of CoNLL* 2002: 167-170.

Chang, J. T., H. Schutze, and R. B. Altman. 2004. "GAPSCORE: Finding gene and protein names one word at a time." *Bioinformatics*, 20(2): 216-225.

Collins, M. 2002. "Ranking algorithms for named-entity extraction: boosting and the voted perceptron." *Proceedings of ACL 2002.*

CoNLL(Conference on Computational Natural Language Learning). [online]. [cited 2010.12.15]. 〈http://www.cnts.ua.ac.be/conll2002/ner/〉. 〈http://www.cnts.ua.ac.be/conll2003/ner/〉.

CRF(Conditional Random Fields). [online]. [cited 2010.12.15]. 〈http://en.wikipedia.org/wiki/Conditional_random_field〉.

DARTS(Double-ARray Trie System). [online]. [cited 2010.12.15]. 〈http://chasen.org/~taku/software/darts/〉.

Elsner, M. and E. Charniak. 2007. A Generative Discourse-New Model for Text Coherence. Tech Report CS-07-04.

Extended Named Entity Hierarchy. [online]. [cited 2010.12.15]. 〈http://nlp.cs.nyu.edu/ene/〉.

Grosz, B. J., A. K. Joshi, and S. Weinstein. 1995. "Centering: A framework for modeling the local coherence of discourse." *Computational Linguistics*, 12(2): 203-225.

Lafferty, J., A. McCallum, and F. Pereira. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." *Proceedings of International Conference on Machine Learning*, 282-289.

LDC. 2008. ACE(Automatic Content Extraction) English Annotation Guidelines for Entities, ver 6.6, Linguistic Data Consortium.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz. 2004. "Building a large annotated corpus of English: The Penn Treebank." *Computational Linguistics*, 19(2): 313- 330.

Miller, G. A. 1995. "WordNet: A lexical database for English." *Communications of the ACM*, 38(11): 39-41.

MUC(Message Understanding System). [online]. [cited 2010.12.15]. 〈http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html〉.

Penn Treebank(The Penn Treebank Project). [online]. [cited 2010.12.15]. 〈http://www.cis.upenn.edu/~treebank/〉.

Sasaki, Y., S. Montemagni, P. Pezik, D. Rebholz-Schuhmann, J. McNaught. and S. Ananiadou. 2008. "BioLexicon: A lexical resource for the biology domain." *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine*.

Sekine S.(2010) Personal Communications. [online]. [cited 2010.12.15]. 〈http://nlp.cs.nyu.edu/oak/〉.

Tanabe, L. and W. J. Wilbur. 2002. "Tagging gene and protein names in biomedical text." *Bioinformatics*, 18(8): 1124-1132.

The Gene Ontology Consortium. 2008. "The

Gene ontology project in 2008." *Nucleic Acids Research*, 36(Database issue): D440-D444.

Tjong Kim Sang, E. F. 2002. "Memory-based shallow parsing." *Journal of Machine Learning Research*, 2(March): 559-594.

Watanabe, Y., M. Asahara, and Y. Matsumoto. 2007. "A Graph-based approach to named entity categorization in Wikipedia using conditional random fields." *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 649-657.

Weischedel, R., S. Pradhan, L. Ramshaw, T. Khleif, M. Palmer, N. Xue, M. Marcus, A. Taylor, C. Greenberg, E. Hovy, R. Belvin, and A. Hoston. 2007. OntoNotes Release 2.0, BBN Technologies.

Zhang, J., D. Shen, G. Zhou, J. Su, and C. L. Tan. 2004. "Enhancing HMM-based biomedical named entity recognition by studying special phenomena." *Journal of Biomedical Informatics*, 37(6): 411-422.

Zhou, G., J. Zhang, J. Su, et al. 2004. "Recognizing names in biomedical texts: A machine learning approach." *Bioinformatics*, 20(7): 1178-1190.