ISBN 978-89-6211-480-593560

과학기술핵심개체 탐지를 위한 테스트컬렉션 구축과 정보추출기술 개발 및 성능 평가

최윤수(KISTI)
최성필(KISTI)
정창후(KISTI)
윤화묵(KISTI)
류범종(KISTI)

ISBN 978-89-6211-480-593560

과학기술핵심개체 탐지를 위한 테스트컬렉션 구축과 정보추출기술 개발 및 성능 평가

최윤수(KISTI)
최성필(KISTI)
정창후(KISTI)
윤화묵(KISTI)
류범종(KISTI)

KiSTi 한국과학기술정보연구원
Korea Institute of Science and Technology Information

목 차

1. 서론

인터넷의 발달과 더불어 대용량 데이터를 실시간으로 처리하여 필요한 지식을 발견하기 위한 정보추출 기술들이 핵심적인 분야로 인식되고 있다. 정보추출은 크게 (1) 개체명 인식(named-entity recoginition), (2) 대용어 참조 해소(coreference resolution), (3) 관계 추출(relation extraction)의 세 가지 요소기술로 세분화 된다. 본 연구에서는 이 중 개체명 인식과 대용어 참조해소 부분을 위한 테스트컬렉션을 구축하고 관련 기술을 통합 개발한다.

개체명 인식은 문서내의 원소를 찾아서 기 정의된 범주로 분류하는 작업으로, 특정 도메인에 따라 개체명인식기(Named Entity Recognition System)가 개발된다. 개체명인식은 인명, 지명, 기관명 같은 일반적인 범주와 더불어, 생의학분야에서는 유전자와 유전자와 관련된 산출물 등과 같이 분야특화된 범주가 사용된다. BBN[1]은 29개의 범주와 64개의 하위범주를 제안하였고, Sekine은 200개의 하위범주를 갖는 확장계층도[2]를 제안하였다.

본 연구에서는 일반적인 범주인 인명, 지명, 기관명에 대한 부분만 개체명으로 정의하고, 생의학분야에 특화된 개체명은 기술용어 및 분야특화명으로 구분되는 전문용어로 분류한다.

대용어 참조해소는 한 명사구와 명사구를 참조하는 개체로 해소하는 작업으로 자연어의 하이퍼링크라 볼 수 있으며, 일반적으로 한번 출현한 단어의 반복을 피하기 위해 사용된다. 개체명과 전문용어가 인식된 후 이들 간의 관계추출 작업을 수행할 때 더 많은 자원을 확보하기 위해 대용어 참조해소는 중요한 작업이다. 본 연

1) http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html
2) http://nlp.cs.nyu.edu/ene

구에서는 대명사를 참조해소의 대상으로 한다.

현재까지 개체명인식분야는 신문기사 및 방송기사 등을 중심으로 연구되었고, 전문용어인식분야는 생의학분야에서 유전자, 단백질과 관련된 용어들에 대하여 많은 연구가 행해졌다. 개체명 인식기술과 전문용어인식기술이 서로 독립적인 영역에서 연구되었고, 전문용어분야에서도 그 범위는 매우 제한적이다.

과학기술문헌의 경우 일반적인 개체명과 전문용어가 혼재되어 있는 형태로 구성되므로, 기존의 연구결과물을 이용하여 접근하기 위해서는 2단계로 작업을 수행해야 하기 때문에, 그 결과물을 통합하는 과정의 불편함이 있고, 처리속도에서 많은 제약이 따른다. 본 연구에서는 개체명과 전문용어를 통합하여 핵심개체로 정의하고, 대상 문헌으로 한국과학기술정보연구원이 보유하고 있는 해외학술지 데이터베이스 중에서 생의학분야 문헌과 과학기술 뉴스기사를 수집하였다. 수집된 문헌들을 이용하여 테스트컬렉션 구축 작업을 수행하였고, 이를 대상으로 개체명과 전문용어를 동시에 인식하고, 이를 참조하는 대용어에 대한 참조해소를 수행하는 통합된 플랫폼을 구현한다.

2. 관련연구

2.1 테스트컬렉션 구축

정보추출 분야에 기계학습 모델을 이용하기 위해서는 학습과 테스트 및 평가를 위한 테스트컬렉션 구축이 필수적인 작업이다. 1990년대 초반 미국 정보는 정보 추출 연구의 평가 및 활성화를 위해 MUC(Message Understanding Conference)을 주관하였고, 이는 ACE 프로젝트로 발전하였다.

<표 1>은 MUC7와 ACE 2005에서 개체명 인식을 위해 사용된 분류이다. MUC에서 ACE로 발전하면서 개체명에 관한 새로운 분류가 추가되고, 더 상세하게 변화됨을 알 수 있다. MUC와 ACE에서 테스트컬렉션 구축을 위해 사용한 자료는 방송기사, 신문기사, 웹로그, 유즈넷, 전화통화 등으로, 일반인들이 쉽게 접근할 수 있는 분야로 구성된다.

<표 1> 개체명 인식 분류

분류	MUC7	ACE 2005
entity names (개체명)	persons, organization, locations	persons, organizations, locations, facilities, geographical/social/political entities, vehicle, weapon
temporal expressions (시간 표현)	dates, times	date, times
value	monetary values,	contact-Info, numeric, crime, job-title, sentence

	percentages	
event	-	life, movement, transaction, business, conflict, contact, personnel, justice

인명, 지명, 기관명으로 대표되는 개체명인식과 달리, 생의학분야에서는 생의학 관련 용어들에 대한 인식과 관련된 많은 연구들이 수행되었다. 생의학분야에서는 주로 유전자와 단백질과 관련된 전문용어를 추출하는 기술에 관한 연구가 수행되었고, <표 2>는 생의학분야에서 대표로 인식되는 테스트컬렉션이다.

<표 2> 생의학분야 전문용어 인식

컬렉션	AIMed	BioInfer	HPRD50	IEPA	LLL
건 수	1,955	1,100	145	486	77
분 류	Protein Gene	Protein Gene RNA	Protein Gene	Chemicals	Protein Gene

2.2 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] .

개체명인식은 정보추출(Information Extraction)의 한 분야로써 문서내에서 개체명을 추출하고 기정의된 분야(인명, 지명, 기관명 등)로 분류하는 작업을 말한다. 개체명 인식에 관한 연구는 MUC-6(Message Understanding Conferences)[3]에서 유래되어 최근에는 개체명을 약 200여 개로 나누어 질의응답 시스템의 문장 분석에 널리 사용되고 있다.[4] MUC-6에 참가한 많은 시스템은 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(Conference on Computational Natural Language Learning) 2002[5]와 2003[6]을 통해서 많은 발전이 있었다. 이 대회에 참가한 대부분의 시스템은 기계학습 방법을 이용하였으며 영어의 경우에 약 89%의 정확률을 보인다. 기계학습 방법에서는 주로 BIO 태그(B: 개체명의 시작, I: 개체명의 중간, O: 관계없음)를 이용하는데 <표 3>은 CoNLL에서 사용한 예이다.

<표 3> CoNLL에서 사용된 BIO 태그

토큰	BIO 태그	비고
U.N.	I-ORG	기관명으로 인식
official	O	
Ekeus	I-PER	인명으로 인식
heads	O	
for	O	

3) http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
4) Extended Named Entity Hierarchy, http://nlp.cs.nyu.edu/ene/
5) http://www.cnts.ua.ac.be/conll2002/ner/
6) http://www.cnts.ua.ac.be/conll2003/ner/

Baghdad	I-LOC	지명으로 인식
.	O	

(Black, Vasilakopolos 2002)는 90년대 초반 Eric Brill이 처음 소개한 변환기반 학습(Transformation-based learning)을 이용하였다[2]. 변환기반 학습은 형태소 부착, 영어 전치사구 접속문제해결, 기저구 인식, 철자 수정 등 자연어처리의 다양한 분야에 사용된 기계학습기법으로 첫 번째 단계에서 개체명을 인식하는 작업을 수행하고, 두 번째 단계에서 인식된 개체명 후보들에 대한 적합한 분류를 제공하는 방법이다.

(Carreras, Marques, Padro 2002)는 CoNLL 2002에서 개체명인식을 위해 잘 알려진 BIO 태그 외에, 개체명의 시작과 종료 경계를 인식하기 위한 Open-Close&I 과 Global Open-Close 방식을 도입하였고, 개체명분류를 위해 AdaBoost 이진 분류 알고리즘을 사용하였다[4].

(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 웹데이터를 이용하였고 ME 태거의 85.3%의 F1-척도에 비해 87.9%의 F1-척도를 보여주었다[6].

(Watanabe, Asahara, Matsumoto 2007)은 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[21].

2.3 전문용어 인식 기술

전문용어는 초기에 용어사전을 이용해서 관리되고 보급되어 왔다. 그러나 과학기술의 급속한 발전으로 전문분야와 전문용어도 또한 급속히 생성되고 있어 용어사전만으로는 도저히 전문용어를 파악하여 관리할 수 없는 상황이다(오종훈, 최기선, 2006)[27].

과학기술문헌 중 생의학분야의 전문용어들은 문맥에 따라 상이한 대상을 나타낼 수 있고(예, ferritin은 단백질 또는 실험을 의미한다), 한 개체가 여러 개의 전문용어로 표현될 수 있고(예, PTEN, MMAC1은 동일한 유전자를 의미한다), 신규용어들이 신속하게 생성되므로 전문용어 인식이 더욱 어렵다.[7)]

생의학 분야의 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

7) 본 연구에서 대상 데이터를 과학기술분야 중 생의학 분야로 제한하였기 때문에, 생의학 분야 전문용어 추출기술에 초점을 맞춘다.

entropy model) 등이 있다.

(Tanabe, Wilbur 2002)는 규칙기반 접근방법을 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 66.7를 보여준다[18].

(Chang, Schutze, 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[5].

(Zhou, Zhang, Su 2004)는 은닉마코프 모델(HMM : Hidden Markov Model)을 이용하였다. 대문자시작 정보, 접두사와 접미사 정보, 품사정보, 시작 단어 등을 전문용어후보 추출을 위한 자질로 선택하였고, GENIA 말뭉치 2.1을 사용하여 학습및 실험을 한 결과, 66.5%의 정확률과 66.6%의 재현율을 보였다[25].

현재까지 개발된 생의학 분야 전문용어 인식기술은 전체적으로 F1-척도 75%와 85% 사이를 보여준다.

2.4 대용어 참조해소 기술

대용어는 대용의 기능을 담당하는 어휘로서 선행어보다 간결한 형식을 사용하여 반복되는 성분을 대신하는 것으로 명확한 진술이나 단어의 반복적인 사용을 피하고자할 때 사용하는 말이며, 대명사가 그 대표적인 예이다. 문장의 의미를 정확히 파악하기 위해서

는 문장에 사용된 대용어가 이전 문장/대화의 어떤 사물/행위를 가리키는지를 구별하여야 하는데 이러한 과정을 대용어 처리 혹은 참조해소라 한다.

대용어 처리에 대한 연구는 꽤 오랜 역사를 가지고 있으나 실용화된 시스템을 사용하는 경우는 거의 없다. 과거에는 중심화 이론(centering theory)를 바탕으로 하는 규칙기반의 연구가 주로 진행되었으나 최근에 와서는 기계학습을 이용한 대용어 처리에 관한 연구도 활발히 진행되고 있다(Elsner and Charniak, 2007). 대부분의 연구들의 성능은 뛰어나지 않지만 관계추출분야에서 대용어를 그대로 사용할 수 없고, 대용어가 가리키는 개체를 정확히 파악해서 이 정보를 이용한다면 관계추출의 성능을 크게 개선할 수 있다.

대용어 참조해소 방법은 규칙기반에 의한 방법(Grosz, Joshi 1995),(Carbonell, Brown 1988), (Lappin, H. Leass, 1994)과 기계학습에 의한 방법(Soon, Ng, Lim 2003)(Yang, Su, Tan 2008)이 있다[3.9,12].

규칙기반에 의한 방법은 대용어에 대하여 대용어 후보들을 수집하여 성(gender), 수(number), 등의 제약조건을 만족하는 후보들 중에서 경험적으로 가능성이 높은 후보를 선택한다. 규칙기반 시스템의 경우 여러 유형의 대용어 관계를 규칙화하는 것이 쉽지 않다는 점에서 많은 제약이 있다[12].

(Lapain & Leass, 1994)는 RAP(Resolution of Anaphora Procedure)를 제안하였다. RAP는 제약조건에 만족하는 후보들 중 구의 속성이나 위치 문장거리 등에 대해 상이한 가중치를 할당하여 가장 높은 점수를 획득한 후보를 선택하는 시스템이다.

기계학습 방법은 학습 말뭉치로부터 대용어 관계를 결정하는 규칙을 학습하는 접근방법이다. 대용어와 대용어가 태깅된 학습말뭉

치로부터 유용한 자질들을 선택하고 추출하여, 모델을 학습한다. 학습된 모델을 만드는 방법은 크게 두 가지로 구분된다. 첫 번째 방법은 대용어와 선행어 후보가 대용 관계에 속하는지를 결정하는 이진 분류 방법(Soon, Ng, Lim 2003), (Pnnzetto, Strube 2006) 이다[17]. 두 번째 방법은 선행어 후보들에 대해서 점수를 계산하고 그 점수에 따라 선행어 후보들을 순위화하여 가장 높은 순위에 있는 후보를 선택하는 방법(Lappin, Leass 1994)이다[12].

(Kilicaslan, Guner, Yildrim 2009)는 SVM(Support Vector Machine)을 이용한 터키어의 참조해소 시스템을 제안하였으며 약 73%의 정확율을 보인다. (Denis, Baldridge 2009)는 순위화 모델을 제안하였고, 영어에 대하여 72.4%의 정확율을 보인다[7,10]. (Nguyen, Kim , Tsujii 2008)은 GENIA 말뭉치에 대하여 최대 엔트로피 모델을 이용해서 약 71.43%의 정확도를 보인다[15].

3. 과학기술 핵심개체 신규탐지 시스템

3.1 핵심개체 정의

본 연구에서 대상 데이터를 생의학분야 문헌으로 선정하였고, 생의학분야 문헌에서 중요한 역할을 담당하는 개체명들과 전문용어들을 통합한 핵심개체들을 <표 4>와 같이 10개의 분류로 정의하였다.

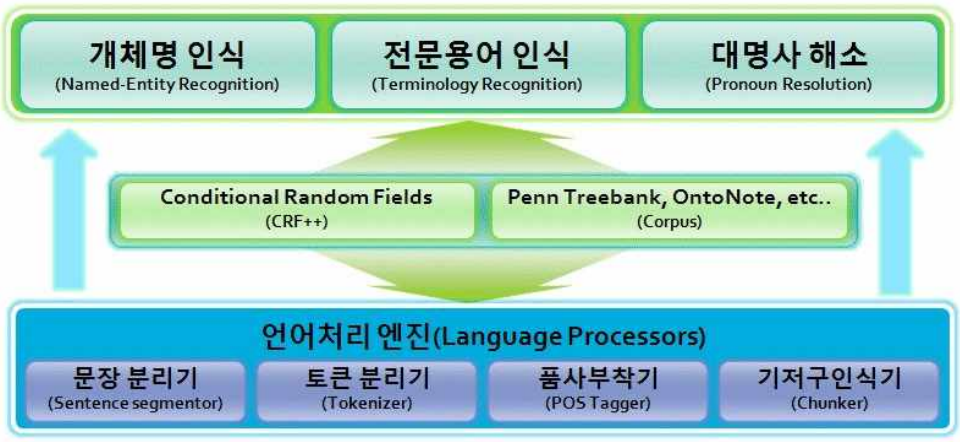
<표 4> 핵심개체 분류표

상위분류	상세분류 (범주)	비고
일반개체명	Person	인명
	Location	지명
	Organization	기관명
기술명	TechTerm	기술용어(장비, 치료법, 수술 등)
	Others	기타(중요한 용어이지만 다른 분류에 속하지 않는 용어)
분야특화명	Gene	유전자
	Protein	단백질
	Disease	질병
	Organism	유기체
	Drug	약명

일반개체명은 개체명인식에서 주로 사용되는 인명(Person), 지명(Location), 기관명(Organization)으로 분류하였으며, 전문용어는 생의학분야에 특화된 분야특화명과 장비, 치료법, 수술 등에 포함되는 기술용어와 기타로 분류하였다.

3.2 전체 시스템 구성도

과학기술문헌에서 개체명과 전문용어를 인식하고, 대용어에 대한 참조해소를 수행하는 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 인식기, 대용어 참조해소기는 하부의 언어처리과정에서 추출된 각각의 자질들 또는 통합된 자질들을 사용한다.



<그림 1> 핵심개체 자동인식 시스템 구성도

전체적인 작업은 (1)언어처리 단계, (2)전문용어 및 개체명 인식 단계, (3)대용어 참조해소 단계)로 진행된다.

첫째, 언어처리 단계에서는 입력된 문장을 문서단위로 분리하고, 의미있는 단위로 토큰을 인식하고, 토큰에 대한 적절한 품사를 부착하고, 기저명사구를 인식한다.

둘째, 개체명인식기에서 인명, 지명, 기관명을 추출하고, 전문용

어 인식기에서 기술명(2개분야) 및 분야특화명(5개분야)을 추출한 뒤, 그 결과를 하나의 문서로 통합한다.

셋째, 대용어 참조해소기에서 문서내에서 출현하는 대명사들을 대상으로 선행어를 발견하는 작업을 수행한다. 선행어가 발견된 대명사에 대해서 선행어에 대한 정보(개체명, 전문용어)를 동일하게 적용하고, 선행어에 대한 참조정보를 함께 기록한다.

4. 테스트컬렉션 구축(KEEC 2009)

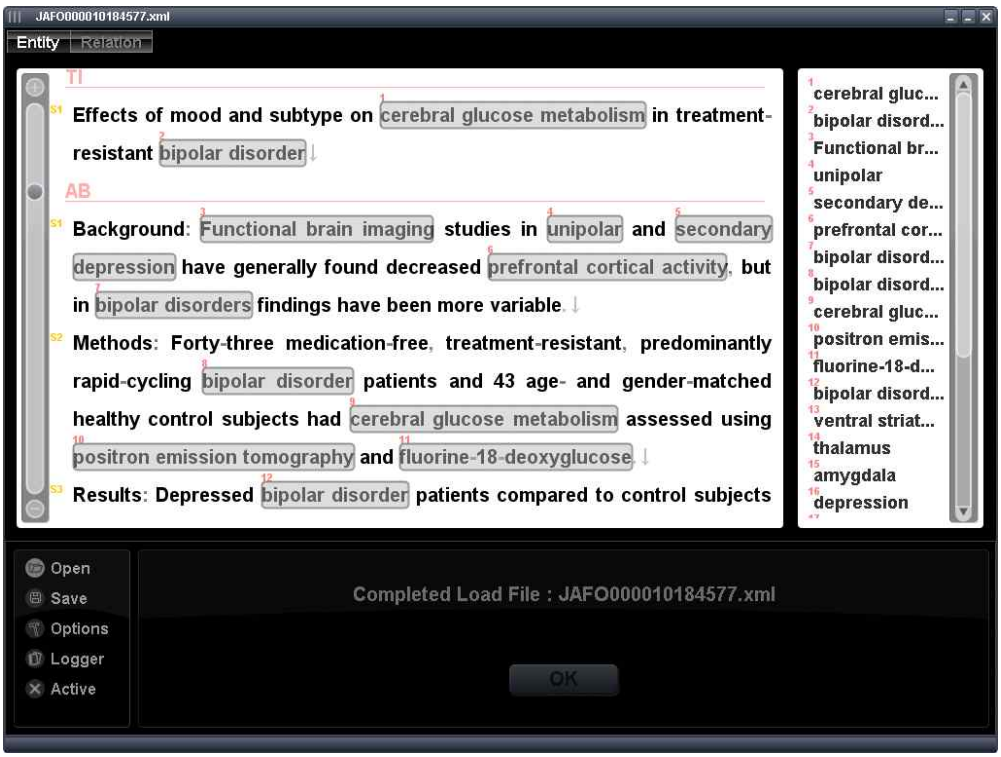
본 연구에서 핵심개체 인식 대상을 과학기술문헌 중에서 생의학 분야로 한정하였다. 생의학 분야에 관한 문헌은 과학기술 뉴스와 한국과학기술정보연구원에서 보유하고 있는 해외학술지로 선정하였다. 기계학습 모델을 이용한 핵심개체 인식 기술을 위해 학습과 테스트를 위한 테스트컬렉션(KISTI Entity Extraction Collection 2009-KEEC 2009)를 다음과 같이 구축하였다.

<표 5> 테스트컬렉션 DTD

doc.dtd

```
<?xml version="1.0"    encoding="EUC-KR"?>

<!-- 문서(DOC)는    내용(TEXT)과 내용 안에 존재하는 관계들의 집합
(NRLIST)으로    구성된다. -->
<!ELEMENT DOC (TEXT, NRLIST)>
<!ATTLIST DOC
        did       CDATA    #REQUIRED>    <-- 문서 아이디    -->

<!-- 내용(TEXT)은    제목(TI)과 초록(AB)으로 구성된다. -->
<!ELEMENT TEXT (TI, AB)>

<!-- 제목(TI)은    문장(S)으로 구성된다. -->
<!ELEMENT TI (S+)>

<!-- 초록(AB)도    문장(S)으로 구성된다. -->
<!ELEMENT AB (S+)>

<!-- 문장(S)는    과학기술핵심개체(NE)로 구성된다. -->
<!ELEMENT S (NE*)>
```

```
<!-- 과학기술핵심개체(NE)는    실제 개체를 태깅한다. -->
<!ELEMENT NE (#PCDATA)>
<!ATTLIST NE
        eid       CDATA    #REQUIRED     <!-- 개체 아이디 -->
        co_ref    CDATA    #IMPLIED      <!-- 대용어 참조 -->
        class     CDATA    #REQUIRED     <!-- 미리 정의된 개체 -->
        nn        CDATA    #REQUIRED>    <!-- 대상 개체의 원형    -->
```



<그림 2> 테스트컬렉션 구축 지원 도구

과학기술 뉴스는 해외사이트[8]에서 생의학분야 중 2000년도 이후의 문서에서 문서크기가 상위 80% 이상에 해당되는 문서들을 년도

8) http://www.eurekalert.org/

별로 임의로 선정하여, 전체 11,185건을 수집하였다.해외학술지는 SCI급, 인용지수, 초록크기 등 여러 가지 요소들을 조합하여 수집하였다. 첫째, 해외학술지 중에서 인용지수(Impact Factor)를 기준으로 상위 50종을 우선 선별하였다. 둘째, 동일한 종에서 개별 초록의 크기가 평균초록 크기의 90%이상인 문서를 선정하였다. 셋째, 발행년도가 2000년 이후인 최신 문서를 선정하였다. 넷째, 선정된 종에서 종별로 각 25%의 문서를 선정하여 최종적으로 10,310건의 문서를 수집하였다.

선정된 문서는 <표 5>의 DTD를 갖는 XML문서 형식으로, <표 4>에서의 분류를 사용하여 태깅작업이 수행된다.

테스트컬렉션 구축은 전문가 2인에 의해 수행되었고, 서로 교차하여 검토 비교하여 오류를 최소화 하였다. 테스트컬렉션 구축 시 발생되는 철자오류 및 태깅오류 등을 방지하고, 작업 속도를 높이기 위하여 <그림 2>와 같은 구축지원도구를 사용하였다.

<표 6> 테스트컬렉션 구축 결과

문서수	문장수	핵심개체	개체포함문장
354	8,303	21,142	7,032

테스트컬렉션 구축도구는 문장 분리 및 합병, 핵심개체 지정 및 취소, 핵심개체 추천, 분류코드 지정, 오류 검증 등의 기능 등을 제공하고, 완성된 문서를 <표 5>의 DTD에 맞는 XML문서로 저장한다. 또한 최종 작업에 대한 문서수, 문장수, 일일별 통계등을 제공한다. <표 6>은 본 연구에서 구축한 테스트컬렉션 결과이다.

5. 언어처리 엔진을 위한 말뭉치 및 사전 구축

기계학습 모델을 사용하는 경우, 학습을 위한 말뭉치(테스트컬렉션)의 규모가 성능에 큰 영향을 미친다. 본 연구에서 구축한 KEEC 2009는 핵심개체인식을 위한 개체명, 전문용어, 대용어에 대한 태깅을 부착한 테스트컬렉션이고, 이 기술의 바탕이 되는 언어처리 엔진(문장분리, 토큰분리, 품사부착, 기저구 인식 등)을 위한 말뭉치가 필요하다.

본 연구에서는 적합한 외부 말뭉치를 선정하고, 과학기술핵심개체 인식을 위한 형태로 변환작업을 수행하고, 관련 사전을 구축한다.

선정된 말뭉치는 Penn Treebank(Marcus et al, 2004), OntoNote(Weischedel, 2007), Message Understading Conference (MUC)(Chinchor 1997), Automatic Contect Extract (ACE)(LDC, 2008)이고 사전 구축은 크게 두 종류의 사전이 구축되었다. 하나는 Wikipedia 말뭉치와 ACE 말뭉치로부터 구축된 개체명 사전이고 다른 하나는 전문용어 인식을 위해 사용된 Gene Ontology(The Gene Ontology Consortium, 2008)이다. 이들 말뭉치의 사용 분야는 <표 7>과 같으며, 각각의 말뭉치에 대한 자세한 설명은 아래의 각 절에서 간단하게 기술한다.

<표 7> 사용 말뭉치의 현황

구 분	모델 학습	사전 구축
문장 분리	Penn Treebank	
토큰 분리	Penn Treebank	
품사 부착	Penn Treebank	
기저구 인식	Penn Treebank	

개체명 인식	OntoNote, Penn Treebank	MUC, ACE, Wikipedia
전문용어 인식	KISTI Term Corpus, Penn Treebank, GENIA Corpus	Gene Ontology, BioLexicon, MEDLINE
대명사 참조해소	OntoNote, Penn Treebank	

5.1 Penn TreeBank

Penn Treebank는 Penn Treebank Project의 산출물로 현재 3번째 버전까지 공개되어 있다. Penn Treebank Project는 자연어에 대한 언어학적 구조를 분석하기 위한 목적을 가진 프로젝트로서 문장의 구문구조와 의미론적인 정보, 그리고 품사정보를 갖는 말뭉치들을 제작했다. 이 프로젝트는 3차례에 걸쳐 <표 8>과 같이 전체 4개의 말뭉치들을 구축하였다.

<표 8> Penn Treebank의 구성

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

본 연구에서는 주로 Wall Street Jouranl과 The Brown Corpus를 주로 사용하였다. Wall Street Journal은 OntoNote에서 대명사 참조해소에 관련된 정보가 부착되어 학습 자료로서 매우 중요한 역할을 한다.

5.2 OntoNote 말뭉치

OntoNote 말뭉치는 OntoNote Project의 산출물이다. OntoNote Project는 BBN Technologies, 콜로라도 대학, 펜실베니아 대학 등이 협력하여 진행되었으며, 자연언어 처리 프로그램의 말뭉치를 제작하는 것을 목적으로 진행되었다. 이 말뭉치는 몇 가지 특징을 가지고 있는데, 그 첫 번째는 여러종류(뉴스기사, 전화대화, 웹블로그, 방송 등)의 말뭉치로 이루어져 있는 것이고 두 번째는 다양한 언어(영어, 중국어, 아랍어)로 이루어졌다는 것이다. 이 말뭉치들은 개체명에 대한 정보를 담고 있으며, 참조해소, WSD(Word Sence Disambiguation), 구문분석 등에 대한 정보도 담고 있다. OntoNote의 경우 말뭉치의 종류가 너무 다양하기 때문에 <표 9>를 통해 본 연구에 사용된 영어에 대한 말뭉치의 정보를 표시하였으며 앞에서 언급하였지만 주로 개체명 사전 구축과 대명사 참조해소의 학습 말뭉치로 사용되었다.

<표 9> OntoNote의 구성

말뭉치	내 용
ABC	American Broadcasting Company news
CNN	Cable News Network news
NBC	National Broadcasting Company news
VOA	Voice of America articles
WSJ	Wall Street Journal articles
MNB	MSNBC News
PRI	Public Radio International News

<표 10>은 OntoNote로부터 추출된 개체명 사전의 표제어 수이다. 전체 개체명 수는 10,227개이며 이 중에 인명이 4,141개, 지명이

2,192개, 기관명이 3,894개이다. 이 개체명 수는 말뭉치 내에서 중복되지 않는다.

<표 10> OntoNote에서 추출된 개체명 수

말뭉치	문장수	단어수	PER	LOC	ORG
ACE	1,234	21,936	147	123	46
CNN	6,173	103,795	656	384	370
MNB	662	12,024	88	36	34
NBC	652	13,450	92	99	55
PRI	2,161	47,569	307	217	176
VOA	2,158	48,430	402	293	230
WSJ	14,256	353,195	2,449	1,040	2,983
총계	27,296	600,399	4,141	2,192	3,894

5.3 MUC & ACE 말뭉치

MUC(Message Understanding Conferences) 말뭉치는 MUC-6과 MUC-7에 사용된 말뭉치들의 집합이다. 이 말뭉치 들은 Wall Street Journal Articles를 기본으로 하여 몇 가지 다른 뉴스들을 추가한 것으로 OntoNote와 비슷한 형태로 이루어져 있다. ACE 말뭉치는 TIDES Extraction(2003~2005)에 사용된 학습 데이터이다. 이 말뭉치들은 New York Times, Associated Press Worldstream 등의 기사와 CNN, ABC, VOA, PRI, MNB, NBC 등의 뉴스로 이루어져 있다. ACE 말뭉치는 개체명 정보를 포함하고 있으며, 각 개체명 사이의 관계정보 또한 포함하고 있다. 본 시스템에서는 앞에서 언급한 대로 MUC과 ACE에서는 개체명 사전 구축하는데 주로 사용되었으며 구축된 사전의 개체명 수는 <표 11>과 같다. 구축된

개체명의 총 수는 19,951개이다.

<표 11> MUC와 ACE에서 추출된 개체명 수

말뭉치	문장수	단어수	PER	LOC	ORG
MUC	43,039	369,680	2,888	1,157	2,759
ACE	NA	NA	7,579	976	4,592
총계	43,039	369,680	10,467	2,133	7,351

5.4 Wikipedia 말뭉치

위키피디아는 2001년 비영리 단체인 위키미디어 재단에 의해 만들어졌으며 현재까지 운영되고 있다. 위키피디아는 배타적인 저작권을 가지고 있지 않기 때문에 사용에 제약을 받지 않는다. 현재 300만여 개의 단어를 보유하고 있는데 현재 필요로 하는 개체명이 다수 포함되어 있으므로 위키피디아의 정보를 이용하였다.

위키피디아에서 개체명을 효율적으로 추출하기 위하여 인명, 지명, 기관명에 해당하는 카테고리를 분류하였다. 인명(person)에 관련되는 카테고리는 8,928개 지명(location) 4,470개, 기관명(organizaiotn)은 1,812개를 추출하였다. 이 카테고리 데이터를 이용하여 각각의 카테고리에 포함되는 단어를 1차적으로 분류하였다. 1차적으로 분류한 단어에 대하여 정확한 데이터를 만들기 위하여 2차적으로 위키피디아에서 각각의 단어에 대한 첫번째 문장을 확인하여 PLO에 대한 단어(인명 18,424개. 지명 16,010개, 기관명 14,110개)들을 분류하였다(<표 12> 참조).

<표 12> Wikipedia 말뭉치에서 추출된 개체명 수

말뭉치	PER	LOC	ORG
Wikipedia	18,424	16,010	14,111

5.5 BioLexicon

BioLexicon은 생의학 분야에서 특히 생물학 분야에 더 적합한 전문용어를 수집 및 통합한 용어집이라고 할 수 있다. 일반적으로 워드넷(WordNet)은 생물학 분야와 생물학 전문용어에 특화된 자원이 아니라 여러 분야를 아우르는 범용적으로 사용되는 어휘집이라 볼 수 있다. 그리고 다른 생의학 분야에 특화된 자원들은 생물학보다는 특히 의학 분야 쪽으로 치중된 경우가 많다. 그래서 보다 생물학 분야에 적당한 자원의 요구가 있어왔다. BioLexicon은 2006년부터 2009년까지 BOOTStrep 프로젝트의 결과로 생물학에 초점을 맞춰 추출된 용어집이다. 또한 용어의 수집뿐 아니라 용어의 관계(상하계층관계, 부분전체관계, 등), 용어의 이형태 정리까지 되어있는 용어집이다. BioLexicon은 홈페이지[9]에 공개되어 있다.

5.6 Gene Ontology

유전자 온톨로지(Gene Ontology)는 진핵생물의 유전자와 관련된 정보를 담고 있는데, 생물학적으로 다양한 목적에 대해 다루는 생물학적 과정(biological process) 온톨로지, 생화학 수준에서 유전자 생산물에 대하여 다루는 분자 기능(molecular function) 온톨로지, 유전자 생산물의 위치에 대하여 다루는 세포 요소 (cellular component) 온톨로지의 세 가지 세부 온톨로지로 구성되어 있다. 각 온톨로지는 'is-a'와 'part-of' 관계로 이루어진 방향성 비순환 그래프 (directed acyclic graphs) 형태를 갖게 된다. 현재 생물학적

9) http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html

과정, 분자 기능, 세포 요소 온톨로지는 각각 17,129개, 2,451개, 8,646개의 개념을 포함하고 있고, 총 28,226개의 개념을 포함한다. 유전자 온톨로지는 현재 계속 구축되고 있으며, 수작업으로 구축되고 있기 때문에 신뢰도가 높으나 구축 속도가 느린 단점이 있다. 아래 그림은 유전자 온톨로지의 예이다. 그림과 같이 유전자 온톨로지는 약 99%의 개념에 대해 정의(definition) 정보를 가지고 있다. 유전자 온톨로지는 홈페이지[10]에 공개되어 있다.

5.7 MEDLINE

MEDLINE(Medical Literature, Analysis, and Retrieval System Online)은 미국 국립의학도서관(U.S. National Library of Medicine)에서 제공하는 세계에서 가장 우수한 생의학 관련 데이터베이스로서 동물학, 영양학, 약물학, 의학, 수의학, 정신의학, 의료공학, 병리학 등에 대해 다루고 있다. MEDLINE은 Index Medicus, Index to Dental Literature, International Nursing Index의 전체 내용을 포함하고 있으며, 1966년 이후의 전 세계 5,000여 연속간행물 기사에 대한 색인과 초록을 제공한다. 현재 하루에 약 18,000,000개의 문서가 있으며, 한 달에 약 40,000개의 문서가 추가되고 있다. MEDLINE은 홈페이지[11]에 공개되어 있다

5.8 GENIA 말뭉치

GENIA 말뭉치는 MEDLINE 데이터베이스를 대상으로 "human

10) http://www.geneontology.org
11) http://www.nlm.nih.gov

AND blood cell AND transcription factor"라는 질의를 이용해 검색한 5,000 여 개의 요약문들의 일부를 대상으로 생의학 전문가들이 개체명 태그를 부여한 말뭉치이다. 이 말뭉치는 현재 (GENIA corpus version 3.02) <표 13>과 같이 37개의 세부 분야를 사용하고 있으며, 2,000여 개의 요약문을 대상으로 약 93,000 여 개의 전문용어가 태깅되어 있다. GENIA 말뭉치는 홈페이지[12]에 공개되어 있다.

<표 13> GENIA 말뭉치의 분야 정보

GENIA 말뭉치 분야 정보	
amino_acid_monomer peptide protein_N/A protein_complex protein_domain_or_region protein_family_or_group protein_molecule protein_substructure protein_subunit	nucleotide polynucleotide DNA_N/A, DNA_domain_or_region DNA_family_or_group, DNA_molecule DNA_substructure RNA_N/A, RNA_domain_or_region RNA_family_or_group, RNA_molecule RNA_substructure
other_organic_compound organic inorganic atom carbohydrate lipid	virus mono_cell, multi_cell body_part tissue cell_type, cell_component, cell_line other_artificial_source
other_name	coordinated

12) http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

6. 언어처리엔진

개체명과 전문용어에 대한 자동인식 그리고 대용어에 대한 선행어 자동탐지를 위해 기계학습모델을 이용하였다. 기계학습모델을 구현하기 위해서는 필요한 학습 자질을 추출하기 위해 다양한 언어적인 분석이 필요하고, 이를 위해 문장분리, 토큰분리, 품사부착, 기저구 인식 등의 언어처리 모듈을 개발하였다. 기존에 공개된 모듈들은 서로 다른 컴퓨터 언어들을 사용하고, 입력과 출력의 형식이 서로 상이하여 전체 시스템의 효율적인 통합과 성능을 위해 직접 개발하였다.

언어처리 단계는 문장분리(sentence segmentation), 단어분리(tokenization), 품사 부착(POS tagging), 기저구 인식(chunking)으로 구성된다. 문장 분리는 입력된 문서를 문서 단위로 분리한다. 토큰 분리는 의미있는 기본 단위로 토큰을 사용하는데 형태소 분석보다는 조금 더 큰 단위이고 공백 단위보다는 작은 단위이다. 예를 들어, 문장 부호의 분리가 그 대표적이다. 토큰 분리도 단순한 문제가 아니다. "Mr."에서와 같이 "."은 문장 부호가 아니므로 분리되어서는 안된다. 이처럼 동일한 기호라도 주변 문맥에 따라 서로 다른 결과를 생성할 수 있다.

품사 부착은 토큰에 주어진 문장에 가장 적합한 품사를 할당하는 작업이다. 토큰에 품사를 부착하기 위해서는 토큰 주변의 문맥이 중요한 역할을 담당한다. 기저구 인식은 주로 명사구를 인식하는 작업이다. 영문에서 명사구는 중심어를 기준으로 앞이나 뒤에서 수식어구(형용사구, 또 다른 명사구, 전치사 구 등)로부터 수식을 받는 형태를 말한다. 어떤 명사구의 범위를 정확하게 결정하는 문제는 완전 구문분석(full-parsing) 문제로 볼 수 있다. 또한 명사구

의 내부 구조를 결정할 경우에 중심어들의 수식관계를 결정하기 위하여 많은 어휘 지식이 요구된다. 그러나 정보검색이나 정보추출 분야에서 복잡한 언어적인 지식을 이용하면, 실시간 처리 뿐만 아니라 강인한 시스템을 구현하기가 어렵기 때문에, 비교적 간단한 구문지식을 이용해서 명사구를 인식할 수 있는 기저명사구를 인식한다.

언어처리엔진의 기계학습모델을 학습 및 테스트하기 위해 <표 8>의 말뭉치를 사용하였다. Penn Treebank는 자연어에 대한 언어학적 구조를 분석하기 위한 목적을 위해 진행된 프로젝트의 산출물로 문장의 구문구조와 의미론적 정보, 품사정보를 담고 있다.
문장분리기에서는 InitCap(첫글자가 대문자), AllCap(모두 대문자), CapPeriod(대문자 뒤에 종결문자), LowerCase(모두 소문자), ConatinDigit(숫자 포함), Prev_CC(이전 2문자), Next_CC(이후 2문자) 7개의 자질을 사용하였고, 현재단어, 다음단어에 이 자질들을 적용하였다. Penn Treebank의 Brown 말뭉치와 WSJ 말뭉치가 문장분리 학습을 위해 사용되었고, 학습과 테스트를 9:1 비율(857,050 단어 / 92,864 단어)로 나누어 사용하였다. 성능평가 결과 99.76%의 문장 분리 정확률을 보여주었다. 약어나 인명의 머리글자(“B.", "C" 등)에서 일부 오류가 발생하였다.

토큰분리기에서는 Char(현재 자신의 문자), Pre1Next1Cap(앞뒤가 대문자), SpecialChar(특수기호), Prev_CCO(자신포함 이전 3문자), Next_OCC(자신포함 이후 3문자) 자질을 사용하고, 현재단어에만 자질들을 적용하였다. 토큰 분리기의 학습을 위해 Penn Treebank의 Brown 말뭉치를 9:1의 비율(4,976,447 문자 / 550,762 문자)로 분리하여 사용하였다. 성능평가 결과 모든 문자 중 토큰 분리를 정확하게 인식하는 비율이 99.98%로 오류가 거의 발생하지 않

았다.

품사부착기는 Word(단어), Suffix3(접미사 3문자), Prefix2(접두사 2문자)를 자질로 취하였고 이전 2개 단어와 현재 그리고 다음 2개 단어에 자질들을 적용하였다. 품사부착을 위한 말뭉치도 Penn Treebank의 Brown 말뭉치를 사용하였고, 9:1(학습:991,410 토큰 / 평가 : 108,800 토큰)로 나누었다. 모든 단어 중에서 품사를 정확하게 부착한 비율은 96.2% 였다.

기저구인식기는 품사부착기에서 사용한 자질외에 POS(품사정보), InitCap(첫글자가 대문자) 자질을 선택하여 이전 2개 단어, 현재, 다음 2개단어와 이전 2개 품사, 현재품사, 다음 2개품사에 자질들을 적용하였다. 기저구 인식 또한 Penn Treebank 의 Brown 말뭉치를 9:1의 비율(학습:412,584 토큰 / 평가 : 45,867 토큰)로 분리하여 사용하였다. 성능평가는 재현율과 정확률 두 가지를 측정하였다. 기저구 인식의 재현율은 96.57%, 정확률은 96.85%로 나타났다.

<표 14> 언어처리엔진 정확률

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

실험을 위해 말뭉치의 1/10을 학습을 위해 사용하였고, 9/10을 테스트를 위해 사용하였다. <표 14>는 본 연구에서 개발한 언어처리엔진에서 사용한 말뭉치 종류와 정확률을 보여준다. 4개의 모듈 모두 높은 수준의 정확률을 보여주었다.

7. 개체명 인식기

MUC-6와 CoNLL 2002, 2003에서 제안된 개체명인식 시스템들은 2장에서 언급한 BIO 태그를 이용한 기계학습방법을 이용하였다. 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 15>는 선택된 자질 집합을 보여준다.

<표 15> 개체명 인식의 자질 집합

자질이름	내 용	적 용 범 위
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 있는지	

<표 16> 개체명 자질 값의 부호화

문 자	부호와의 의미
B	기본 개체명 사전에 첫 단어로 출현
I	기본 개체명 사전에 첫 단어가 아닌 단어로 출현

P	기본 개체명 사전에 인명으로 출현
L	기본 개체명 사전에 지명으로 출현
O	기본 개체명 사전에 기관명으로 출현

<표 15>에서 보는바와 같이 사전(Dic)자질을 제외하고 다른 자질들은 6장에서 설명한 언어처리엔진에서 사용하는 자질들과 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 <표 16>에서와 같이 “BIPLO″ 템플릿을 제작하여 이를 해결하였다. ”BI″는 개체명의 처음과 중간을 나타내고, “PLO″는 인명, 지명, 기관명을 구분하는 태그이다.

일반적으로 기계학습기반 엔진들은 학습데이터가 충분할 경우 좋은 성능을 보여준다. 개체명인식 성능 향상을 위해 <표 17>에서 보는 바와 같이, OntoNote 말뭉치, MUC 말뭉치 등 다양한 언어자원을 사용하여 개체명 사전을 구축하였다..

<표 17>에서 사용한 외부 말뭉치들은 개체명정보를 포함하지만, 본 연구의 개체명인식기에서 사용한 자질들을 위한 문장분리 정보, 토큰분리 정보, 품사 정보가 결여되어 있다. 개체명인식기의 성능을 높이기 위해, 6장에서 소개한 언어처리모듈을 이용하여 문장분리 정보, 품사 정보 등을 외부 말뭉치에 추가하여 본 연구에서 사용할 수 있도록 확장하였다.

<표 17> 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

<표 18> 실험에 사용된 말뭉치의 종류

말뭉치이름	의 미	크기(단어)	비 고
말뭉치 1	OntoNote WSJ	292,823	Penn Treebank 사용
말뭉치 2	OntoNote 전체	458,451	언어처리 엔진 사용
말뭉치 3	OntoNote 전체, MUC-6&7	822,344	언어처리 엔진 사용

개체명 시스템은 두 가지 방법으로 구현되었다. 하나는 개체명의 경계를 인식한 후에 개체명의 종류를 분류하는 경계분류 순차 시스템이고 다른 하나는 경계인식과 분류를 동시에 수행하는 경계분류 동시 시스템이다. <표 19>는 <표 18>의 말뭉치를 각 시스템에 적용하여 성능을 평가한 결과이다.

<표 19> 개체명 인식의 성능 평가

적용 말뭉치	F1 척도	
	경계분류 순차 시스템	경계분류 동시 시스템
말뭉치 1	84.00	89.34
말뭉치 2	90.43	92.09

말뭉치 3	98.66	99.21

<표 19>에서 보는바와 같이 말뭉치의 규모가 확장함에 따라, 경계분류 순차시스템과 경계분류 동시 시스템 모두 더 좋은 결과를 나타내고, 경계인식과 개체명분류를 순차적으로 하는 경우보다 동시에 수행하는 경계분류 동시 시스템의 성능이 우월함을 알 수 있다. 본 연구에서는 이 결과를 바탕으로 통합시스템을 위한 개체명 인식기를 경계분류 동시 시스템으로 적용하였다.

8. 전문용어 인식기

전문용어는 전문적 개념을 지칭하는 어휘 또는 어휘의 집합을 말한다. 이러한 전문용어는 각 분야의 전문문서에 사용되지만, 문서 내의 단어 또는 어휘가 전문용어인지 아닌지를 선별하는 작업은 쉬운 일이 아니다. 본 연구에서는 생의학 분야의 전문용어 자질을 바탕으로 CRF 기계학습모델을 이용한 전문용어 인식 방법을 구현한다.

전문용어는 핵심개체 중 분야특화명과 기술명을 의미하며, Gene(유전자), Protein(단백질), Oranism(유기체), Disease(질병), Drug(약명), TechTerm(기술용어), Others(기타) 7개의 범주로 분류된다.

대상 데이터를 생의학분야로 선정하였기 때문에, 전문용어 인식기의 성능을 높이기 위하여 기 구축되어 있는 생의학 분야 말뭉치인 BioLexicon, Gene Ontology, GENIA를 이용하였다.

사용된 말뭉치들은 본 연구에서 선정한 7개의 범주와 일치하는 부분도 있지만 세분화된 범주나 전혀 관계없는 범주 또한 존재하기 때문에, 본 연구에서 제안한 핵심개체의 범주로 변환하는 작업을 선행하였다. <표 20>은 BioLexicon 말뭉치 범주와 대응되는 범주를 보여준다.

<표 20> BioLexicon 범주 및 대응 범주

BioLexicon 범주	entry 개수	variant 개수	본 연구 제안 범주
cell	842	1,400	Organism
chemicals	19,637	106,302	Others
enzymes	4,016	11,674	Protein

diseases	19,457	33,161	Disease
genes and proteins	1,640,608	3,048,920	대상 클래스 없음
molecular role concepts	8,850	60,408	대상 클래스 없음
operons	2,672	3,145	Gene
protein complexes	2,104	2,647	Protein
protein domains	16,940	33,880	Protein
species	482,992	669,481	대상 클래스 없음
transcription factors	160	795	대상 클래스 없음

전문용어 인식을 위한 기계학습모델의 자질들은 <표 21>에서 보는바와 같이 10가지가 선정되었다. 통계적으로 볼 때 생의학분야 전문용어 중 소문자로만 구성된 용어가 절반 이상이지만, 특정질병, 단백질, 약명, 유전자의 이름에 해당하는 전문용어는 대문자, 숫자, 특수기호 등을 포함하는 경우가 많다(예: 5HT2a, BHC110/LSD1, CaMKII).

이러한 생의학분야 전문용어의 특성을 반영하기 위해, 대소문자가 섞여있는지 여부를 묻는 자질(Mixed)과 숫자, 특수기호의 포함여부를 묻는 자질(ContainDigit, ContainSpChar)을 사용하였다. 알파벳 20자 이상의 긴 용어들의 경우 약어로 표현되는 경향을 반영하기 위해, 모두 대문자인지 여부를 묻는 자질(AllCap)을 사용하였다.

<표 21> 전문용어 인식을 위한 자질 집합과 자질 값

자질	수	비고	표현 형식
POS	36	품사정보	NN, NNS, NP, NPS, JJ,..
AllCap	2	모든 문자가 대문자	+, -
Mixed	2	대소문자 혼용 여부	+, -
ContainDigit	2	숫자 포함 여부	+, -

CotainSpChar	2	특수기호 포함 여부	+, -
Prefix2	-	접두사 2문자	Di, Al, Ch, pn, sy, an ...
Suffix3	-	접미사 3문자	dia, ity, ion, mia, ncy, ...
FirstTerm	2	전문용어 시작단어 단서어휘 여부	+, -
MiddelTerm	2	전문용어 중간단어 단서어휘 여부	+, -
EndTerm	7	전문용어 종료단어 단서어휘 여부	Disease, Theraphy, System ..

<표 22> 질병명과 약명에 관련된 접두사와 접미사

접사	의미	예
-algia	pain	talalgia *ankle*
-cele	hernia	gastro*cele stomach*
-dynia	pain, swelling	uro*dynia urine*
-gen	producing, beginning	carcino*gen cancer*
-oma	tumour	adeno*ma gland*
-osis	abnormal condition	dermatosis skin
sulfa-	antibiotic	sulfacetamide
cef-, ceph-	cephalosporin antibiotic	cefaclor; cephalexin

생의학분야 전문용어의 또 다른 특징은 중 하나는 그리스 라틴 어원을 사용한다는 점이다. 이러한 용어들은 접두사, 접미사가 동일한 형태소로 구성되며 특별한 의미를 갖는다. <표 22>는 특정 접두사, 접미사가 가지는 의미와 예를 보여준다.

접사의 길이는 유동적이므로, 특징만 뽑아낼 수 있도록 접미사는 3개의 문자, 접두사는 2개의 문자를 자질(Prefix2, Suffix3)로 사용하였다.

2 어절 이상의 복합어로 구성된 전문용어에 대해서 전문용어 사

전을 활용하였다. 각 분야의 전문용어 사전마다 어절별로 나누어서 가장 앞에 나오는 어절(F단서어휘), 중간에 나오는 어절(M단서어휘), 마지막에 나오는 어절(E단서어휘)로 나누어 빈도별로 정리한 것이 단서어휘이다(<표 23> 참조).

<표 23> 전문용어 단서어휘

단어	F 단서어휘	M 단서어휘	E 단서어휘	총계
therapy	0	1	63	66
treatment	1	2	22	27
surgery	2	1	13	18
technique	0	0	11	12

단서어휘는 전문용어 인식 시스템에서 중요한 역할을 담당 한다. 특히 복합어인 전문용어 중 마지막 단어와 관련된 단서어휘는 경계인식 외에, 분야 결정을 위한 중요한 자질이 되므로, 단서어휘 사전에 포함되어 있는지에 대한 여부뿐 아니라, 분야정보도 자질값에 포함하였다.

단서어휘의 형태는 대소문자, 숫자, 특수기호, 약어, 복수형태 등의 유무와 차이로 인해 여러 가지 이형태들이 나타난다.

예) (Hodgkin Disease), (HODGKIN DISEASE),
(Hodgkin's Disease),
(Hodgkin's disease), (Disease, Hodgkin)

이러한 요소들은 전문용어 인식 시스템의 재현율을 떨어뜨리는 원인이 되므로, 본 연구에서는 전체가 대문자, 숫자, 기호인 단어를 제외하고 나머지 단어에 대하여 일반화 작업을 수행하였다. 위에서

제외한 단어는 약어로 볼 수 있는 단어이며, 일반화 작업으로 인해 다른 단어와 중의성이 발생할 수 있기 때문에 제외하였다. 일반화 작업은 모든 단어를 소문자로 변환하고 복수형 단어를 단수형으로 교체하는 작업이다.

전문용어 인식기도 개체명인식기와 동일하게 전문용어에 대한 경계인식과 인식된 전문용어에 대하여 분류를 수행하는 두 가지 시스템으로 구현되었다. 성능평가를 위해서 한국과학기술정보연구원에서 구축한 테스트컬렉션(KEEC 2009[13])를 이용하였다. 성능평가는 ㉮외부 사전을 이용한 경우와 ㉯단서어휘를 일반화한 경우를 조합하여 (1) ㉮, ㉯ 모두 사용하지 않은 경우, (2) ㉯만 사용한 경우, (3) ㉮, ㉯ 모두 사용한 경우에 대하여 각각 수행하였다.

<표 24> 전문용어 경계인식의 성능 평가

경계	정확률			재현율			F1-척도		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
B, I	**87.23**	75.94	83.92	67.61	84.59	**96.42**	76.17	80.03	**89.73**

<표 24>는 전문용어 경계인식의 실험결과이다. 정확률은 (1)의 경우가, 재현율과 F1-척도는 (3)의 경우가 가장 좋은 결과를 보였다. (1)의 정확률은 87.23%로 가장 높은 반면 재현율이 다른 경우와 비교하여 현저하게 낮음을 보였다. (1)의 경우 외부사전과 단서어휘를 사용하지 않았기 때문에 전문용어 인식률은 낮지만, 인식된 전문용어에 대한 경계는 정확히 인식했음을 의미한다. 사전과 단서어휘를 함께 사용한 (3)의 경우 정확률, 재현율 모두에서 전반적으로

13) KEEC 2009(KISTI Entity Extraction Collection 2009) : 한국과학기술정보연구원이 보유하고 있는 해외학술지 중 생의학관련분야 문헌을 선정하여 테스트컬렉션을 구축하였으며, 핵심개체 21,142를 포함하는 8,303 문장으로 구성되어 있다.

높은 성능을 보여준다.

<표 25>는 전문용어 분야인식의 실험결과이다. 경계인식의 경우와 마찬가지로 외부사전과 단서어휘를 사용한 (3)의 경우에 가장 좋은 실험결과를 보여주었다. 사전이나 단서어휘를 사용하지 않은 (1), (2)의 경우에는 실험수치가 상당히 낮은 경우를 볼 수 있다. 정확률 측정 (2)의 Gene39.24%, 재현율 측정 (1)의 Gene은 22.92%로 매우 낮았다.

전문용어의 경계인식과 분야인식의 두 경우 모두, 외부 사전과 단서어휘를 이용한 경우에 가장 좋은 F-1 척도를 보여준다. 본 연구의 결과는 2.2에서 소개한 생의학 분야 전문용어 인식에 대한 세계 최고 수준에 근접(F1-척도 75%~85%사이)하는 실험결과를 보였다.

<표 25> 전문용어 분야분류의 성능 평가

분류	정확률			재현율			F1-척도		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Disease	87.78	82.67	82.49	74.19	79.50	95.87	80.42	81.05	88.68
Drug	60.53	51.22	90.48	25.00	34.24	92.94	35.36	41.04	91.69
Gene	84.62	39.24	71.88	22.92	64.58	95.83	36.07	48.82	82.14
Organism	84.13	83.59	73.91	74.65	77.70	95.07	79.11	80.54	83.16
Protein	64.71	47.85	86.89	49.62	75.19	79.70	69.35	58.48	83.14
TechTerm	69.35	57.91	82.06	65.71	87.14	87.14	67.48	69.58	84.53
Others	64.73	55.94	75.69	46.50	70.19	86.29	54.12	62.26	80.64
전 체	76.47	65.65	**79.29**	59.32	73.11	**91.06**	66.81	69.18	**84.77**

9. 대용어 참조해소기

본 연구에서 제안된 시스템은 크게 두 부분으로 구성된다. 하나는 대용어로서 대명사를 인식하는 시스템이고 다른 하나는 인식된 대명사의 선행어를 찾는 시스템이다. 전자를 대용어 인식 시스템(anaphoric pronoun identifier)이라고 하고 후자를 선행어 결정 시스템(pronoun resolver)이라고 한다. 품사가 대명사(PRP)라고 해서 모두 대용어는 아니다. 즉 문서 내에 어떤 선행어를 가리키지 않는 대명사가 존재한다는 것이다. 예를 들면 문장 “It is important not to give up”에서 ‘It’은 선행어를 가지지 않는 대명사이다. 대용어 인식 시스템은 품사가 대명사(PRP)인 단어 중에서 대용어를 찾아내는 시스템이다.

<표 26> 대용어 인식 시스템의 자질 집합

종 류	자 질 집 합
어 휘	w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, $w_{i-1}w_i$, w_iw_{i+1}
품 사	t_{i-2}, t_{i-1}, t_i, t_{i+1}, t_{i+2}, $t_{i-1}t_i$, t_it_{i+1} $t_{i-2}t_{i-1}t_i$, $t_{i-1}t_it_{i+1}$, $t_it_{i+1}t_{i+2}$

대용어 인식 시스템에 의해서 인식된 대용대명사(anaphoric pronoun)를 대상으로 선행어를 결정하는 시스템이 선행어 결정 시스템이다. 아래의 예문을 보면 2개의 대용대명사 ‘them’과 ‘it’이 있으며 각각의 선행어는 ‘Those figure’와 ‘the government’이다.

<E id=1 Those figures> are almost exactly what the government proposed to legislators in September. If **<E id=2 the**

government> can stick with **<E id =1 them>**, **<E id=1 it>** will be able to halve this year's 120 billion ruble (US $193 billion) deficit.

<표 27> 선행어 결정 시스템의 자질 집합

종 류	자 질
대명사	- 인칭대명사인지 아닌지 - 소유격인지 아닌지 - 3인칭인지 아닌지 - 재귀대명사인지 아닌지 - 대문자를 포함했는지 아닌지 - 앞/뒤 단어의 품사 - 앞뒤 단어를 포함한 품사열(5개)
후 보	- 구성 단어의 수 - 대명사인지 아닌지 - Is indefinite NP - Is Demonstrative NP - 앞/뒤 단어의 품사 - 자신의 품사 - 이전의 텍스트에서 자신이 나온 횟수
관 계	- 대명사와 후보 간의 문장 거리 - 대명사와 후보 간의 단어 거리

대용어 참조를 해소를 위한 자질은 대용대명사 자신과 주변의 어휘 정보들과 품사 정보를 사용한다. 여러번의 실험을 통해서 <표 26>과 같은 자질 집합을 선택하였다. <표 26>에서 w_i는 대용대명사이고 t_i는 대용대명사의 품사이다. 색인이 음수이면 w_i의 이전 단어이고 양수이면 다음 단어이다.

선행어 결정 시스템에 사용되는 자질 집합은 대용 대명사에 대한 자질, 선행어에 대한 자질, 대용 대명사와 선행어 사이의 관계 자질 세 가지로 분류한다. 각각에 대해 품사나 격, 위치 정보 등을

조합하여 자질들을 생성하게 된다. 하지만 자질들의 수가 너무 많고, 기계학습의 학습 속도 문제로 인해 이러한 자질의 모든 조합에 대해 실험하는 것은 사실상 불가능하므로, 기존 연구들을 참고하여 사용할 수 있는 자질들의 목록을 작성하고, 결정 트리에서 사용되는 정보 이득(information gain)을 이용하여 선행어 결정 시스템을 위한 자질 집합을 <표 27>과 같이 결정하였다.

본 연구에서 선행어 결정을 위해 TCM(twin-candidate model)을 사용한다. TCM은 대용대명사와 각각의 후보 선행어들 사이의 관계에 의해서 최종 선행어를 결정하는 것이 아니라 두 후보 선행어들끼리 경쟁할 수 있도록 모델링된 것이다. TCM에 의해서 모델이 되면 최종적으로 선행어를 결정하기 위해서 후보들끼리 경쟁을 하게 되는데 경쟁 방법에는 크게 두 가지 방법이 존재한다.

첫 번째는 승자진출전(tournament) 방법으로 모든 후보들 중 가장 먼저 2개의 후보를 선택하고 경쟁을 붙이게 된다. 이 경쟁에서 살아남은 후보와 아직 선택되지 않은 후보 중 1개를 선택하여 다시 경쟁하는 형태이다. 모든 후보들이 선택되어 남아있는 후보가 없어지면 경쟁은 끝이 나고 최후에 살아남아 있는 후보가 정답으로 결정되는 시스템이다.

두 번째는 연맹전(league) 방법으로 모든 후보들에 대해 경쟁을 붙여서 가장 승률이 좋은 후보를 선택하는 시스템이다. 경쟁에 대한 정답 태그는 총 3가지로 “00”, “10”, “01” 이 그것이다. “00”은 둘 다 정답이 아닌 경우이고 “10”은 앞의 것이, “01”은 뒤에 것이 경쟁에서 살아남은 것을 표시하는 태그이다. <표 28>은 승사진출전 방식의 실제 경쟁하는 예를, <표 29>는 연맹전 방식의 실제 예를 보여준다.

<표 28> 승자진출전을 통한 선행어 결정

대명사	후보들	정답
[6 them]	[1 Those figures], [2 the government]	10
	[1 Those figures], [3 legislators]	10
	[1 Those figures], [4 September]	10
	[1 Those figures], **[5 the government]**	10
[7 it]	[1 Those figures], [2 the government]	01
	[2 the government], [3 legislators]	10
	[2 the government], [4 September]	10
	[2 the government], [5 the government]	01
	[5 the government], [6 them]	10

대용어 참조 해소 시스템에서는 대용대명사가 가리키는 선행어를 정확하게 찾는 비율을 성능 척도로 사용하며, 선행어 결정 정확률 A_a이라고 한다. 여기서 n_a는 전체 문서에 인식된 올바른 대용대명사의 수이고 n_a^c는 대용대명사들 중에서 선행어를 정확하게 찾아낸 대용대명사의 수이다.

$$A_a = \frac{n_a^c}{n_a} \times 100$$

<표 30>은 연맹전과 승자 진출전에 대한 성능 평가 결과이다. 모두 대용대명사 결정 시스템을 통해 결정된 대용대명사를 대상으로 선행어 결정 시스템을 적용한 결과로써 대용대명사 결정 시스템의 에러가 그대로 전파된 결과이다.

<표 29> 연맹전을 통한 선행어 결정

대명사	후보들	정답
[7 it]	[1 Those figures], [2 the government]	01
	[1 Those figures], [3 legislators]	00
	[1 Those figures], [4 September]	00
	[1 Those figures], [5 the government]	01
	[1 Those figures], [6 them]	00

	[2 the government], [3 legislators]	10
	[2 the government], [4 September]	10
	[2 the government], [5 the government]	01
	[2 the government], [6 them]	10
	[3 legislators], [4 September]	00
	[3 legislators], [5 the government]	01
	[3 legislators], [6 them]	00
	[4 September], [5 the government]	01
	[4 September], [6 them]	00
	[5 the government], [6 them]	10

선행어	점수
[1 Those figures]	0
[2 the government]	4
[3 legislators]	0
[4 September]	0
[5 the government]	5
[6 them]	0

<표 30> 대용어 참조해소 시스템의 정확률

구분	대용대명사	대용대명사 시스템 결과	선행어를 정확하게 인식한 대용대명사의 수	정확률 (%)
승자진출전	676	610	386	57.1
연맹전	676	610	254	37.5

<표 30>을 살펴보면 승자진출전에 비해서 연맹전이 상당히 낮은 정확률을 보이는 것을 알 수 있다. 이는 연맹전의 특성상 모든 후보들 간의 경쟁이 일어남에 따라 정답 태그 중 “00”의 비율이 너무 높아져서 학습 결과가 ″00″쪽으로 치우치는 현상 때문이다.

<표 31> 선행어 결정 시스템의 정확률

대용대명사의 수	선행어를 정확하게 인식한 대용대명사의 수	정확률 (%)
676	526	83.13

<표 31>은 대용대명사 결정 시스템의 결과를 이용하지 않고 대용대명사 만을 대상으로 선행어 결정 시스템을 적용시킨 결과이다. <표 31>을 살펴보면 대용대명사 시스템에서 전파되는 에러율이 상당히 높은 것을 알 수 있다(정확률 83.13%). 이러한 이유로 대용대명사 결정 시스템을 사용하지 않고, 모든 대명사를 대상으로 선행어를 결정하되, 대용대명사가 아닐 경우는 최종 정답 태그가 “00” 되도록 선행어 결정 시스템을 구성하였다. 최종 시스템의 성능은 <표 32>와 같다. <표 30>에서 살펴본 대용대명사 결정 시스템을 거친 선행어 결정 시스템 보다 약 6%정도 높은 정확률을 보인다. 본 연구에서는 최종적으로 대용 대명사 결정 시스템을 거치지 않는 승자 진출전 방식을 선택하였다.

<표 32> 대용대명사 결정 시스템을 거치지 않은 시스템 정확률

구분	대용대명사의 수	선행어를 정확하게 인식한 대용대명사의 수	정확률 (%)
승자진출전	676	426	63.01

10. 결론

본 연구에서는 독립적으로 개발되어오던 개체명인식기술과 전문용어인식기술을 하나의 작업으로 통합하여, 과학기술 문헌에 포함된 핵심개체인 개체명(named entity)과 전문용어(terminology)를 자동으로 인식하고 이들을 가리키는 대용어를 결정하기 위해 테스트컬렉션을 구축하고 관련 시스템을 개발하였다.

- ❖ 일반적인 개체명과 분야 종속적인 전문용어들에 대한 분류를 핵심개체로 정의. 일반 개체명은 인명(Person), 지명(Location), 기관명(Organization) 3개 분류로 지정. 전문용어는 질병명(Disease), 약명(Drug), 유전자명(Gene), 생물조직명(Organism), 단백질명(Protein), 기술명(TechTerm), 기타(Others) 7개 분류로 지정
- ❖ 대상 데이터를 기반으로 과학기술 핵심개체 인식기술을 위한 테스트컬력션 구축. 테스트컬렉션 구축 작업 지원을 위한 구축 지원 도구 개발. 전체 2만여개의 핵심개체를 포함하는 7천여 문장의 테스트컬렉션 구축
- ❖ 개체명인식, 전문용어인식, 대용어 참조해소 기술들의 하부 엔진으로 문장분리기, 토큰분리기, 품사부착기, 기저구 인식기 등의 하부처리엔진 개발
- ❖ 개체명에 대한 분류를 순차적으로 수행하는 경계분류 순차 시스템과 동시에 수행하는 경계분류 동시 시스템을 구현하여 성능을 비교
- ❖ 생의학분야의 전문용어 특성들을 파악하고, 외부 말뭉치들을 가공하여 전문용어 인식기 개발

❖ 개체명과 전문용어가 인식된 문서에서 대용어가 참조하는 핵심개체를 해소하는 대용어 참조해소기 개발

❖ 하부처리 엔진, 개체명 인식기, 전문용어 인식기, 대용어 참조해소기에 대한 성능 평가를 수행

[참고문헌]

[1] Ananiadou, S. and Nenadic, G. (2006) “Automatic terminology management in biomedicine”, Text Mining for Biology and Biomedicine, pp. 67-97.

[2] Black, W. J. and Vasilakopoulos, A. (2002) “Language-Independent Named Entity Classification by Modified Transformation-Based Learning and by Decision Tree Induction”, Proceedings of CoNLL-2002, Taipei, Taiwan, pp. 159-162.

[3] Carbonell, J., Brown, R. (1988) "Anaphora resolution: a multi-strategy approach", Proceedings of COLING, pp. 96-101.

[4] Carreras, X., Màrques, L. and Padró, L. (2002) “Named Entity Extraction using AdaBoost”, Proceedings of CoNLL-2002, Taipei, Taiwan, pp. 167-170.

[5] Chang, J. T., Schutze, H. and Altman, R. B. (2004), ‘'GAPSCORE: Finding gene and protein names one word at a time’', Bioinformatics, Vol. 20(2), pp. 216 - 225.

[6] Collins, M. (2002) “Ranking Algorithms for Named-Entity Extraction: Boosting and the Voted Perceptron”, Proceedings of ACL 2002, University of Pennsylvania, PA.

[7] Denis, P., Baldridge, J.(2009) "A ranking approach to pronoun resolution", Proceedings of IJCAI-07, pp. 1588-1593.

[8] Elsner, M. and Charniak, E. (2007) A Generative Discourse-New Model for Text Coherence, Tech Report CS-07-04, Department of Computer Science, Brown University.

[9] Grosz, B.J., Joshi, A.K. and Weinstein, S. (1995) Centering: A Framework for Modeling the Local Coherence of Discourse, Computational Linguistics , vol. 12, no. 2, pp.203-225.

[10] Kilicaslan, Y., Guner, E., Yildirim, S.(2009) "Learning-based

pronoun resolution for Turkish with a comparative evaluation", Computer Speech Language, vol. 23, no. 3. pp. 311-331.
[11] Lafferty, J., McCallum, A. and Pereira, F. (2001) "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", Proceedings of International Conference on Machine Learning, pp. 282-289.
[12] Lappin, S., Leass, H. (1994) "An algorithm for pronominal anaphora resolution", Computational Linguistics, vol. 20. no.4. pp. 535-561.
[13] LDC (2008), ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, ver 6.6, Linguistic Data Consortium.
[14] Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. (2004) "Building a large annotated corpus of English: The Penn Treebank", Computational Linguistics, vol. 19, no.2, pp. 313-330.
[15] Nguyen, N., Kim, J., Tsujii, J.(2008) "Challenges in pronoun resolution system for biomedical text", Proceedings of LREC'08, pp. 2408-2412.
[16] Sasaki, Y., Montemagni, S., Pezik, P., Rebholz-Schuhmann, D., McNaught, J. and Ananiadou, S. (2008) "BioLexicon: A lexical resource for the biology domain", Proceedings of the Third International Symposium on Semantic Mining in Biomedicine.
[17] Soon, W., Ng, H., Lim, D.(2003) "A machine learning approach to coreference resolution of noun phrases", Computational Linguistics. vol. 27. no. 4, pp. 521-544.
[18] Tanabe, L. and Wilbur, W. J. (2002), ''Tagging gene and protein names in biomedical text'', Bioinformatics, Vol. 18(8), pp. 1124 - 1132.

[19] The Gene Ontology Consortium (2008), “The Gene Ontology project in 2008”, Nucleic Acids Research vol. 36 pp. D440 - 444.
[20] Tjong Kim Sang, E. F. (2002) “Memory-Based Shallow Parsing”, Journal of Machine Learning Research, volume 2 (March), pp. 559-594.
[21] Watanabe, Y., Asahara, M. and Matsumoto, Y. (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, pp. 649 - 657, Prague.
[22] Weischedel, R., Pradhan, S. Ramshaw, L., Khleif, T., Palmer, M., Xue, N., Marcus, M. Taylor, A. Greenberg, C., Hovy, E., Belvin, R. and Hoston, A. (2007) OntoNotes Release 2.0, BBN Technologies.
[23] Yang, X., Su, J. and Tan, C. L. (2008) “A twin-candidate model for learning-based anaphora resolution”, Computational Linguistics, vol. 34, no. 3, pp. 327-356.
[24] Zhang, J., Shen, D., Zhou, G., Su, J. and Tan, C.-L. (2004) “Enhancing HMM-based biomedical named entity recognition by studying special phenomena”, J. Biomed. Inform., vol. 37, pp. 411-422.
[25] Zhou, G., Zhang, J., Su, J. et al. (2004), "Recognizing names in biomedical texts: A machine learning approach", Bioinformatics, Vol. 20(7), pp. 1178-1190.
[26] 국립국어원 (2007), 전문용어 연구, 태학사.
[27] 오종훈, 최기선 (2006) 기계학습에 기반한 생의학분야 전문용어의 자동인식, 정보과학회논문지: 소프트웨어 및 응용, 제33권, 제8호, pp. 718-729.