http://dx.doi.org/10.5392/JKCA.2012.12.07.463

기술 지식 자동 추출을 위한 테스트 컬렉션 구축

Construction of Test Collection for Automatically Extracting Technological Knowledge

신성호, 최윤수, 송사광, 최성필, 정한민
한국과학기술정보연구원 소프트웨어연구실

Sungho Shin(maximus74@kisti.re.kr), Yun-Soo Choi(armian@kisti.re.kr),
Sa-Kwang Song(esmallj@kisti.re.kr), Sung-Pil Choi(spchoi@kisti.re.kr),
Hanmin Jung(jhm@kisti.re.kr)

요약

지난 10년간 인터넷과 컴퓨팅 기술의 발전, 모바일 기기와 센서들의 진화, 페이스북이나 트위터와 같은 소셜 네트워크의 출현 등으로 정보량은 급속도로 늘어나고 있다. 대용량의 데이터와 이로 인해 파생되는 방대한 정보는 그것을 얻고자 하는 사람들에게 한계를 느끼게 한다. 따라서 방대한 정보 속에서 의미있는 지식을 추출하기 위한 시스템 기반의 연구가 활발히 시도되고 있다. 이로 인해 지식 추출 시스템의 중요성이 날로 강조되고 있지만, 정확성과 효율성 측면에서 여전히 많은 과제가 있다. 지식 추출 시스템의 성능을 향상시키기 위해서는 시스템을 평가하기 위한 테스트 컬렉션이 중요하다.

본 논문에서는 기술 지식의 자동 추출을 위해 개발된 시스템을 평가하기 위한 테스트 컬렉션을 소개한다. KEEC/KREC(KISTI Entity Extraction Collection/KISTI Relation Extraction Collection)라 명명된 테스트 컬렉션에 대한 구축 절차 및 기준과 구축된 테스트 컬렉션의 특징을 제시한다. 특히 테스트 컬렉션의 주요한 평가 기준이 되는 정확도를 높이기 위해 태깅 지원 도구를 활용한 전문가 태깅 방식을 사용하는 것이 주요 특징이다. 태깅 지원 도구를 활용한 전문가 태깅은 시스템에 의한 자동 태깅 도구들 또는 사람이 태깅을 하되, 지원 도구 없이 태깅하는 방법보다 태깅의 정확도를 높여준다. 구축된 KEEC/KREC은 실제로 과학기술 문헌에 존재하는 PLOT(Person, Location, Organization, Technology) 간 연관관계 추출 성능 평가를 위해서 사용되었고, 의미있는 연구결과를 도출하는데 기여하였다.

■ 중심어 : | 기술 지식 | 테스트 컬렉션 | 개체 추출 | 관계 추출 |

Abstract

For last decade, the amount of information has been increased rapidly because of the internet and computing technology development, mobile devices and sensors, and social networks like facebook or twitter. People who want to gain important knowledge from database have been frustrated with large database. Many studies for automatic knowledge extracting meaningful knowledge from large database have been fulfilled. In that sense, automatic knowledge extracting with computing technology has been highly significant in information technology field, but still has many challenges to go further. In order to improve the effectives and efficiency of knowledge extracting system, test collection is strongly necessary.

In this research, we introduce a test collection for automatic knwoledge extracting. We name the test collection KEEC/KREC(KISTI Entity Extraction Collection/KISTI Relation Extraction Collection) and present the process and guideline for building as well as the features of . The main feature is to tag by experts to guarantee the quality of collection. The experts read documents and tag entities and relation between entities with a tool for tagging. KEEC/KREC is being used for a research to evaluate system performance and will continue to contribute to next researches.

■ keyword : | Technological Knowledge | Test Collection | Named Entity Extraction | Relation Extraction |

접수번호 : #120618-005
접수일자 : 2012년 06월 18일
심사완료일 : 2012년 07월 04일
교신저자 : 최성필, e-mail : spchoi@kisti.re.kr

I. 서 론

오늘날 지구상에는 매일 2,100억 개의 e메일이 발송되고 있고, 미국의 3개 방송사(ABC·NBC·CBS)의 10년간 방송분량 만큼의 동영상이 매일같이 YouTube에 올라오는 등 하루 동안 접하는 정보량이 100년 전 사람들이 평생 접할 정보를 뛰어넘는 정보 폭주의 시대에 살고 있다. 대용량의 데이터와 이로 인해 파생되는 방대한 정보는 그 속에서 유용한 정보 또는 의미있는 지식을 얻고자 하는 사람들에게 한계를 느끼게 한다. 따라서 방대한 정보 속에서 의미있는 지식을 추출하기 위한 시스템 기반의 연구가 활발히 시도되고 있다.

지식 추출은 대량의 문헌을 대상으로 이루어지므로, 시스템에 의한 자동 추출 방식이 일반적이다. 지식 추출 시스템의 성능을 향상시키기 위해서는 시스템을 평가하기 위한 테스트 컬렉션이 중요하다. 테스트 컬렉션이 부족하거나 부정확하다면, 개발된 시스템에 대한 검증이 어렵고, 지식 추출 기술의 발전은 더디게 될 것이다. 특히 급속한 데이터량의 증가를 보이는 과학기술분야에서의 지식 추출은 더욱 중요한 의미를 가진다. 따라서 과학기술 분야 기술 지식의 자동 추출을 위해 개발된 시스템을 평가하고 성능을 향상시키기 위한 테스트 컬렉션 구축이 많이 이루어져야 한다.

Cranfield 컬렉션, CACM 컬렉션, NPL 컬렉션 등 이전의 테스트 컬렉션은 주로 정보 검색을 위해 많이 사용되었고, 이 분야에서 테스트 컬렉션을 이용한 실험은 오랜 역사를 지니고 있다. 반면, 지식 추출을 위한 테스트 컬렉션 구축은 비교적 최근에 많이 이루어지고 있다.

본 논문에서는 과학기술 분야 기술 지식의 자동 추출을 위한 테스트 컬렉션(이하 KEEC/KREC)을 구축하기 위한 절차 및 기준과 구축된 테스트 컬렉션의 특징을 제시한다. 특히, 테스트 컬렉션 구축에 있어서 일반적인 방식이라 할 수 있는 자동 또는 반자동의 구축 방법을 사용하지 않고, 태깅 지원 도구를 활용하여 전문가(숙련자)에 의해 태깅을 통해 테스트 컬렉션의 정확도를 높이고자 노력하였다.

II. 관련 연구

1. 기술 지식

정보와 지식의 차이 관점에서 볼 때, 기술 정보는 필요에 맞게 가공되지 않은 것이고, 기술 지식은 필요에 맞게 가공된 것이다. 불특정 다수의 정보를 선택하고 조합하여 주어진 문제의 시간과 공간에 적합한 해결책으로 활용이 가능할 때 비로소 기술 지식이라 한다[1].

기술 지식과 관련된 기술의 유형을 몇 가지로 구분해 보면, 지식의 구조화, KDD(Knowledge Discovery in Database), 지식모델링, 지식계량화 등 4가지로 나뉜다[2]. 첫째는 지식의 구조화인데, 정보나 지식을 개인화된 정보로 표현하기 위한 일련의 정보 코드화 기술이다. 이 분야는 사용자 제작 DB, 분석형 DB, 지능형 인포메이션 등이 포함된다. 또한 태크케스트, 델파이방법 등 전문가의 지식을 표현하는 방법도 여기에 속하게 된다. 둘째, KDD이다. DB 마이닝, 텍스트 마이닝, 온톨로지 및 시맨틱기술 등이 포함된다. 정보를 자동으로 분류하고 스크리닝, 기계적 추출 등의 영역이 여기에 해당된다. 일반적으로 IT기술을 기반으로 정보로부터 자동적으로 관련 정보나 지식을 찾아내는 기술이다. 셋째, 지식모델링이다. 테크놀로지 인텔리전스, 비즈니스 인텔리전스, 의사결정지원모델 등 정보의 관계분석을 통해서 새로운 지식으로 표현하기 위한 로직과 모델 등이 해당된다. 지식모델링은 KDD나 지식구조화, 지식계량화와 연계된다. 지식모델링은 IT 등의 기술로 시스템적인 접근이 가능하도록 구현된다. 넷째, 지식계량화 및 가시화이다. 정보와 지식의 흐름을 분석하여 지식의 현재 수준과 포지셔닝을 분석하고 미래의 유용한 지식의 패턴을 인식하게 된다. 많은 경우에 지식모델링과 IT기술을 통하여 구현된다.

위의 네 가지 기술지식은 기술지식에 대한 포괄적인 정의를 담고 있다. 데이터, 정보 및 전문가의 지식 등으로부터 유용한 지식을 활용하기까지의 단계를 이르는 일련의 기술영역으로 정의하고 있다. 본 논문에서의 기술지식은 기술문헌 및 웹문서들에서 빈번하게 출현하는 기술개체 및 개체 간 관계정보로 한정한다.

2. 기술 지식 추출

지식추출은 대량의 문헌을 대상으로 지식을 추출하는 개념이므로, 자동화된 추출에 관한 연구가 일반적이다. [그림 1]은 시스템에 의한 일반적인 지식 추출 프로세스이다[3][4]. 텍스트 문서는 문장 또는 단락 단위로 청킹되고, 규칙 또는 패턴들이 개체 식별을 위해 사용된다. 개체 추출 단계에서, 시스템은 추출 대상이 되는 개체와 관련 있는 주변 단어들을 검색한다. 관계 추출을 위해서, 시나리오 수준의 추출 패턴이 사용된다. 어떤 시스템들은 추출된 개체들의 정확도를 파악하기 위해 전체 데이터로부터 수집된 통계치들을 사용하기도 한다. 추출된 정보에 대해서는 중복 제거나 통합되는 등 후처리 작업이 수행된다. 추출된 개체들은 관계에 독립적이어서, 개체 태깅은 다른 추출 작업을 위해 공유될 수 있다.

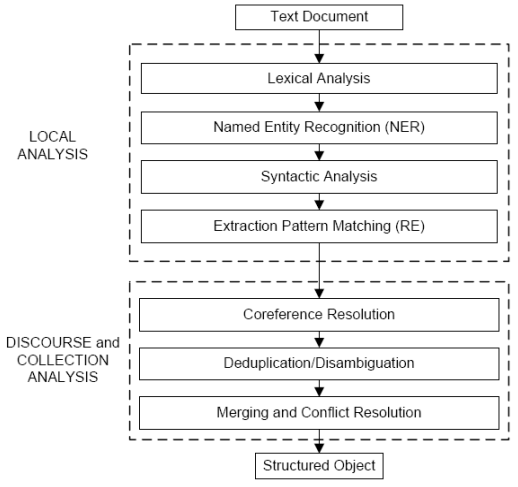


그림 1. 지식 추출 프로세스

기술 지식 추출에서 개체명 인식과 관계 인식은 중요한 개념으로 연구되고 있다. Bikel[5]은 문서 내 개체를 식별하기 위해 표준 은닉 마르코프모델의 변형 기법을 사용한다. 이들은 고유명사의 의미 분별 문제를 주어진 단어열(W)에 대해 가장 적합한 개체명 분류를 찾는 문제라고 할 수 있기 때문에 식 1을 푸는 것과 같다고 설명한다.

$$Pr(NC|W) = Pr(W,NC) / Pr(W) \quad (1)$$

은닉 마르코프 모델은 조건부 확률을 풀기 위한 것으로써, MUC-7(7th Message Understanding Conference) 표준에 기반한 개체명 분류(NameClass, NC), 개체명으로 분류되지 않은 내부적인 분류, 문장에서 시작과 끝을 나타내는 특별한 분류로 구성된다. 단어와 NC들은 아래의 3가지 절차로 생성된다.

가. 이전 단어와 NC의 조건하에 NC의 선택
나. 현재와 이전 NC의 조건하에 NC의 첫 번째 단어 생성
다. NC 내부에서 이전 단어들의 조건하에 단어 생성

이 3가지 과정은 관측된 단어열이 생성될 때까지 반복하게 되고, Viterbi 알고리즘을 사용하여 효율적으로 전체 가능한 개체명 분류의 공간에 전술한 서식을 최대화하기 위한 탐색을 수행하면서 개체명을 인식해 나간다.

관계 추출 기법은 처리 기법에 따라 크게 (1) 규칙기반 방법(rule-based methods), (2) 자질 기반 방법(feature-based methods), 그리고 (3) 커널 기반 방법(kernel-based methods)으로 분류할 수 있다[6].

규칙기반 방법은 문서로부터 관계를 추출하기 위해 문장의 문법이나 특정 관계를 나타내는 패턴을 사용한다. 패턴은 실제 문서에서 추출하고자 하는 정보를 포함하고 있는 문장의 일부를 관계와 비교할 때 사용되는 것이다.

자질 기반 방법은 최대 엔트로피 모델(Maximum Entropy Model)을 기반으로 다양한 형태의 어휘적, 구문적, 의미적 자질들을 이용하여 관계 추출을 시도한다. 이를 기반으로 지지벡터기계를 활용하여 더 확장되고 세분화된 자질 정보를 관계추출에 적용할 수 있다.

이와 유사하게 모든 세부 자질을 종류별로 구분하고 이를 개별적인 선형 커널로 구성하여 최종적으로 혼합 커널로 결합하는 기법도 제안되고 있다. 이 방법은 커널 함수를 직접 고안하여 적용하였다는 점에서 커널 기반 기법으로 분류될 수도 있으나, 커널의 구조가 단순하고 대부분 자질 벡터로 변환될 수 있는 점에 근거하여 자질 기반 방법으로 분류하였다. 두 개의 구문 분석 트리에 대

한 유사도를 재귀적으로 측정하는 연속 부분 트리 커널(contiguous subtree kernel)과 희소 부분 트리 커널(sparse subtree kernel)은 대표적인 커널 기반 기법의 대표적인 구문 트리 커널 기법에 속한다.

3. 테스트 컬렉션

주로 정보과학 분야, 특히 정보 검색을 위해 많이 사용되었고, 이 분야에서 테스트 컬렉션을 이용한 실험은 오랜 역사를 지니고 있다. 정보검색 시스템의 질적 성능을 객관적이고 공정하게 평가하기 위해 외국에서는 오래전부터 테스트 컬렉션을 구축하고 정보검색 시스템의 질적 성능(효율성)을 평가하기 위해 질의 집합 및 적합성 정보를 같이 제공하고 평가를 하고 있다[7].

정보 검색에 있어서 테스트 컬렉션에 대한 구축 연구는 이준호의 연구를 참고하였다[8]. 정보 검색에 대한 연구는 Cranfield I이라고 불리는 색인에 대한 실험과 함께 시작되었으며, 그 후로 30년이 넘는 동안 실험은 검색 기법의 개발에 있어서 필수적인 요소로 인식되어 왔다.

미국의 NIST(National Institute of Standards and Technology)의 후원으로 1992 년에 처음으로 개최된 학술 대회 TREC(Text REtrieval Conference)에서 1백만 건을 초과하는 문서들을 대상으로 대용량 테스트 컬렉션의 구축을 시작하였으며, 이후 매년 테스트 컬렉션에 포함되는 문서들의 수를 증가시키고 있다. 학술 대회 TREC에서는 이러한 테스트 컬렉션을 사용하여 실험실에서 개발된 시스템들 뿐만 아니라 산업체에서 개발된 상용 시스템들도 평가하여 그 결과를 발표하고 있다.

일본에서도 테스트 컬렉션의 중요성을 인식하여 정보기관인 NACSIS(National Center for Information Systems)가 주관이 되어 대규모 테스트 컬렉션 구축 사업을 추진 중이다. 또한 NTT Data Corporation에서 BMIR-J1(Banchmark for Information Retrieval Systems for Japanese texts Ver.1)과 BMIR-J2 (Banchmark for Information Retrieval Systems for Japanese texts Ver.2)라는 테스트 컬렉션을 개발하였으며, BMIR-J1은 600건의 문서와 60개의 질의로 구성되어 있고, BMIR-J2 는 경제학 및 공학 분야에서 5,080 건의 신문기사와 60 개의 질의를 포함하고 있다.

정보 검색용 테스트 컬렉션은 일반적으로 문서 집합, 질의 집합 그리고 각 질의에 대한 적합 문서 리스트로 구성된다. 이들 중 검색의 대상이 되는 문서 집합은 테스트 컬렉션 구축에 있어서 가장 기본적인 요소이다. 문서 집합의 구성에 있어서 고려해야 할 사항은 다양한 분야 및 크기의 문서들로 문서 집합을 구성해야 한다는 것이다. 즉, 정보 검색 기술은 많은 경우에 통계적인 방법이나 언어 처리 기술을 사용하는데, 이들은 모두 문서의 종류에 많은 영향을 받는다. 또한, 문서와 질의의 유사도 계산에 핵심적 역할을 하는 가중치 기법들 중 일부는 특정 크기의 문서들에 높은 유사도를 부여하는 특성을 지니고 있기 때문에 가중치 기법의 성능 평가를 위해서라도 다양한 크기의 문서들로 문서 집합을 구성하는 것이 바람직하다.

정보 추출을 위한 테스트 컬렉션 구축은 비교적 최근에 많이 이루어지고 있다. 정보 추출 분야에 기계학습 모델을 이용하기 위해서는 학습과 테스트 및 평가를 위한 테스트 컬렉션 구축이 필수적이다. MUC(Message Understanding Conference)나 ACE(Automatic Content Extraction)와 같은 테스트 컬렉션은 신문기사, 뉴스 등으로 한정되어 있고, 이것을 사용하기 위해서는 많은 비용을 지불해야 하기 때문에 대다수의 일반 연구자들은 자체적으로 테스트 컬렉션을 구축하여 성능 평가를 수행하고 있다. 정창후 외 3명은 반자동화된 처리 과정을 거쳐서 규모 있는 관계 추출용 테스트 컬렉션을 구축하는 프레임워크를 제안한다[9]. 그리고 개발된 프레임워크를 이용하여 실제적으로 과학기술 문헌에 존재하는 기술용어 간 연관관계 추출 시스템의 성능 평가를 위한 테스트 컬렉션을 구축하고 결과를 분석한다. [그림 2]와 같이 반자동 테스트 컬렉션 구축 프레임워크는 문헌의 구문적 특성과 의미적 특성을 시스템적으로 처리하여 후보 트리플을 생성하는 자동 처리 과정과 후보 트리플 중에서 가장 적합한 트리플을 구축자가 최종적으로 선택하는 수동 처리 과정으로 이루어진다.

본 논문에서는 자동 또는 반자동의 구축 방법을 사용하지 않고, 테스트 컬렉션의 정확도를 높이기 위해, 태깅 지원 도구를 활용하여 전문가(숙련자)에 의해 태깅이 이루어졌다. 전문가에 의한 태깅 방식은 다음 장에서 상세

히 기술한다.

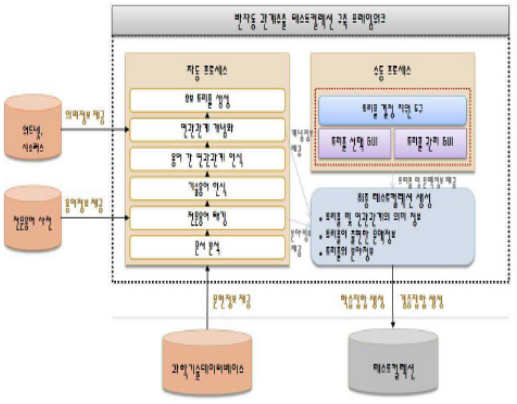


그림 2. 반자동 테스트 컬렉션 구축 프레임워크 구조

III. 테스트 컬렉션 구축

1. 구축 절차

KEEC/KREC 구축을 위한 절차는 [그림 3]과 같다. 먼저 테스트 컬렉션 구축의 대상이 되는 원시 데이터를 수집한다. 텍스트 기반의 전자 문서가 주 대상이며, 논문이나 특허의 원문이나 초록이 될 수 있다. 본 논문에서는 KISTI NDSL의 저널 중 환경/에너지 분야 281,069건(193종)과 PUBMED 저널 중 환경/에너지 분야 91,000건(105종)을 원시 데이터로 수집하였다. 기계적인 추출 작업이 아닌 전문가에 의한 태깅에 기반하였기 때문에, 원시 데이터 전체를 대상으로 하는 것은 시간과 비용이 많이 든다. 따라서 원시 데이터 중에서 전체 데이터의 특성을 반영할 수 있는 기준을 마련하고, 이 기준 하에 무작위로 후보 데이터셋을 선정한다. 선정된 문서들을 대상으로 태깅 지원 도구를 활용하여, 전문가에 의해 개체와 관계를 태깅하게 된다. 이때 사전에 마련된 태깅 기준에 의해 태깅 작업이 수행된다. 전문가들은 상호 검증(peer review)을 거침으로써, 태깅 작업의 정확도를 높일 수 있다. 태깅된 정보는 지원 도구에 의해 XML파일로 저장된 후, 개체정보와 관계정보로 저장된다.

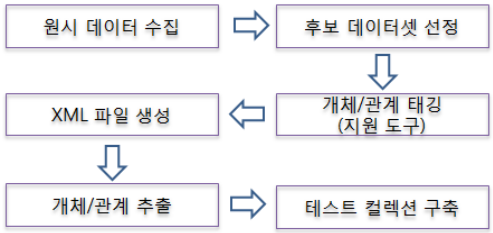


그림 3. KEEC/KREC 구축 절차

2. 개체/관계 태깅 기준

개체 추출은 Green Technology(GT) 분야를 대상으로 하여, PUBMED, NDSL 데이터의 문장에 대해서 사전에 지정해 놓은 클래스[표 1]인 PLOT(Person, Location, Organization, Technology), 환경전문용어(Environmental Technical term), 환경작용용어(environmental influence factor)에 해당되는 개체에 대해서 태깅한다. 기술용어(Technology Term), 생물유전자(Gene), 생물단백질(Protein), 질병(Disease), 생체 구성요소(organ), 약명(Drug), 기타(Others)에 대해서도 문장에 포함된 개체가 있으면 태깅한다.

표 1. 개체 유형(class) 및 표현

구분	개체명	영문 표현
PLOT	사람	Person
	위치	Location
	조직	Organization
	기술용어	Tech Term
바이오	유전자	Gene
	단백질	Protein
	질병	Disease
	생명체의 구성요소	Organ
	약명	Drug
환경	환경작용인자	environmental influence factor
	환경전문용어	environmental technical term
기타	기타	Others

대상 데이터의 문장 속에서 각각의 클래스에 해당되는 개체를 태깅하는 것이 기본이다. 각 클래스별로 개체를 부여하는 방법은 다음과 같다.

- 사람 : 인명은 태깅하고, 직업이나 사람을 가리키는 보통 명사는 태깅하지 않는다. 직위, 직책(Dr. 등)과

함께 쓰이는 이름일 경우, 직위/직책은 빼고 태깅한다.

- 위치 : 지명이나 장소 등을 태깅한다. 만일, 약자로 표시된 경우는 약자를 태깅한 뒤, NN에 전체 이름을 모두 기입한다.
- 조직 : 기관, 조직 이름을 태깅한다. 만일 약자로 표시된 경우는 약자를 태깅한 뒤, NN에 전체 이름을 모두 기입한다.
- 기술용어 : 기술적 행위가 포함된 용어를 태깅한다. 그 기술로써 치료가 가능하고 상태를 보완하는 등 따위의 행위적 의미를 포함하는 용어를 기술용어라 정의하고 태깅한다.
- 유전자 : 유전자에 해당하는 용어를 태깅한다.
- 질병 : 질병의 이름을 태깅하나, 병의 증상, 징후 등도 광범위하게 질병으로 간주하여 병명으로 태깅한다.
- 약명 : 약명(고유명사)과 질병치료제를 통칭하는 약명칭 또한 태깅하여도 무방하다.
- 환경작용인자 : 그 문서에서 환경에 영향을 미치는 용어를 태깅한다. 그 영향이 나쁘거나 좋은 것은 상관하지 않는다.
- 환경전문용어 : 환경 분야(Green Technology)에서 중요하게 사용되는 용어라고 판단될 경우 태깅한다. 그 명사가 보통 문서에서는 일반명사라고 간주될지 모르지만, 환경 분야에서는 전문용어라고 판단되는 경우는 태깅하여 준다.

환경을 대상으로 한 전문용어는 일반명사로 보이는 전문용어도 많지만 환경에 영향을 미치거나 환경에 영향으로 변화되는 대상은 환경전문용어로 볼 수 있다고 판단하여 개체를 태깅할 수 있다. 그러나 환경에 의해서 영향을 받는 생물(예. 왜가리, 암소, 돼지 등)은 환경전문용어로 볼 수 없으므로, 개체로 추출하지 않는다. 환경전문용어 중에 기술적인 요소를 포함하는 용어가 있을 때에는 환경용어가 아닌 기술용어(Technology Term)로 태깅하도록 한다.

위와 같은 기준에 의해 개체를 추출 후, 추출된 개체 간에 결정한 관계 클래스[표 2]를 기준으로 관계를 태깅한다. 관계 태깅을 할 경우에는 문장 속에서 유추하는 것이 아니고, 문장 속의 동사나 어구에서 명확하게 드러나는 관계만을 태깅한다.

표 2. 관계 유형(class) 및 표현

관계명	영문 표현
사업적 관계	business
혈연적 관계	family
고용	hire, engage, employ
관리	manage, deal, care, handle
일원/회원	membership
설립	establish, set up, found, launch
근무	be on duty, work for, serve
위치	locate, place, site
활용/적용	use, utilize, utilise, apply, employ
변화/변경	change, alter, modify
야기/유발/자극 (원인-결과 관계)	induce, stimulate, cause, have, get, make
창조/생산	produce, make, create
획득	get, acquire
분석	analyze, analyse, study, examine, canvass, canvas
포함	include
억제	suppress, stamp down, inhibit, subdue, conquer, curb
증가	increase
방지	prevent, keep
치료	remedy, curative, cure, therapeutic
활성화	activate
영향	affect, impact, bear upon, bear on, touch on, touch
의존	depend
반응/응답	react, respond
동일	equal, be
부분	part, portion, component part, component
소유	own, have, possess
제공	yield, give, afford
도움	help, assist, aid
연구	research
전달	convey, transmit, communicate
해결	solve, work out, figure out, puzzle out, lick, work
발견	discover
결합/연결	bond, unite
해체/분해	dissolve, disintegrate, break down
강화	fortify
약화	weaken
감소	decrease
결과-원인 관계	result from, stem from
조절/조정	modulate, regulate, control
이동	move
악화	aggravate, deteriorate, degenerate
유지, 지속	maintain, sustain
무관계	be not related with
유관계	be related with
불명확	be unclear

관계 태깅의 기준은 아래와 같다. 아래에 명시되지 않은 관계들은 유사한 방법으로 정의될 수 있다.

- 치료관계 : 한 개체가 다른 개체에 대해서 치료하는 개념이 명시될 때 치료관계로 태깅한다.
- 원인-결과 관계 : 두 개체가 원인 - 결과로 명시되면 원인-결과 관계로 태깅한다.
- 결과-원인 관계 : 두 개체가 결과 - 원인으로 명시되면 결과-원인 관계로 태깅한다.
- 동일 관계 : 두 개체가 동일하다고 명시되면 동일 관계로 태깅한다.
- 위치 관계 : 두 개체가 위치의 관계를 이루면 위치 관계로 태깅한다.

특이사항으로는 두 개체 간의 관계에서 앞의 개체가 뒤의 개체에 대해 수동적으로 관계를 형성할 때, Passive로 처리한다는 것이다. 또한 두 개체가 관련이 있으나, 위의 클래스 포함되지 않을 경우, 의미적으로 유추하여 관계를 설정하지 말아야 한다.

3. 구축 방법 및 결과

본 논문에서는 테스트 컬렉션 구축을 위해 전문가(숙련자)에 의한 태깅 방식을 사용하였다. 태깅 지원 도구를 활용한 전문가 태깅은 시스템에 의한 자동 태깅 도구들 또는 사람이 태깅을 하되, 지원 도구 없이 태깅하는 방법보다 태깅의 정확도를 높여주는 역할을 한다. 태깅 지원 도구가 없다면, 작업자는 하드카피로 된 원시 문서를 직접 읽거나 모니터 상에서 해당 문서를 보면서 문서의 여백이나 별도의 메모창에 태깅을 해야 한다. 이 경우 취합하는 과정에서 태깅 정보를 유실할 수 있다. 또한 문장 단위로 파싱된 상태가 아니기 때문에, 작업자의 집중을 더 필요로 한다. 작업자는 더 쉽게 피로해지고, 태깅의 품질이 낮아질 가능성이 높다. 시간도 더 오래 걸리게 된다. 자동 태깅 도구들은 사람에 의한 작업보다 효율적일 수는 있지만, 기계가 하는 일이기 때문에 사람에 의한 태깅보다 정확도가 떨어질 수 있다. 따라서 태깅 지원 도구는 전문가의 판단에 의한 태깅이라는 면에서 더 정확한 결과를 얻을 수 있으며, 태깅 도구 없이 하는 방법보다 정확성과 효율성을 갖추었다고 할 수 있다.

전문가에 의한 태깅 작업에 사용된 지원 도구는 [그림 4]와 같다. 지원 도구는 텍스트 파일을 읽어들여, 문장 단위로 파싱을 하여 사용자에게 보여준다. 사용자는 각각의 문장을 보면서, 개체 추출 기준에 따라, 개체를 태깅한다. 태깅할 단어를 선택하면, 개체 클래스를 볼 수 있고, 그중 적합한 클래스를 선택할 수 있다. 관계 태깅도 유사한 방법으로 할 수 있다. 태깅된 개체를 드래그로 연결하면 관계 클래스를 볼 수 있고, 적절한 관계를 선택할 수 있다.

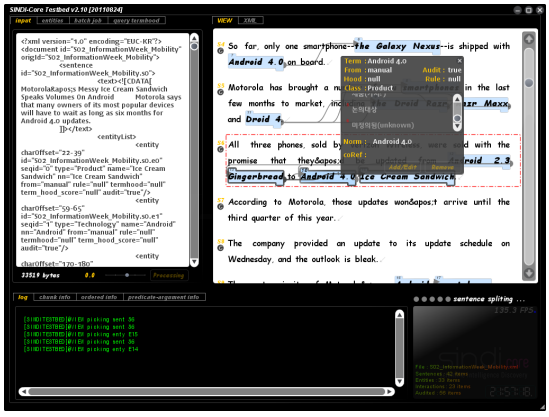


그림 4. 태깅 지원 도구

KISTI NDSL의 저널 중 환경/에너지 분야 281,069건(193종)과 PUBMED 저널 중 환경/에너지 분야 91,000건(105종)을 대상으로 무작위로 추출된 1,088개의 코퍼스의 크기는 5.74MB이며, 파일의 타입은 XML이다. 각 파일은 최대 35KB부터 최소 2KB까지 다양하며, 파일 당 평균 5.28KB의 크기를 가진다. 코퍼스로 선정된 1,088개 문서에서 각 문서별로 최종 태깅 된 결과는 [표 3]과 같다.

표 3. KEEC/KREC 구축 결과

구분		건수
원시 데이터		KISTI NDSL 281,069건 PubMed 91,000건
코퍼스	문서 수	1,088개
	문장 수	14,306개
기술 지식	개체 태깅 수	10,023개
	관계 태깅 수	1,566개

각 파일이 가지고 있는 문장의 수도 다양하다. 최대 50개 문장을 가진 파일이 있고, 최소 2개의 문장을 가진 파일도 있다. [표 4]는 한 파일에 포함되어 있는 문장수의 분포를 나타낸다. 문장수의 분포의 분포를 10단계로 나누었고, 각 단계별로 5개의 문장을 할당하였다. 문서수의 합은 문장수 11~15개와 6~10개에서 최고를 기록하였고, 개체수와 관계수도 비슷한 분포를 보였다.

표 4. KEEC/KREC 구축 결과 분석

문장수 분포	문서수합	개체수합	관계수합
46~50	5	148	9
41~45	9	261	21
36~40	10	219	19
31~35	18	381	47
26~30	34	673	111
21~25	84	1340	213
16~20	155	1764	267
11~15	299	2745	482
6~10	326	2006	325
1~5	148	486	72

[그림 5]는 문장수의 분포에 따른 해당 분포 내의 문서수합, 개체수합, 관계수합을 표현한 것이다. 이 그림에서 문장수 11~15과 6~10 부근에서 문서수합, 개체수합, 관계수합이 가장 높고, 비슷한 분포를 보이는 것을 알 수 있다. 따라서 개체수합, 관계수합은 문서수합에 비례함을 알 수 있고, 이는 문서가 많을수록 태깅된 개체수와 관계수가 많았음을 의미한다. 이를 통해 전문가에 의한 태깅 작업이 어느 정도는 제대로 이루어졌다고 판단할 수 있다.

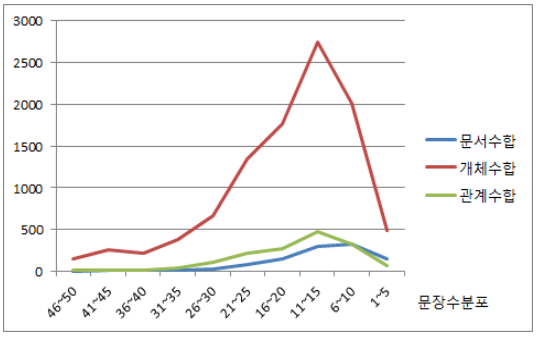


그림 5. 문서수합, 개체수합, 관계수합

이와 같은 방법으로 구축된 KEEC/KREC 데이터의 포맷은 XML이며, [표 5]와 같은 구조를 가지고, 실제 사용 예는 [그림 6]과 같다.

표 5. KEEC/KREC 데이터 포맷

엘리멘트	설명
document	최상위 엘레멘트
sentence	문서내의 한 문장을 표현하는 엘레멘트
text	한 문장의 내용
entityList	한 문장에서 추출된 후보 용어 목록
entity	한 문장에서 추출된 후보 용어
pair	개체 간 관계 표현
chunk	품사 정보를 이용하여 구성되는 명사구 등의 후보
chunkList	후보명사구 목록
parsingResultList	문장 파싱결과 목록
parsingResult	문장 파싱 결과

```
<?xml version="1.0" encoding="EUC-KR" ?>
<document id="6266218" origId="6266218">
    <sentence id="6266218.s0">
        <text><![CDATA[
        ]]></text>
        <entityList>
            <entity charOffset=".." .... />
            ....
        </entityList>
        <chunkList>
            <chunk charOffset=".." .... />
            ....
        </chunkList>
        <parsingResultList>
            <parsingResult parser="enju" structure="ordered"><![CDATA[
                ....
            ]]></parsingResult>
            <parsingResult parser="enju" structure="predicate-argument"><![CDATA[
                ....
            ]]></parsingResult>
        </parsingResultList>
        <pair e1="NDSL.d56.s10.e0" e2="NDSL.d56.s10.e1" id="NDSL.d56.s10.p0"
        interaction ="induce" psv="0" />
    </sentence>
    <sentence id="....">
        ....
    </sentence>
</document>
```

그림 6. KEEC/KREC 데이터 예시

Ⅳ. 테스트 컬렉션 활용

본 논문에서 구축된 KEEC/KREC은 실제로 과학기술 문헌에 존재하는 PLOT 간 연관관계 추출 성능 평가를 위해서 사용되었다.

정창후 외 3명은 KEEC/KREC을 기반으로 PLOT 간 연관관계 자동 분류에 대한 성능 평가를 실시하였다[10]. 혼합 커널의 보다 정확한 성능 비교를 위해서 우선 일반 구문 트리 커널과 술어-논항 구조의 패턴 유사도 커널을

각각 단독으로 사용한 경우에 대해서 실험하고, 최종적으로 두 방법을 결합한 상태로 사용한 경우의 성능 측정 결과를 비교한다.

[표 6]은 PLOT 간 연관관계 추출 실험에 대한 성능 평가 결과를 보여준다. 트리 커널과 술어-논항 구조 패턴 유사도 커널을 단독으로 수행했을 때는 트리 커널의 성능이 술어-논항 구조 패턴 유사도 커널보다 더 좋은 것을 확인할 수 있다. 하지만 트리 커널 단독으로 사용하는 것보다는 술어-논항 구조 패턴 유사도 커널을 결합하여 혼합 커널을 구성하였을 때는 더 나은 성능을 보이는 것을 확인할 수 있다. 따라서 기존의 트리 커널은 술어-논항 구조 패턴 유사도 커널과 결합하여 더 나은 성능을 발휘한다는 사실을 알 수 있다.

표 6. PLOT 간 연관관계 추출 성능

커널 종류	micro-averaged F-score(%)	macro-averaged F-score(%)
술어-논항 구조 패턴 유사도 커널	64.60	33.69
구문 트리 커널	68.78	38.09
혼합 커널	74.72	42.33

위 사례는 KEEC/KREC을 활용하여 지식 추출 도구의 성능을 평가하고, 의미있는 연구결과를 얻을 수 있다는 사실을 보여주고 있다. 테스트 컬렉션의 구축 목적이 시스템의 성능 평가 등 다양한 실험에 활용되는 것이기 때문에, KEEC/KREC 데이터를 활용하여 유의미한 실험 결과를 얻었다는 것 자체가 해당 컬렉션 데이터의 유용성 및 활용성을 검증하는 것이라고 할 수 있다. 향후에도 KEEC/KREC 데이터를 가능한 많은 연구자들과 공유하여, 기술 지식 자동추출 기술의 향상에 조금이라도 기여할 수 있도록 노력하고자 한다.

V. 결론

정보의 폭주 시대에 의미있는 기술 지식을 효율적으로 추출하기 위한 연구는 계속되고 있다. 하지만, 품질이 보장된 테스트 컬렉션이 없다면, 기술의 진보는 더디게 될 것이다. 특히 급속한 데이터량의 증가를 보이는 과학기술분야에서의 지식 추출은 더욱 중요한 의미를 가진다.

본 논문에서는 과학기술 분야 기술 지식의 자동 추출을 위한 테스트 컬렉션을 구축한다. 테스트 컬렉션을 구축하기 위해서는 체계적인 구축 절차가 필요하고, 기술 지식의 특성 상 개체 및 관계 태깅을 위한 기준이 필요하다. 실제 태깅을 위해서는 일반적인 자동화된 태깅 방법이 아닌, 전문가에 의한 태깅 방법을 시도하였다. 또한 태깅 지원 도구를 활용하였다. 자동 태깅 도구들은 사람에 의한 작업보다 효율적일 수는 있지만, 기계가 하는 일이기 때문에 사람에 의한 태깅보다 정확도가 떨어질 수 있다. 따라서 태깅 지원 도구는 전문가의 판단에 의한 태깅이라는 면에서 더 정확한 결과를 얻을 수 있으며, 태깅 도구 없이 하는 방법보다 정확성과 효율성을 갖추었다고 할 수 있다. 즉, 기존의 테스트 컬렉션들은 시스템에 의한 자동으로 구축되고, 일부분 사람에 의해 검증하는 방식이지만, KEEC/KREC은 처음부터 전문 도구를 사용하여 전문가에 의해 수동으로 구축함으로써, 정확도를 높이고자 노력하였다.

이와 같은 과정을 통해 구축된 KEEC/KREC은 실제 과학기술 지식 추출 도구의 성능을 평가하기 위한 연구에 활용된다. 향후에도 지속적으로 KEEC/KREC 데이터를 가능한 많은 연구자들과 공유하여, 기술 지식 자동추출 기술의 향상에 기여하고자 한다.

참 고 문 헌

[1] 1 http://ko.wikipedia.org

[2] 한국과학기술단체총연합회, 새로운 연구・비즈니스 분야로 등장하는 지식기술, The Science & Technology, 2012(2).

[3] R. Grishman, “Information extraction: Techniques and challenges. In Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology,” International Summer School, pp.10-17, 1997.

[4] E. Agichtein, “Scaling Information Extraction to Large Document Collections,” IEEE, 2005.

[5] D. Bikel, "Nymble:A High-Performance Learning Name-Finer," In proceedings of 5th Conference on Applied Natural Language Processing, p.194, 1997.

[6] 최성필, 정창후, 최윤수, 맹성현, "평면적 어휘 자질들을 활용한 확장 혼합 커널 기반 관계 추출," 정보과학회논문지 : 소프트웨어 및 응용, 제36권, 제8호, pp.642-652, 2009(8).

[7] 강현규, 전홍석, 오염덕, "정보 검색 시스템 평가를 위한 한텍(HANTEC) 적합성 정보의 평가 및 수정 구축," 한국정보기술학회논문지, 제9권, 제4호, pp.167-172, 2011(4).

[8] 이준호, *정보검색이론*, 숭실대학교, 2003(3).

[9] 정창후, 최성필, 이민호, 최윤수, "기술용어 간 관계 추출의 성능평가를 위한 반자동 테스트 컬렉션 구축 프레임워크 개발," 한국콘텐츠학회논문지, 제10권, 제2호, pp.1-8, 2010(2).

[10] 정창후, 최성필, 최윤수, 송사광, 전홍우, "술어-논항 구조의 패턴 유사도를 결합한 혼합 커널 기반 관계 추출," 한국인터넷정보학회 논문지, 제12권, 제5호, pp.73-85, 2011(10).

저 자 소 개

신 성 호(Sungho Shin) 정회원

- 2000년 : 경북대학교 경영학과(학사)
- 2002년 : 경북대학교 경영학과-경영정보시스템(석사)
- 2002년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 정보추출, 지식공학, 시맨틱웹, MIS

최 윤 수(Yun-Soo Choi) 정회원

- 1993년 : 충남대학교 컴퓨터공학과(학사)
- 1995년 : 충남대학교 컴퓨터공학과(석사)
- 1995년 ~ 현재 : 한국과학기술정보연구원 선임연구원

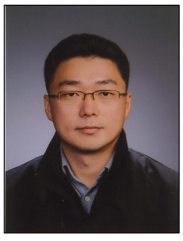
<관심분야> : 정보검색, 텍스트마이닝

송 사 광(Sa-Kwang Song) 정회원

- 1999년 : 충남대학교 컴퓨터과학과(석사)
- 2011년 : 한국과학기술원 전산학과(박사)
- 2010년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 텍스트마이닝, 자연어처리, 정보검색, 시맨틱웹

최 성 필(Sung-Pil Choi) 정회원



- 1998년 : 부산대학교 전자계산학과(석사)
- 2012년 : 한국과학기술원 정보통신공학과(박사)
- 1998년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 기계학습, 정보검색, 자연어처리, 정보추출, 텍스트마이닝

정 한 민(Hanmin Jung) 정회원

- 2003년 : 포항공과대학교 컴퓨터공학과(박사)
- 2004년 ~ 현재 : 한국과학기술정보연구원 책임연구원
- 2004년 ~ 현재 : 과학기술연합대학원대학교 겸임교수

<관심분야> : 시맨틱웹, 정보검색, 자연어처리, HCI