# 기술용어 간 관계추출의 성능평가를 위한 반자동 테스트 컬렉션 구축 프레임워크 개발

## Development of a Framework for Semi-automatic Building Test Collection Specialized in Evaluating Relation Extraction between Technical Terminologies

**정창후, 최성필, 이민호, 최윤수**
한국과학기술정보연구원 정보기술연구실

Chang-Hoo Jeong(chjeong@kisti.re.kr), Sung-Pil Choi(spchoi@kisti.re.kr),
Min-Ho Lee(cokeman@kisti.re.kr), Yun-Soo Choi(armian@kisti.re.kr)

**요약**

관계 추출 시스템의 중요성이 날로 부각되면서 이러한 시스템을 평가하기 위한 테스트 컬렉션의 구축이 중요한 과제로 떠오르고 있다. 본 논문에서는 반자동화된 처리 과정을 거쳐서 규모 있는 관계 추출용 테스트 컬렉션을 구축하는 프레임워크를 제안한다. 그리고 개발된 프레임워크를 이용하여 실제적으로 과학기술 문헌에 존재하는 기술용어 간 연관관계 추출 시스템의 성능 평가를 위한 테스트 컬렉션을 구축하고(관계유무 파악 및 관계분류 식별을 검사할 수 있는 1,707건의 문장 규모) 결과를 분석한다. 제안된 방법론은 정형화되고 시간이 많이 소요되는 문서분석 작업을 처리과정별로 자동화함으로써 구축에 들어가는 비용을 최소화할 수 있고, 시스템의 알고리즘을 기반으로 동작하기 때문에 구축자의 성향에 따른 편차를 줄이고 일관된 결과물을 얻을 수 있다. 또한 문헌 집합(과학기술 전 분야에 걸친 30,858,830건의 학술 데이터베이스) 및 용어 사전(16개 분야 253,603건 규모의 전문용어) 선정 시 특정 분야에 편중되지 않도록 노력함으로써 균형 잡히고 객관화된 테스트 컬렉션을 생성할 수 있다.

■ 중심어 : | 테스트 컬렉션 | 관계 추출 | 기술 용어 | 프레임워크 |

**Abstract**

Due to the increase of the attention on relation extraction systems, the construction of test collections for assessing their performance has emerged as an important task. In this paper, we propose semi-automatic framework capable of constructing test collections for relation extraction on a large scale. Based on this framework, we develop a test collection which can assess the performance of various approaches to extracting relations between technical terminologies in scientific literatures. This framework can minimize the cost of constructing this kind of collections and reduce the intrinsic fluctuations which may come from the diversity in characteristics of collection developers. Furthermore, we can construct balanced and objective collections by means of controlling the selection process of seed documents and terminologies using the proposed framework.

■ keyword : | Test Collection | Relation Extraction | Technical Terminology | Framework |

접수번호 : #091103-003
접수일자 : 2009년 11월 03일
심사완료일 : 2010년 02월 02일
교신저자 : 최윤수, e-mail : armian@kisti.re.kr

## I. 서론

인터넷의 발전으로 인해 방대한 정보가 유통되면서 사용자의 정보에 대한 요구도 다양해지고 있다. 기존의 정보 검색이나 정보 분류를 뛰어넘어 이제는 정보에 대한 요약 및 핵심 정보 추출과 같은 좀 더 세밀한 정보의 가공을 요구하고 있는 추세이다. 이러한 흐름의 일환으로 텍스트 문서 집합에 존재하는 개체들 사이의 의미적 연관관계를 추출하는 관계 추출 시스템의 중요성이 날로 부각되고 있다. 하지만 시스템 개발과 관련하여 현재 큰 문제점이 되고 있는 것은 이러한 시스템의 성능을 평가할 지표가 부족하다는 것이다.

관계 추출 시스템의 객관적인 비교 평가는 문서에서 중요하게 인식되는 핵심개체와 이들 간의 연관관계로 이루어진 트리플 집합이 제대로 갖추어졌을 경우에 가능하다. 다시 말해서, 관계 추출 시스템의 객관적인 신뢰도 평가를 위해서는 체계적으로 구축된 테스트 컬렉션이 필요하다. 이러한 테스트 컬렉션은 해당 분야의 연구뿐만 아니라 상용화 시스템의 성능을 평가하여 적절한 시스템을 선택하는 데에도 매우 중요한 역할을 하므로 관련 기술의 발전뿐만 아니라 궁극적으로는 정보 유통에 있어서의 경쟁력 강화에도 필수적인 역할을 수행한다[1]. 따라서 정보 시스템을 개발할 때에는 응용 분야에 맞게 구축된 테스트 컬렉션을 사용하여 시스템의 평가를 수행하는 과정이 필연적으로 따라오게 된다[2-7].

본 논문의 구성은 2장에서 관계 추출 시스템의 성능 평가를 위한 기존의 테스트 컬렉션 구축 연구에 대해서 살펴보고, 3장에서는 다양한 언어 자원을 기반으로 한 반자동 테스트 컬렉션 구축 프레임워크를 제안한다. 다음으로 4장에서 프레임워크를 이용하여 실제적으로 기술용어 간 관계 추출 테스트 컬렉션을 구축한 내용에 대해서 설명하고, 마지막으로 5장에서 결론 및 향후 연구에 대해서 기술한다.

## II. 관련 연구

국제적인 관계 추출 평가 대회인 ACE(Automatic Collection Extraction)[8]에서는 1990년 초부터 관계 추출 시스템의 평가를 위한 다양한 테스트 컬렉션을 구축해오고 있다. MUC[9]의 성공적인 연구결과에 고무된 NIST와 DARPA는 본격적으로 보다 고차원적인 정보 추출 기법을 위한 기반 인프라 구축을 시도하였으며, 그 결과 ACE 검증 컬렉션이 매년마다 구축되고 이를 기반으로 수행된 많은 연구결과를 바탕으로 워크숍을 개최하고 있다. 현재까지 일반에게 공개된 학습 집합은 2002년부터 2005년까지 구축된 버전이며 LDC(Linguistic Data Consortium)[10]를 통해서 유료로 배포하고 있다. 그러나 MUC나 ACE와 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[11-13]. 이러한 방법으로 테스트 컬렉션을 생성하는 접근법의 문제점은 투입할 수 있는 인력 및 자원의 제약으로 인하여 테스트 컬렉션의 규모와 대상 분야가 작아질 수밖에 없고, 검증하는 작업도 한계가 있기 때문에 구축자의 개인적인 견해가 상당 부분 반영될 소지가 있다는 것이다. 이러한 경우에 해당 테스트 컬렉션을 사용하여 평가한 결과를 일반화하기에는 다소 무리가 있다.

시스템 혹은 관련 기술을 평가할 때 일정 수준 이상의 규모로 구축된 테스트 컬렉션을 사용하는 것은 정확한 평가를 위해 필수적이다[1]. 본 연구에서는 반자동화된 처리 과정을 거쳐서 과학기술문헌으로부터 규모 있는 관계 추출용 테스트 컬렉션을 구축하는 프레임워크에 대해서 설명한다. 프레임워크를 이용함으로써 구축에 들어가는 비용을 최소화할 수 있고 구축자의 성향에 따른 결과물의 편차를 줄일 수 있다. 또한 문헌 집합 및 용어 사전 선정 시 특정 분야에 편중되지 않도록 노력함으로써 균형 잡히고 객관화된 테스트 컬렉션을 생성할 수 있다. 테스트 컬렉션 구축 시 필요한 기반 데이터로 과학기술 전 분야에 걸친 30,858,830건의 해외 학술 데이터베이스[1]를 사용한다. 또한 기술용어 탐색 및 매칭을 위해서 16개 분야 253,603건 규모의 전문용어 사전[2]을 사용한다. 이를 기반으로 문서 내에 출현하는 기

1) NDSL, http://www.ndsl.kr

술용어와 이들 간의 의미적 연관관계 후보 집합을 자동으로 인식하여 구축자에게 제공함으로써 균형 잡히고 규모가 있는 관계 추출 테스트 컬렉션을 구축할 수 있다. 테스트 컬렉션은 서로 의미 있는 관계를 형성하는 기술용어 집합, 용어 간의 상관성을 설명하는 연관관계 집합, 그리고 기술용어에 대한 연관관계의 적합성 판정을 거친 트리플 집합 등으로 구성된다.

본 논문에서는 정형화되고 시간이 많이 소요되는 작업을 자동화함으로써 구축에 들어가는 비용을 최소화할 수 있는 프레임워크 기반 테스트 컬렉션 구축 방법론을 제안한다. 대규모의 학술 데이터베이스와 다양한 분야의 전문용어 사전, 그리고 최신의 기계학습 알고리즘을 활용하여 규모 있고 실용적인 테스트 컬렉션을 구축한다는 점에서 기존의 테스트 컬렉션 구축 방법론과 차별성이 있다.

## III. 테스트 컬렉션 구축 프레임워크 개발

과학기술 문서에 출현하는 기술용어 및 이들 간의 연관관계를 수동으로 설정하는 일은 매우 어려운 작업이다. 적용 대상이 특정 분야에 한정된 경우라면 해당 분야 전문가에 의해서 기술용어 식별이나 연관관계 설정 작업이 이루어질 수 있지만, 이 역시도 매우 까다로운 작업이며 세분화된 설정 기준(미리 정의된 연관관계 집합, 관계설정 방법 및 판단기준 등)과 분야 전문가의 어휘적 판단 능력 등이 요구된다.

본 연구에서는 이러한 난점들을 극복하고 보다 폭넓은 분야에 속하는 기술용어 간의 연관관계를 처리하기 위해서 용어 쌍을 포함하는 문장 내에서의 관계표현 디스크립터를 워드넷(WordNet)[14]의 상위어(hypernym) 관계를 이용하여 개념을 일반화시킨 후에 연관관계로 활용하는 방법을 사용한다. 이러한 방법을 통하여 기술용어 간의 후보 연관관계를 자동으로 제시하고 이들 중에서 가장 적합한 관계를 구축자가 최종적으로 선택하도록 한다.

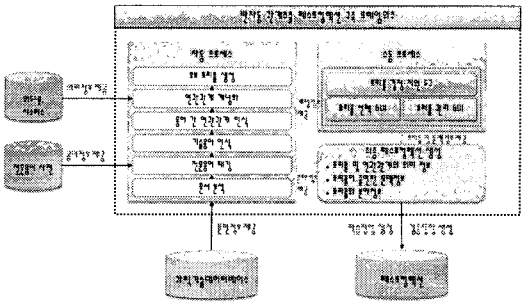


그림 1. 테스트 컬렉션 구축 프레임워크의 구조

[그림 1]과 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. 이 두 작업을 거쳐서 테스트 컬렉션이 생성되는데, 테스트 컬렉션은 트리플 및 연관관계의 의미 정보, 트리플이 출현한 문맥 정보(트리플이 추출된 문장), 트리플의 분야 정보 등을 포함한다. 테스트 컬렉션을 구축하는 과정을 좀 더 구체적으로 살펴보면 다음과 같다.

① 문서 분석: 원본 데이터베이스를 분석하는 기능으로 문서에 대한 구문 분석과 더불어 품사 부착(tagging), 기저구 인식(chunking) 등의 작업을 수행한다. 이 과정에서 어휘변형을 해소하고 복합어 처리를 위한 다양한 특수 규칙이나 알고리즘을 사용한다.

② 전문용어 부착: 문헌에 존재하는 전문용어를 식별하는 기능으로 16개 분야 253,603건 규모의 전문용어 사전을 사용한다.

③ 기술용어 인식: 문서에서 중요한 의미를 가지고 있는 기술용어를 식별하고 이를 추출 및 정제하는 기능으로 기술용어를 식별하기 위해서 CRF(Conditional Random Fields)와 같은 기계학습 알고리즘을 사용한다[15-17].

④ 기술용어 간 연관관계 인식: 식별된 기술용어 간

2) KISTI에서 구축한 전문용어 사전으로 건축공학(2,653), 금속공학(1,233), 기계공학(56,880), 물리학(11,901), 산업공학(755), 생물학(73,562), 수학(5,519), 의학(181,825), 전기전자공학(1,243), 전산학(3,157), 지구과학(7,338), 지리학(5,916), 토목공학(655), 화학(19,436), 화학공학(451), 환경공학(936)으로 구성되어 있음. 괄호 안은 기술용어 개수를 나타냄(분야 간 중복 허용).

의 연관관계를 파악하는 기능으로 문장의 구문 패턴을 이용하여 용어 쌍이 가지고 있는 관계를 인식한다. 패턴분석 이후에 용어 쌍이 연관관계를 가질 수 있는 형태이면 연관관계를 표현하는 디스크립터를 추출한다.

⑤ 연관관계 개념화: 워드넷과 같은 의미망을 활용하여 획득된 디스크립터의 개념을 추상화하고 이를 의미적으로 클러스터링하는 작업으로 의미 확장을 통해서 다양한 후보 연관관계를 생성한다. 다시 말해서 다양한 의미를 가질 수 있는 디스크립터를 의미망을 활용하여 각각 최상위 레벨의 의미로 추상화시킨 후에 해당 의미들을 기술용어 간의 후보 연관관계로 활용하는 것이다.

⑥ 후보 트리플 생성: 기술용어와 추상화된 다양한 후보 연관관계를 이용하여 후보 트리플을 생성하는 기능으로 후보 연관관계의 종류에 따라서 후보 트리플의 종류가 결정된다.

⑦ 최종 트리플 검증: 후보 트리플 집합의 적합성 판정을 통하여 최종 트리플을 결정하는 기능으로 테스트 컬렉션 구축자는 연관관계의 의미 정보 및 문맥 정보 등을 참조하여 해당 기술용어 간의 관계를 가장 잘 설명할 수 있는 연관관계를 트리플 결정 지원 도구를 사용하여 최종적으로 선택한다.

이와 같은 프레임워크를 이용하여 구축된 테스트 컬렉션의 품질을 보증받기 위해서는 다음의 세 가지 조건을 만족해야 한다. 먼저 기반 데이터로 대용량의 데이터베이스가 필요하다. 문헌 내에 존재하는 용어 및 관계 추출의 재현율을 높이고 결과의 변동성을 최소화하기 위해서는 대용량의 데이터를 가지고 문서 분석 작업을 수행할 필요가 있다. 다음으로 여러 분야를 포괄하는 전문용어 사전이 필요하다. 문헌에 존재하는 다양한 용어를 인식하기 위해서는 분야 정보를 다양화할 필요가 있다. 마지막으로 객관적인 적합성 판정을 유도할 수 있는 표준화된 처리 과정이 갖춰져야 한다. 구축할 때마다 편차가 심한 결과를 생성한다면 프레임워크의 신뢰성이 떨어지게 된다. 본 연구에서는 위의 세 가지 전제 조건을 만족시키기 위해서 첫 번째로 과학기술 전 분야에 걸친 30,858,830건의 해외 학술 데이터베이스(NDSL 데이터)를 사용하였고, 두 번째로 16개 분야 253,603건 규모의 전문용어 사전을 사용하였다. 그리고 세 번째로 각종 언어처리와 구문분석, 기계학습을 시스템적으로 처리하여 기술용어와 연관관계로 이루어진 후보 트리플 집합을 효과적으로 생성하였다. 동시에 트리플이 추출된 문맥 정보, 연관관계의 의미 정보, 의미에 맞게 사용된 예제문과 같은 정보를 추가적으로 제공하여 구축자가 연관관계를 결정하는 작업을 좀 더 쉽고 정확하게 수행하도록 하였다.

프레임워크를 이용한 테스트 컬렉션 구축은 정형화되고 시간이 많이 소요되는 작업을 자동화함으로써 구축에 들어가는 비용을 최소화한다. 이러한 구축 과정에서 품질 좋은 결과물을 생성하기 위해서는 자원이 처리되는 과정뿐만 아니라 그 자원 자체 또한 품질이 우수해야 한다. 본 논문에서는 위와 같이 체계적으로 갖추어진 문서 처리 과정과 질 좋은 기반 데이터를 사용해서 양질의 테스트 컬렉션을 생성하도록 노력하였다.

## IV. 프레임워크를 이용한 기술용어 간 관계 추출 테스트 컬렉션 구축

반자동 테스트 컬렉션 구축 방법은 문장 내의 디스크립터가 표현하는 개념을 의미망을 활용하여 상위 개념으로 일반화시키면서 다양성을 줄이고 집약성을 확보하는 절차를 거쳐서 테스트 컬렉션을 구축한다. 본 장에서는 실제적으로 프레임워크를 이용하여 후보 트리플을 생성하고 구축자가 최종적으로 연관관계를 선택하여 최종 트리플을 결정하는 과정에 대해서 설명한다.

### 1. 후보 트리플 집합 생성

테스트 컬렉션 구축 시 자동으로 처리되는 시스템적인 처리 과정의 결과는 서로 연관성을 가지는 기술용어 쌍과 이것들 사이의 후보 연관관계 집합이다. [표 1]은 이러한 예를 보여준다.

**표 1. 후보 연관관계가 제시된 기술용어 쌍의 예**

| 기술용어 | 후보 연관관계 | 기술용어 |
|---|---|---|
| interstitial_lung_disease | (keep, maintain, hold)<br>(persist, remain, stay)<br>(be) | tropical_pulmonary_eosinophilia |
| inner_limiting_membranes | (think, cogitate, cerebrate)<br>(act, move)<br>(examine, see)<br>(analyze, study, examine) | atomic_force_microscopy |
| innate_immune_responses | (make, create)<br>(act, move)<br>(trigger) | pattern_recognition_receptors |
| inhaled_nitric_oxide | (change, alter, modify)<br>(oppress, suppress, crush)<br>(inhibit, bottle_up, suppress)<br>(make, create)<br>(appoint, charge)<br>(have, have_got, hold)<br>(move, displace) | pulmonary_vascular_resistance |

[표 1]에서 볼 수 있듯이, 각 기술용어 쌍 별로 연관관계들이 복수로 지정된다. 따라서 테스트 컬렉션 구축 과정은 이러한 후보 연관관계 중에서 가장 적절한 관계를 지정하는 작업으로 정의될 수 있다. 이때 구축자는 두 기술용어 간의 관계를 보다 세밀하게 분석하기 위해서 기술용어 포함 문장들을 참고하게 된다. 본 연구에서는 기술용어로서의 전문성 정도가 비교적 강한 3단어 이상으로 구성된 용어 집합을 대상으로 테스트 컬렉션을 구축하였다. 본 연구에서 사용된 3단어 이상 기술용어 쌍은 총 6,144개이며, [그림 2]와 같이 각 쌍마다 용어가 표시된 참고 문장이 함께 제공된다.

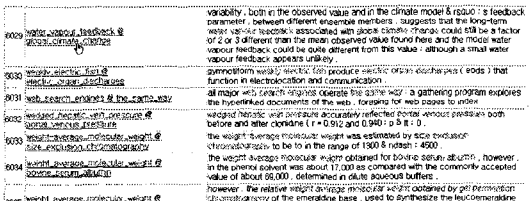



**그림 2. 기술용어 쌍과 대응 참고 문장 예시**

## 2. 트리플 적합성 판정 방법 및 절차

[표 1]에서 살펴본 바와 같이, 트리플은 기술용어의 쌍과 관련된 연관관계의 후보 집합으로 이루어진다. [표 2]는 이 집합으로부터 트리플 유효성을 판정하기 위해서 기술용어에 대한 분석과 후보 연관관계에 대한 최종 연관관계를 선택하는 작업을 정의하고 있다.

**표 2. 기술용어 간의 연관관계 설정**

| 관계 구분 | 내용 |
|---|---|
| 관계설정 불가 | - 두 기술용어가 S+V+O 형태를 구성하지 못함<br>- 기술용어의 전문성이 결여됨 |
| 관계를 찾지 못함 | - 두 기술용어가 S+V+O 형태를 구성하고 있으나 후보 연관관계가 적절하지 못한 경우 |
| 관계설정 성공 | - 두 기술용어가 S+V+O 형태이고, 후보 연관관계가 적절한 경우<br>- 연관관계에 대한 "수동", "능동"에 대한 구분<br>- 연관관계에 대한 "긍정", "부정"에 대한 구분 |

두 기술용어 사이에 나타나는 디스크립터(동사/동사구)를 분석하여 발생할 수 있는 "관계 구분"은 [표 2]에서와 같이 "관계설정 불가", "관계를 찾지 못함", "관계설정 성공"의 3가지 형태로 나누어진다.

"관계설정 불가"는 두 기술용어와 연관관계가 S+V+O형태로 이루어 지지 않는 경우(보기1)와, 두 기술용어 중 하나라도 기술용어로 판단되지 않는 경우(보기2)가 있다.

보기1) between may 1988 and november 1996 , 28 patients with **cervical spinal cord injury** underwent **lower urinary tract** reconstruction .
보기2) the ***capillary filtration coefficient*** *showed* ***no significant change*** over time , but decreased by 5-10 % following the albumin infusion .

보기1)에서 기술용어로 인식된 "cervical spinal cord injury"와 "lower urinary tract"는 두 용어 사이에 위치한 연관관계 "underwent"와 관련이 없고, underwent(V)는 patients(S)와 lower urinary tract(O)에 대한 연관관계로 판단되므로 "관계설정 불가"로 처리된다. 보기2)는 두 개의 기술용어가 S+V+O를 구성하고 있지만 두 번째 기술용어로 인식된 "no significant change"가 전문성이 결여되는 것으로 판단되어 "관계설정 불가"로 처리된다.

"관계를 찾지 못함"은 두 기술용어와 연관관계가 S+V+O형태로 구성되어 있지만, 후보 연관관계에서 적절한 연관관계를 찾지 못하는 경우(보기3)이다. 여기에 해당되는 연관관계는 추후 새로 정의할 필요가 있지만 현 단계에서는 미분류로 가정한다.

보기3) **body mass index** *was positively associated with* **systolic blood pressure** in both groups .
후보연관관계군 : 1. act,move 2. think,cognitive,cerebrate 3.join,fall_in,get_together

"관계설정 성공"은 두 기술어와 연관관계가 S+V+O 형태를 구성하고 후보 연관관계에서 적절한 연관관계를 선택할 수 있는 경우로서, 이때에는 두 용어 간의 관계가 "능동"인지 "수동"인지와 "긍정"인지 "부정"인지에 대한 설정도 함께 수행한다. "능동", "수동"에 대한 판단을 수행할 때에는 연관관계와 두 기술용어의 위치를 파악하여 신중하게 결정해야 한다. 보기4)는 "central nervous system @ the immune system"의 두 기술용어와 관련된 문장의 예를 보여준다. 이 문장에서 인식된 디스크립터는 "affect"로서 문장에서는 "능동"으로 사용되었지만 제시된 기술용어 쌍의 순서와 거꾸로 되어있기 때문에 연관관계는 "수동"으로 설정해야 한다.

보기4) evidence is presented that **the immune system** *can affect* **central nervous system** functioning , leading to changes in learning .

테스트 컬렉션 구축을 위해 자동으로 생성된 트리플 중에서 2,800건을 선정하여 구축자에게 할당하였고 관계 설정 작업을 돕기 위한 트리플 결정 지원 도구를 제공하여 전체적인 작업 시간을 단축하였다.

[그림 3]은 트리플 결정 지원 도구의 관계 설정 기능을 이용하여 "bone mineral density"와 "quantitative computed tomography"의 두 기술용어를 입력하고 검색한 화면이다. 구축자는 기술용어를 포함하는 문장들을 살펴보고 관계설정이 가능한지를 우선 결정한다. 관계설정이 가능하다고 판단되면 후보 집합으로 설정된 관계 중에서 가장 적합하다고 판단되는 관계를 선택하고 "저장"한다. 트리플 결정 지원 도구는 구축자가 관계를 결정할 때 기술용어에 대한 의미를 쉽게 파악할 수 있도록 기술용어에 대한 한글 대역어에 대한 검색화면도 함께 제공한다. [그림 3]의 상단에 있는 "bone_mineral_density @ quantitative_computed_tomography" 링크를 클릭하면 [그림 4]와 같이 두 기술용어에 대한 한글 대역어를 보여 준다.

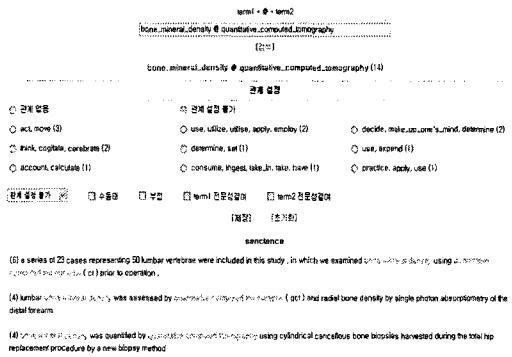



**그림 3. 관계 설정을 위한 웹 페이지**

| bone_mineral_density | quantitative_computed_tomography |
|---|---|
| (의학) 골무기질밀도 | (의학) 정량의 컴퓨터 단층촬영법 |
| (의학) 골미네랄밀도 | (의학) 정량적전산화단층촬영법 |
| (의학) 골수 억제 | (의학) 정량적전산화단층촬영술 |
| (의학) 골염량 | |
| (의학) 골전해질 밀도 | |
| (의학) 베커근육 발육미상 | |

**그림 4. 기술용어에 대한 한글 대역어 검색 화면**

## 3. 구축 완료된 테스트 컬렉션 분석

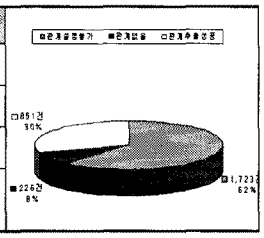
트리플 결정 지원 도구를 이용하여 구축된 테스트 컬렉션은 [표 3]과 같은 형식으로 텍스트 파일로 저장된다.

**표 3. 테스트 컬렉션 형식**

| Term1 @ Term2 | 연관관계 | Term1 전문성 결여 | Term2 전문성 결여 | 수동태 | 부정 | sentence |
|---|---|---|---|---|---|---|
| blood_pressure_measurements @ no_significant_change | 관계설정 불가 | 0 | 1 | 0 | 0 | .... |
| ow_alloy_steel @ direct_reduced_iron | produce, make,create | 0 | 0 | 1 | 0 | .... |

**표 4. 관계설정 성공률 현황**

| 관계 | 건수 |
|---|---|
| 관계설정 불가 | 1,723건(62%) |
| 관계없음 | 226건(8%) |
| 관계설정 성공 | 851건(30%) |
| 전체 | 2,800건 |




[표 4]는 전체 2,800건의 관계 쌍으로부터 설정된 연관관계에 대한 현황을 보여준다. 전체 데이터에서 "관계설정 불가"는 62%(1,723건), "관계없음"은 8%(226건), 그리고 "관계설정 성공"은 30%(851건)를 차지하였다.

**표 5. 연관관계 분석 현황**

| 관계 | 건수 | 백분율(%) |
|---|---|---|
| use,utilize,utilise,apply,employ | 121 | 14% |
| change,alter,modify | 62 | 7% |
| induce,stimulate,cause,... | 61 | 7% |
| make,create | 59 | 7% |
| think,cogitate,cerebrate | 44 | 5% |
| analyze,analyse,study,... | 36 | 4% |
| get,acquire | 30 | 4% |
| include | 25 | 3% |
| 기타 | 413 | 49% |

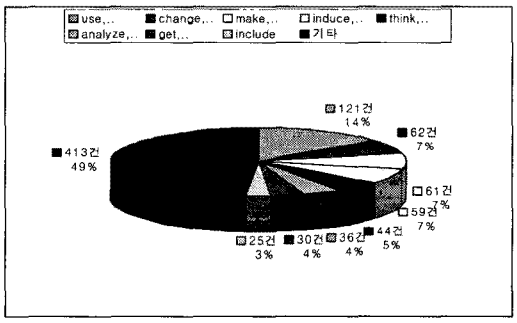



**그림 5. 연관관계 분석 그래프**

[표 5]와 [그림 5]는 "관계설정 성공"으로 선택된 연관관계에서 상위 8개의 관계와 나머지 관계에 대한 건수와 백분율을 보여준다. 전체적으로 106개의 "연관관계"가 설정되었는데 이 중에서 가장 많이 출현한 관계는 "use, utilize, utilise, apply, emply" 관계로서 전체 851회 중 121회로 14%를 차지하였고, 이를 포함한 상위 8개의 연관관계가 전체 연관관계의 50%이상을 차지하였다. 끝으로 한 번만 출현한 연관관계는 40개로 전체 연관관계에서 37.7%를 차지하고 있지만 실제 건수는 4.7%에 불과하였고, 10회 미만 출현한 연관관계는 84개로 전체 연관관계의 79.2%를 차지하고 있지만 실제 건수는 212건으로 24.9%만을 차지하였다.

## V. 결론 및 향후 연구

본 논문에서는 워드넷과 같은 언어자원을 기반으로 한 반자동 관계 추출 테스트 컬렉션 구축 프레임워크에 대해서 설명하였다. 그리고 개발된 프레임워크를 적용하여 과학기술 문헌에 존재하는 기술용어 간 연관관계 추출 시스템의 성능 평가를 수행하는 테스트 컬렉션을 실제적으로 구축하고 결과를 분석해보았다. 구축된 테스트 컬렉션은 관계유무 파악 및 관계분류 식별을 검사할 수 있는 1,707건의 문장 규모로 구성되어 있다. HPRD50[11], IEPA[12], LLL[13]에서 구축한 문장 개수가 각각 145, 486, 77인 것을 감안하면 그 규모가 비교적 크다고 할 수 있다. 또한 자동화된 프로세스를 기반으로 테스트 컬렉션을 구축하기 때문에 추가적으로 규모를 증가시키는 작업도 기존의 방법론에 비해 훨씬 수월하다.

본 논문에서 제시한 테스트 컬렉션 구축 프레임워크의 장점은 정형화되고 시간이 많이 소요되는 문서분석 작업을 처리과정별로 자동화함으로써 구축에 들어가는 비용을 최소화할 수 있고 시스템의 알고리즘을 기반으로 동작하기 때문에 구축자의 성향에 따른 편차를 줄이고 일관된 결과물을 얻을 수 있다는 것이다. 또한 대규모의 학술 데이터베이스(과학기술 전 분야에 걸친 30,858,830건의 NDSL 데이터)와 다양한 분야의 전문용어 사전(16개 분야의 253,603건의 전문용어), 그리고 CRF와 같은 최신의 기계학습 알고리즘을 활용하여 특정 분야에 편중되지 않은 균형 잡히고 객관화된 테스트 컬렉션을 구축할 수 있다는 것이다.

향후 연구로서, 우선 시급한 과제는 후보 연관관계를 제시할 때 선택될 가능성이 가장 높은 후보부터 순차적으로 리스트를 제공하는 것이다. 구축자가 해당 트리플 및 트리플이 포함된 문장의 문맥을 보고 가장 적합한 연관관계를 검사할 때, 유사성이 높은 순서대로 리스트를 제공한다면 연관관계 선택 작업을 좀 더 효과적으로 수행할 수 있다. 따라서 트리플을 포함하고 있는 해당 문맥정보와 후보로 제시된 연관관계 리스트 사이의 의미 유사도를 계산하여 순위화된 리스트를 제공할 필요가 있다. 다음으로 구축된 테스트 컬렉션의 정밀한 검증을 통해서 기존의 구축 방법과의 효용성을 비교·분석하는 작업이 필요하다. 또한 테스트 컬렉션 구축 프레임워크의 다양한 필드 적용을 통한 시스템 안정화 및 기능 개선을 통해서 프레임워크의 신뢰성을 향상시킬 필요가 있다. 끝으로 현재는 3단어 이상의 기술용어를 대상으로 트리플을 생성했는데 앞으로는 용어에 대한 전문성 측정 기준을 마련하여 전문성 정도가 높게 나타나는 한 단어 이상의 모든 기술용어를 대상으로 적용하는 방법에 대한 연구가 필요하다.

## 참 고 문 헌


[1] 맹성현, 이석훈, 이준호, 이응봉, 송사광, "정보 검색 시스템 평가를 위한 균형 테스트 컬렉션 구축", 정보관리학회지, Vol.16, No.2, pp.135-148, 1999.

[2] L. Jimmy and K. Boris, "Building a Reusable Test Collection for Question Answering," Journal of the American Society for Information Science and Technology, Vol.57, No.7, pp.851-861, 2006.

[3] K. Gabriella, L. Mounia, and R. Jane, "Construction of a Test Collection for the Focussed Retrieval of Structured Documents," ECIR 2003, pp.88-103, 2003.

[4] 이경순, 김재호, 최기선, "질의응답시스템의 성능 평가를 위한 테스트컬렉션 구축", 한글 및 한국어 정보처리 학술대회, pp.190-197, 2000.

[5] 이준호, 최광남, 한현숙, 김종원, 남성원, "정보 검색 연구를 위한 KRIST 테스트 컬렉션의 개발", 정보관리학회지, Vol.12, No.2, pp.225-232, 1995.

[6] 김지영, 장동현, 맹성현, 이석훈, 서정현, 김현, "한국어 테스트 컬렉션 HANTEC의 확장 및 보완", 한글 및 한국어 정보처리 학술대회, pp.210-215, 2000.

[7] 김성혁, 서은경, 이원규, 김명철, 김영환, 김재군, "자동색인기 성능시험을 위한 Test Set 개발", 정보관리학회지, Vol.11, No.1, pp.81-102, 1994.

[8] http://ldc.upenn.edu/Projects/ACE

[9] http://www-nlpir.nist.gov/related_projects/muc

[10] http://www.ldc.upenn.edu

[11] K. Fundel, R. Kuffner, and R. Zimmer, "RelEx - Relation extraction using dependency parse trees," Bioinformatics, Vol.23, pp.365-371, 2007.

[12] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele, "Mining MEDLINE: abstracts, sentences, or phrases?," Proceedings of PSB'02, pp.326-337, 2002.

[13] C. Nedellec, "Learning language in logic - genic interaction extraction challenge," Proceedings of LLL'05, pp.31-37, 2005.

[14] A. M. George, "WordNet: A Lexical Database for English," COMMUNICATIONS OF THE ACM, Vol.38, No.11, pp.39-41, 1995.

[15] 배영준, 김재훈, 옥철영, 최윤수, "CRF를 이용한 생물/의학 전문용어 인식", 제21회 한글 및 한국어 정보처리 학술대회, pp.87-91, 2009.

[16] 김형철, 김재훈, 최윤수, "접사 정보를 이용한 영어 미등록어의 품사부착 성능개선", 제21회 한글 및 한국어 정보처리 학술대회, pp.186-190, 2009.

[17] 김형철, 서형원, 김재훈, 최윤수, "CRF를 이용한 대명사 참조해소 시스템", 제21회 한글 및 한국어 정보처리 학술대회, pp.197-201. 2009.

## 저 자 소 개


**정 창 후(Chang-Hoo Jeong)** 정회원

- 1999년 : 충남대학교 컴퓨터과학과 졸업(학사)
- 2002년 : 충남대학교 대학원 컴퓨터과학과 졸업(석사)
- 2003년 ~ 현재 : 한국과학기술정보연구원 정보기술연구실

<관심분야> : 정보검색 및 추출, 분산 데이터마이닝

**최 성 필(Sung-Pil Choi)** 정회원

- 1996년 : 부산대학교 전자계산학과 졸업(학사)
- 1998년 : 부산대학교 대학원 전자계산학과 졸업(석사)
- 2009년 : 한국과학기술원 대학원 정보통신공학과(박사 수료)
- 1998년 ~ 현재 : 한국과학기술정보연구원 정보기술연구실

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

**이 민 호(Min-Ho Lee)** 정회원

- 1998년 : 충남대학교 컴퓨터과학과 졸업(학사)
- 2000년 : 충남대학교 대학원 컴퓨터과학과 졸업(석사)
- 2006년 : 충남대학교 대학원 컴퓨터공학과(박사수료)
- 2000년 ~ 2001년 : 데이콤 중앙연구소 차세대인터넷 개발팀
- 2001년 ~ 현재 : 한국과학기술정보연구원 정보기술연구실

<관심분야> : 정보검색 및 추출, 정보보호, 분산시스템

**최 윤 수(Yun-Soo Choi)** 정회원

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

<관심분야> : 데이터베이스, 정보검색