과학기술분야 용어 간 관계추출 시스템의 평가를 위한 테스트컬렉션 구축

Construction of Test Collection for Evaluation of Scientific Relation Extraction System

최윤수, 최성필, 정창후, 윤화묵, 류범종
한국과학기술정보연구원

Yun-Soo Choi, Sung-Pil Choi, Chang-Hoo Jeong, Hwa-Mook Yoon, Beom-Jong You
Korea Institute of Science and Technology Information

요약

대용량 문서에서 포함된 정보를 추출하는 작업은 정보검색분야 뿐만 아니라 질의응답과 요약분야에서 매우 유용하다. 정보추출 분야 중 관계추출 기술이 중요하게 인식되고 있으나, 기계학습모델을 기반으로 개발하기 위한 학습집합과 개발된 기술을 평가하기 위한 평가집합의 부재로 연구에 난항을 겪고 있다. 본 논문은 한국과학기술정보연구원(KISTI)이 보유하고 있는 해외학술지 데이터를 기반으로 과학기술용어에 대한 관계추출 기술 시스템을 개발하고 평가하기 위한 테스트 컬렉션(KREC2008) 구축을 위한 구축방법 및 절차를 기술한다. 해외학술지 데이터의 초록을 대상으로 기술용어를 추출하였고, 기술용어의 쌍의 관계에 해당되는 단어를 Wordnet에 매핑하여 동사의 개념을 일반화하는 여러 개의 개념화된 후보군을 추출하였다. 평가기준 및 절차 교육이 이루어진 평가자가 개념화된 후보군에서 적합하다고 판단되는 "개념"을 "관계"로 지정하였다. Wordnet을 이용하여 "관계"에 대한 후보군을 생성하였기 때문에, 일관성 있는 관계설정의 품질의 향상시켰고 비전문가도 쉽게 테스트컬렉션을 구축할 수 있는 방법을 제공하였다. 현재 KREC2008은 정보추출 연구자 및 개발자에게 공개되어 있으며, 과학기술분야 관계추출 시스템의 개발 및 신뢰도 평가를 목적으로 하는 학술대회의 연구결과 발표 및 제품 비교 등에 활용될 예정이다.

Abstract

Extracting information in large-scale documents would be very useful not only for information retrieval but also for question answering and summarization. Even though relation extraction is very important area, it is difficult to develop and evaluate a machine learning based system without test collection. The study shows how to build test collection(KREC2008) for the relation extraction system. We extracted technology terms from abstracts of journals and selected several relation candidates between them using Wordnet. Judges who were well trained in evaluation process assigned a relation from candidates. The process provides the method with which even non-experts are able to build test collection easily. KREC2008 are open to the public for researchers and developers and will be utilized for development and evaluation of relation extraction system.

Ⅰ. 서론

인터넷의 발달과 더불어 대용량 데이터를 실시간으로 처리하여 필요한 지식을 발견하기 위한 정보추출 기술들이 핵심적인 분야로 인식되고 있다. 정보추출은 크게 1) 대용어 참조해소(coreference resolution), (2) 개체명 인식(named-entity recognition), (3) 관계 추출(relation extraction)의 세 가지 요소기술로 세분화 되며, 이 중 관계추출 분야는 아직까지 가장 해결하기 어

려운 분야로 남아있다.[1,2,3]

국제적으로 MUC (Message Understanding Conference), ACE (Automatic Content Extraction)에서 기계학습기법을 이용한 지도학습 기반 관계 추출 (supervised relation extraction) 대한 지원이 있어왔다.

1997년도에 개최된 MUC-7에서 처음으로 도입된 '템플릿 기반 관계 추출 (Template Relation Extraction)' 은 태스크에서 본격적으로 기계학습 기반의 관계추출을 위한 학습 집합을 제공하였다. [4].

MUC의 성공적인 연구결과에 고무된 NIST와 DARPA[1]는 본격적으로 보다 고차원적인 정보추출 기법을 위한 기반 인프라 구축을 시도하였으며, 그 결과 ACE 검증 컬렉션이 매년마다 구축되고, 이를 기반으로 많은 연구진들의 연구결과를 바탕으로 워크숍을 개최하고 있다.[5]

그러나, MUC, ACE에서 제공하는 테스트컬렉션은 신문기사, 뉴스 등으로 한정되어 있고, 구입을 위해 많은 비용을 지불해야 한다.

본 연구에서는 KISTI가 보유하고 있는 과학기술문헌을 대상으로 관계추출 시스템을 개발하기 위한 테스트컬렉션 구축을 위한 방법을 제시한다.

Ⅱ. 테스트컬렉션 구축 절차

과학기술 전체분야에 대한 통계정보를 얻기 위해 KISTI가 보유하고 있는 해외학술데이터베이스 30,858,830 건을 대상데이터로 선정했다.

테스트컬렉션 구축은 1) 용어자동추출, 2) 관계후보자동추출, 3) 관계수동지정의 세 단계로 이루어 진다.

1. 기술용어 추출

전체 학술 데이터베이스 30,858,830건에서 관계추출을 위한 자질추출 및 문장추출 작업의 특성상 초록이 포함된 12,666,438(42.9%)건의 서지문헌만을 대상으로 작업을 수행하였다.

원본 데이터베이스를 가공하여 16개 분야 253,603건 규모의 기술용어사전[2] 탐색 및 매칭을 시도한다. 이 과정에서 어휘변형을 해소하고 복합어 처리를 위한 다양한 특수 규칙이나 알고리즘을 사용하여 추출되는 기술용어들의 범위를 확장하였다.

2. 관계후보 추출

기술용어를 포함하고 있는 문장에서 [표 1]에 표현된 세 가지 기본 유형의 기술용어 포함 문장을 추출하였다.

표 1. 기본 유형 문장 추출 결과

두 개의 기술용어가 포함된 기본 유형	문장 개수
기술용어+*동사구*+**기술용어** (NP) (VP) (NP)	2,752,193
기술용어+*동사구*+*전치사*+**기술용어** (NP) (VP) (PP) (NP)	3,646,484
기술용어+*동사구*+*부사*+*전치사*+**기술용어** (NP) (VP) (ADJP) (PP) (NP)	111,740

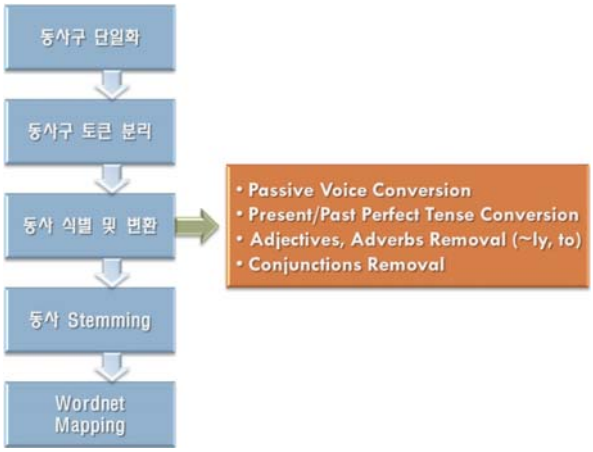
본 연구에서는 위의 세 가지 유형 가운데 가장 단순한 형태인 첫 번째 유형의 문장들에 대해서 후보연관관계 추출을 수행하였다. 첫 번째 유형의 문장들에 대해서 우선적으로 작업을 수행한 근거는, [6]에서 제시하였듯이, 이진 관계를 나타내는 문장집합에 대해서 수작업으로 그 구조를 분석한 결과 약 40% 정도가 첫 번째 유형의 문장구조로 표현되었기 때문이다. 이러한 결과를 바탕으로 두 개의 기술용어들 사이에서 다양하게 표현된 동사구들을 단일화, 정규화하여 이를 워드넷에 사상시키는 작업을 수행하였다. 이를 위한 세부적인 프로세스는 [그림 1]과 같다.

[그림 1]에서 보듯이 동사구 개념화 단계는 총 다섯 개의 세부 프로세스로 구성된다. 동사구 단일화 단계는 반복해서 나타나는 동사구에 대한 단순 단일화 작업을 의미한다. 동사구 토큰 분리 작업은 "has been moved", "was executed"와 같이 다중 어절로 구성된 동사구에 대한 토큰 분리 작업이다. 세 번째 단계인 동사 식별 및 변환 단계에서는 (1) 수동태로 표현된 동사

1) http://www.darpa.mil/

2) 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) : 괄호 안은 기술용어개수 (분야간 중복허용)

에 대한 능동태 변환, (2) 현재, 과거완료형 동사구 변환, (3) Chunking 오류나 품사 태깅 오류로 인한 형용사 및 부사 포함 동사구 필터링, (4) 접속사 제거 등과 같은 필터링 단계를 거치게 된다. 마지막으로 실질적인 워드넷 매핑은 MIT에서 개발한 Java Wordnet Interface (JWI 2.1.4)[3]를 활용하였다.



▶▶ 그림 1. 동사구 개념화 세부 단계

워드넷을 구성하는 synset 집합들은 서로 다양한 관계들로 연결되어 있다. 본 연구에서는 특정 동사에 대한 synset 매핑을 시도할 때 되도록 포괄적인 개념의 synset과 연결시키기 위해서 hypernym 관계를 활용하여, 상위 개념으로의 자동 전이 기반 개념매핑 기법을 적용하였다.

3. 테스트컬렉션 구축

각 기술용어 쌍 별로 [그림 1]에서 제시된 동사구 개념화 기법에 의해 생성된 연관관계들이 [표2]처럼 복수로 지정된다. 테스트 컬렉션 구축 과정은 이러한 후보 연관관계 중에서 가장 적절한 관계를 지정하는 작업으로 정의될 수 있다. 이 때, 구축자는 두 기술용어 간의 관계를 보다 세밀하게 분석하기 위해서 [표 1]의 첫 번째 패턴으로 이루어진 기술용어 포함 문장들을 참고하게 된다.

본 연구에서는 기술용어로서의 전문성 정도가 비교적 강한 3단어 이상으로 구성된 용어 집합을 대상으로 테스트 컬렉션을 구축하였다. 3단어 이상 기술용어 쌍은 총 6,144개이며 각 쌍마다 용어 태깅된 참고 문장이 지정되어 있다.

표 2. 후보 연관관계가 부착된 기술용어 쌍의 예

기술용어쌍	후보 연관관계
inner_limiting_membranes atomic_force_microscopy	1. (think, cogitate, cerebrate) 2. (act, move) 3. (examine, see) 4. (analyze, study, examine)

[표3]은 전문용어의 쌍과 관련연관관계 후보집합으로부터 테스트 컬렉션을 구축하기 위해 기술용어에 대한 분석과 연관관계에 대한 관계를 추출하는 작업을 정의하고 있다.

표 3. 전문용어간의 연관관계 추출

관계구분	내용
관계설정 불가	- 두 전문용어가 S+V+O 형태를 구성하지 못함 - 전문용어의 전문성이 결여됨
관계를 찾지 못함	- 두 전문용어가 S+V+O 형태를 구성하고 있으나 후보 연관관계가 적절하지 못한 경우
관계추출 성공	- 두 전문용어가 S+V+O 형태이고, 후보 연관관계가 적절한 경우 - 연관관계에 대한 “수동”, “능동”에 대한 구분 - 연관관계에 대한 “긍정”, “부정”에 대한 구분

두 전문용어 사이에 나타나는 연관관계(동사/동사구)를 분석하여 발생할 수 있는 “관계구분”은 [표 2]에서와 같이 “관계설정불가”, “관계를 찾지 못함”, “관계추출 성공”의 3가지 형태로 나누어진다.

두 전문용어가 S+V+O의 형태이고, 후보연관관계가 적합한 경우에는 “관계추출 성공”으로, 적합하지 않은 경우에는 “관계를 찾지 못함”으로 정의한다. 여기에 해당하는 연관관계는 추후 새로 정의할 필요가 있지만 현 단계에서는 미분류로 가정하였다.

테스트컬렉션 구축을 위해 후보집합에서 2,800건을 선정하여 비전문가 7명에게 각각 400건씩 할당하였고, 관계추출 작업을 지원하기 위한 사용자 인터페이스를 구성하여 작업시간을 단축하고 작업결과를 표준화하였다.

3) JWI 2.1.4, http://www.mit.edu/~markaf/projects/wordnet/

표 4. 테스트컬렉션 형식

Term1 @ Term2	관계추출	term1 전문성결여	term2 전문성결여	수동	부정	문장
blood_pressure_measurements @ no_significant_change	관계설정불가	0	1	0	0	....
ow_alloy_steel @ direct_reduced_iron	produce,make,create	0	0	1	0	....

관계추출 웹페이지는 구축자가 관계를 결정할 때 전문용어에 대한 의미를 쉽게 파악할 수 있도록 전문용어에 대한 한글대역어에 검색화면도 함께 지원하였다.

Ⅲ. 테스트컬렉션 분석

제공된 사용자 인터페이스를 이용하여 구축된 테스트 컬렉션은 [표 4]와 같은 형식으로 텍스트 파일로 저장된다.

표 5. 관계추출 성공률 현황

관계	건수
관계설정불가	1,723건(62%)
관계없음	226건(8%)
관계추출성공	851건(30%)
전체	2,800건

[표 5]는 전체 2,800건의 관계쌍으로부터 추출된 관계에 대한 현황을 보여 준다. 전체 데이터에서 "관계설정불가" 는 1,723건으로 62%정도의 가장 많은 부분을 차지하고, "관계없음" 은 226건(8%), "관계추출성공 "은 851건으로 전체 데이터 중 약 30%정도의 관계쌍에 대한 연관관계가 정의되었다.

[표 6]은 "관계추출성공" 으로 선택된 연관관계에서 상위 8개의 관계와 나머지 관계에 대한건수와 백분율을 보여 준다. 전체 106개의 "연관관계" 가 추출되었고, 가장 많이 출현한 관계는 "use,utilize,utilise,apply,emply" 관계로써 전체 851회 중121회로 14%를 차지하였고, 이를 포함한 상위 8개의 연관관계가 전체 연관관계의 50%이상을 차지한다. 단 한번만 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 차지하지만 해당 건수는 212건으로 24.9%만을 차지하였다.

표 6. 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	468	49%

Ⅳ. 결론 및 향후 연구

본 연구에서는 전문용어간 관계 추출 시스템의 평가를 위한 테스트컬렉션을 구축하였다. KISIT가 보유하고 있는 과학기술분야 해외학술데이터 30,858,830건에서 초록을 포함하고 있는 12,666,438을 대상으로 전문용어 추출작업을 수행하고, 3어절 이상의 전문용어를 포함하고 있는 문장 중에서 2,800건을 구축대상으로 하였다.

과학기술 전체분야에 대한 전문용어간의 관계를 정의하기 위하여 Wordnet을 이용하여 추출된 후보관계를 제공했다.

일반적으로 테스트컬렉션은 분야전문가에 의해 구축되지만, 본 연구에서는 과학기술전체분야를 대상으로 하였기 때문에, 특정분야전문가를 이용할 수 없었고, 비

전문가가 테스트컬렉션을 구축하기 위해 후보관계를 이용하였다.

특정분야가 아닌 일반적인 분야에 대한 테스트컬렉션 구축작업이나, 관계가 설정되어 있지 않은 분야에서 어떠한 관계들이 존재하는지 알아내기 위해서는 본 연구에서 수행한 Wordnet을 이용하여 관계후보를 제공하는 방법이 유용하다.

본 연구에서 구축된 테스트컬렉션의 품질을 검증하기 위해서는 구축된 테스트컬렉션의 분야별로 분야전문가에게 타당성을 검수하는 작업이 필요하다. 이는 향후과제로 추진될 예정이다.

■ 참 고 문 헌 ■

[1] Bunescu, R. C., Mooney, R. J., "A Shortest Path Dependency Kernel for Relation Extraction", Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, B.C., pp.724-731, 2005.

[2] Culotta, A., Sorensen, J., "Dependency Tree Kernels for Relation Extraction, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics", 2004.

[3] Bunescu, R. C., Mooney, R. J., "Subsequence Kernels for Relation Extraction", Advances in Neural Information Processing Systems, 2006.

[4] MUC. 1987-1998, Message Understanding Conference,(http://www-nlpir.nist.gov/related_projects/muc/)

[5] ACE. 2002-2005, Automatic Content Extraction, http://ldc.upenn.edu/Projects/ACE/

[6] M. Banko and O. Etzioni. "The Tradeoffs Between Open and Traditional Relation Extraction", In Proceedings of ACL 2008.

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

[8] 이준호,최광남,한현숙,김종원,남성원, "정보검색 연구를 위한 KRIST 테스트 컬렉션의 개발", 정보관리학회지, 제12권, 제2호, pp.225-232, 1995.

[9] 맹성현,이석훈,이준호,이응봉,송사광, "정보 검색 시스템 평가를 위한 균형 테스트 컬렉션 구축", 한국정보과학회(제12회), pp210-215, 2000.

[10] 이경순,김재호,최기선, "질의응답시스템의 성능평가를 위한 테스트컬렉션 구축", 한국정보과학회(제12회), pp.190-197, 2000.

[11] Ellen M. Voorhees, Dawn M.Tice, "Building a Question Answering Test Collection", Research and development in information retrieval:SIGIR, pp.200-207, 2000.

[12] Atsushi Fujii, Katunobu Itou, "Building a Test Collection for Speech-Driven Web Retrieval", Speech communication and technology: EUROSPEECH, pp.1153-1156, 2003.