기술 문헌 분석 테스트베드 툴킷 개발

Developing a Test-Bed Toolkit for Scientific Document Analysis

최성필, 송사광, 정한민
한국과학기술정보연구원 소프트웨어연구실

Sung-Pil Choi(spchoi@kisti.re.kr), Sa-kwang Song(smallj@kisti.re.kr),
Hanmin Jung(jhm@kisti.re.kr)

요약

본 논문은 논문, 특허, 연구보고서 등과 같은 다양한 과학 기술 문헌에 포함된 기술 지식을 효과적으로 추출하는데 필요한 텍스트 분석 엔진들의 효과적인 모니터링 및 성능 최적화를 위한 테스트베드 도구를 소개한다. 이 도구는 과학 기술 분야의 전문 용어를 비롯한 인명, 지명, 기관명 등을 자동으로 인식하는 기술 개체 인식 엔진을 위한 테스트베드와 인식된 기술 개체 간의 의미적 연관 관계를 자동으로 추출하는 기술개체 간 관계 추출 테스트베드로 구성되어 있다. 이를 활용함으로써 사용자 및 개발자들은 기술 문헌 분석 엔진의 실행 모니터링은 물론 오류 분석을 효율적으로 수행할 수 있다.

■ 중심어 : | 문헌분석 | 테스트베드 | 자연어처리 | 텍스트마이닝 | 기술지식 |

Abstract

This paper introduces a test-bed toolkit for evaluating and enhancing text analysis engines which extract technological knowledge from articles, patents, reports and so forth. The toolkit consists of two test-beds for technical entity recognition and relation extraction engines, which are capable of identifying technical entities and predicting semantic relation types between the entities. With using the introduced toolkits, users and developers can efficiently perform the execution monitoring and error analysis of the technical text analysis engines.

■ keyword : | Text Analysis | Test Bed | Natural Language Processing | Text Mining | Technology Intelligence |

I. 서론

전 세계적으로 급속하게 생산되고 있는 과학기술분야 학술정보에 대한 효과적 접근은 현재 자연어 처리 및 텍스트 마이닝 분야의 가장 중요한 과제로서 대두되고 있다[1][2]. 다양한 정의나 관점이 있을 수 있으나, 자연어 처리 및 텍스트 마이닝의 가장 중요한 목적 중의 하나는 인간이 언어를 이용하여 텍스트로 표현한 학술 정보를 기반으로 이를 기계적이고 자동화된 공정을

통해서 중요한 핵심 정보를 추출하고 이를 조직화하는 것이라고 할 수 있다.

[1]에서는 학술 문헌에 대한 심층 분석에 대한 기본적인 개념을 확립하였다. 다시 말해서, 독자가 특정 학술 문헌을 읽어 나갈 때, 그 문헌에 나타나는 핵심 용어와 그 용어 간의 관계를 우선적으로 파악한다는 가정 하에서, 이를 자동으로 처리할 수 있는 전반적인 기술을 학술 문헌 분석이라고 정의하였다. 또한 이와 같은 학술 문헌 분석을 수행하기 위해서는 전문용어와 같은

기술개체들을 문헌에서 자동으로 인식할 수 있는 기술 개체 지동인식 엔진과, 인식된 기술 개체간의 관계를 자동으로 추출할 수 있는 기술개체 간 관계추출 엔진이 필요하다고 기술하고 있다[1]. 아래 표는 위에서 기술한 학술 문헌에 출현하는 기술 개체(technology entity)의 종류를 나열하였다.

표 1. 기술 개체 종류 및 설명

기술개체유형	설명
인명	과학기술 연구 종사자, 기술 사업화 종사자, 정책 결정권자 등
지명	국가, 도시, 물리적 위치 등
기관명	대학교, 연구센터, 연구소, 국가기관 등
기술용어	전문용어, 분야용어 등
시간표현	날짜, 연도, 시간 표현 등

[표 1]에서 기술한 개체 유형을 추출하고 이들 간의 의미적 연관 관계를 식별함으로써 최종적으로 도출되는 지식은 기술개체가 주어(subject) 및 목적어(object)가 되고 그 가운데에 관계(relation, predicate)가 연결되어 있는 시맨틱 트리플(semantic triple) 형태로 표현된다.

이러한 시맨틱 트리플을 생성하기 위해 필요한 언어 분석 엔진들은 일반적으로 그 복잡성으로 인해서 분석 진행 과정에 대한 모니터링이 쉽지 않다. 또한 다양한 형태로 나타날 수 있는 오류에 대한 체계적인 분석 또한 힘든 실정이다.

본 논문에서는 학술 문헌 분석을 위해서 개발된 기술 개체 인식 엔진[3]과 기술 개체 간 관계 추출 엔진[4]의 실행 상황을 모니터링하고 분석 오류를 개선시키기 위한 목적으로 개발된 테스트베드 도구를 사용자 인터페이스 관점에서 소개한다.

II. 기술문헌분석 플랫폼 및 테스트베드 개요

본 논문에서 개발된 기술 문헌 분석 테스트베드 도구는 크게 기술개체 자동인식 테스트베드와 기술개체간 관계추출 테스트베드로 구분된다. [그림 1]은 기술 문헌 분석 통합 플랫폼(SINDI Platform)[5] 내에서 이들 두 가지 테스트베드의 위치와 역할을 도식화하고 있다.

그림 1. 기술문헌분석 플랫폼 내에서의 테스트베드의 위치 및 역할

[그림 1]에서 보듯이 기술 문헌 분석 플랫폼은 다양한 기반 사전과 추출 알고리즘을 적용하여 문헌 내에 존재하는 기술 개체를 자동으로 추출하는 SINDI-CORE[3]와 추출된 기술 개체 간의 의미적 연관 관계를 자동으로 추출하는 관계 추출 엔진인 SINDI-LINK[4]로 구성된다. 또한, 이들 기반 엔진의 성능평가 및 내부 기계학습 모듈들을 위한 학습 집합을 구축하기 위한 컬렉션 구축 지원 도구를 비롯하여, 위 두 엔진이 최종적으로 생성하는 기술개체 트리플 집합을 가시화하고 통합 관리하기 위한 통합지원도구가 존재한다.

1. 기술개체인식 엔진 (SINDI-CORE)

본 논문에서 기반하고 있는 기술개체인식 엔진의 가장 큰 특징은 우선 문헌 분석 실행 과정이 단계별로 모듈화되어 있다는 점이다. 문장 분석에서부터 최종적인 개체인식까지의 세부적인 과정이 완벽하게 독립적으로 구현되어 있으며, 이에 따라 대용량 문헌에 대한 모듈별 분산, 병렬 처리가 용이하다. 부가적으로 기술 개체 인식의 범위를 단순 명사구에서 확장하여 구절(전치사구 등)까지도 식별이 가능하다. 또한 1차적으로 인식된 기술개체 후보의 객관성을 평가하기 위해서 웹 통계 정보를 활용한다. 결론적으로, 사전 기반 단순 용어 인식에서부터 웹 통계 정보를 바탕으로 한 신규 용어 인식, 개체명(인명, 지명, 기관명) 인식 기능이 전체적으로 결합된 시스템이다.

그림 2. 기술개체인식 테스트베드 실행 화면 (1)

그림 3. 기술개체인식 테스트베드 실행화면 (2)

2. 기술개체 간 관계 추출 엔진 (SINDI-LINK)

현재 개발된 관계 추출 엔진의 추출 가능 관계 종류는 IS-A, PART-OF, COORDINATING 등을 비롯하여 총 29 종이다. 관계 추출 기법은 지도 학습 기법 (supervised learning)과 패턴 규칙 기법 (pattern rule-based method)이 결합된 형태이다. 패턴 규칙 기법을 위해서 문장 내에서 관계를 표현하는 술어-논항 패턴, 1,160 종을 수집하여 이를 적용하고 있으며, 지도 학습 기법을 위해서는 구문 트리 커널 (parse tree

kernel) 기법을 사용하고 있다.

위에서 설명한 두 가지 엔진들은 내부적으로 문장 분리, 구문 분석, 용어 추출, 구문 패턴 생성 등과 같은 여러 단계의 텍스트 처리 과정을 포함하고 있고, 이들 처리 과정에서 발생하는 오류를 모니터링하고 수정하기 위해서는 이들 엔진과 밀접하게 구동되는 테스트베드가 필수적이다. 본 논문에서는 [그림 1]에서 점선 사각형 내부에 존재하는 두 가지 테스트베드에 대해서 사용자 인터페이스 관점에서 상세히 소개하고자 한다.

그림 4. 기술개체 간 관계추출 테스트베드 실행화면 (1)

그림 5. 기술개체 간 관계추출 테스트베드 실행화면 (2)

III. 기술문헌분석 테스트베드

본 장에서는 앞에서 간략하게 설명한 기술문헌분석 테스트베드에 대한 세부 기능을 살펴본다. 본 논문에서 개발한 테스트베드의 가장 큰 특징은 웹 기반으로 구현되기 힘든 섬세하고 동적인 인터페이스를 위하여 자바 어플리케이션 기반의 클라이언트-서버 구조로 개발되었다는 점과, 단계별 처리 과정을 한 눈에 살펴볼 수 있도록 구성되었다는 점이다.

1. 기술개체인식 테스트베드

SINDI-CORE의 기술개체인식 과정은 크게 (1)문장 분리, (2)언어 분석(품사 태깅, 기저구 인식 등), (3)기술개체후보(개체명, 기술용어) 인식, (4)사전기반 기술개체 인식, (5)통계기반 기술개체 인식 등의 총 5단계 구분된다. 이들 각 과정마다 임시 출력 정보가 도출되며 테스트베드는 이들을 효과적으로 출력하고 다양한 매개변수 조정기능을 제공함으로써 성능 최적화와 동시

그림 6. SINDI 플랫폼 통합 아키텍처

에 오류 모니터링 기능을 수행할 수 있다. [그림 2]와 [그림 3]은 위에서 기술한 기술개체 인식 과정을 세부적으로 출력하는 기술개체인식 테스트베드의 실행화면을 보여주고 있다. 그림에서 보는 바와 같이 엔진 수행 단계별로 도출되는 중간 결과물을 제시함으로써 개발자 혹은 시스템 관리자가 엔진의 수행 상황을 효과적으로 모니터링할 수 있다. 그림의 상부에는 엔진이 생성하는 입출력 데이터를 그대로 출력하고, 하부에는 단계별 모듈 실행 결과가 입력 문헌에 반영되는 과정을 가공하여 출력하고 있다. 그림의 맨 우측에는 최종적으로 기술개체가 식별된 문헌을 동적으로 표현하고 있다.

2. 기술개체 간 관계추출 테스트베드

[그림 4]와 [그림 5]는 기술개체 간 관계추출 테스트베드의 실행화면을 보여준다. 관계추출의 단계는 크게 (1)전처리 단계, (2)구문분석 단계, (3)술어-논항 구조 (Predicat-Argument Structure, PAS) 변환 단계, (4)관계추출 단계로 나눌 수 있다. 관계 추출 기법에 대한 세부적인 사항은 [4]에 상세하게 기술되어 있다.

특정 문장에 포함된 두 기술개체 간의 의미적 관계를

파악하기 위해서 SINDI-LINK는 어휘적, 구문적 패턴을 동시에 활용한다. 이러한 이유로 테스트베드에서는 특정 문헌을 구성하는 개별 문장의 구문분석 결과를 비롯하여 구문적 의존관계를 보다 상세하게 제시함으로써 시스템의 수행 상황을 면밀하게 살펴볼 수 있다. [그림 3]에서 보는 바와 같이, 입력 문헌은 이미 기술개체가 식별 완료된 문서이다. 이들 문서에 대한 구문 분석 결과가 세 번째 단계에서 출력되는데, 구문 트리의 출력 효율성을 극대화하기 위해서 단순 트리 형태로 표현하는 것이 아니라, 문장의 구성 단어에 대한 품사 정보를 중첩 구조(nested structure)로 표현하고 있다. 예를 들어, [그림 5]의 좌측에 표현된 구문 분석 결과에서 "the occurrence of"는 정관사(DT) "the"와 명사(NN) "occurrence", 그리고 전치사(IN) "of"로 구성되어 하나의 명사구(NP)로 인식될 수 있다. 이러한 구조적 정보를 테스트 베드의 화면에서 쉽게 파악할 수 있다.

3. 테스트베드와 SINDI 엔진 간의 상호 연동

[그림 6]은 SINDI 엔진의 전체적인 아키텍처를 보여주고 있다. 입력 문헌에 대해서 사전 및 통계 기반 기술

개체 추출이 완료되면, 특정 문장에 존재하는 두 기술 개체 간의 관계를 표현하는 구문을 술어-논항 구조 형태로 추출한다. 추출된 관계 구문 패턴을 기반으로 미리 구축된 관계 패턴 사전에서 검색하여 가장 유사한 패턴을 찾은 후 이 패턴이 나타내는 관계 유형을 최종 의미적 관계로 판정하여 출력하는 과정을 거치고 있다. 이를 위해서 본 시스템은 다양한 공통 모듈들을 활용하고 있고 테스트베드는 이들 공통 모듈에 대한 입출력 뿐만 아니라 시전 및 통계 기반 개체 식별 모듈은 물론 술어-논항 기반 관계 패턴 매칭 및 관계 추출 모듈과도 연동하고 있다. 이를 위해서는 개별 엔진들의 모듈화가 중요한데, SINDI 플랫폼은 각 요소 모듈들이 모두 개별적으로 실행될 수 있을 정도로 캡슐화(encapsulation)가 잘 되어 있다. 모든 입출력은 XML 형태의 자체 프로토콜[6]을 준수하고 있으며, 언제든지 확장이 가능한 유연한 구조를 보이고 있다.

IV. 결론

본 논문에서는 기술 문헌 분석에 필요한 두 가지 핵심 엔진인 기술개체 인식엔진과 관계 추출 엔진을 위한 테스트베드를 개발하였다. 이로써 엔진 고유의 복잡성으로 인해서 지식 추출의 세부 진행 과정을 파악하지 못하였던 한계점을 극복할 수 있는 기반을 마련하였다. 또한 수동이 아닌 자동으로 분석되는 과정에서 기술적인 한계점으로 인해 발생하는 다양한 분석 오류(언어처리오류, 개체인식오류, 관계추출오류 등)에 대한 파악 및 분석을 통해서 전체 시스템의 성능 향상을 위한 기반을 마련할 수 있다.

향후 연구 방향으로는 우선 개발자 및 시스템 관리자를 위한 다양한 부가 기능을 개발하여야 한다. 이러한 기능에는 언어자원 관리, 반복 학습 기능, 관계 패턴 관리 기능 등이 있다. 부가적으로 단계별로 출력되는 정보의 복잡성이 2차원 평면에 그대로 투영되어 사용자 접근성을 낮게 하는 한계점을 극복하기 위해서 3D 기반의 상세 정보 브라우징 기능을 추가적으로 개발할 필요가 있다. 추가적으로 [7]에서 개발된 분산환경 기반의

지식처리 플랫폼과 연동하여 텍스트 분석의 효율성을 확보하는 연구도 필요하다. 마지막으로 개발되는 전체 학술 지식 추출 및 관리 플랫폼의 활용성을 높이기 위해서 모바일 기기 기반의 시스템을 추가적으로 개발할 계획을 가지고 있다.

참고 문헌

- [1] 최성필 “과학기술분야 시맨틱 전자 문헌의 반자동 구축에 관한 기반 연구,” 한국언어정보학회 2010년 정기학술대회, pp.25-34, 2010.
- [2] R. C. Bunescu and R. J. Mooney, "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
- [3] 최윤수, 송사광, 전홍우, 정창후, 최성필, "과학기술문헌분석을 위한 기계학습기반 범용 전문용어 인식 시스템", 정보처리학회논문지, Vol.18, No.5, pp.329-338, 2011.
- [4] 정창후, 전홍우, 최성필, 최윤수, 송사광, "술어-논항 구조의 패턴 유사도를 결합한 혼합 커널 기반 관계추출", 한국인터넷정보학회논문지, Vol.12, No.5, pp.73-85, 2011.
- [5] 최성필, 최윤수, 전홍우, 정창후, 송사광, 정한민, "SINDI-WALKS: 과학기술지식발견 워크벤치", 한국컴퓨터종합학술대회(KCC), 2012.
- [6] 정창후, 최성필, 이민호, "기술용어 간 관계추출의 성능평가를 위한 반자동 테스트 컬렉션 구축 프레임워크 개발", 한국콘텐츠학회논문지, Vol.10, No.2, pp.481-489, 2010.
- [7] 정창후, 최성필, 윤화목, "그리드 기반의 고성능 과학기술지식처리 프레임워크 개발", 한국콘텐츠학회논문지, Vol.9, No.12, pp.877-885, 2009.
- [8] 최성필, 정창후, 최윤수, 맹성현, "평면적 어휘 자

질들을 활용한 확장 혼합 커널 기반 관계 추출”,
정보과학회논문지, Vol.36, No.8, pp.642-652,
2009.

저 자 소 개

최 성 필(Sung-Pil Choi)

정회원

- 1998년 : 부산대학교 전자계산학 석사
- 2012년 : 한국과학기술원 정보통신공학 박사
- 1998년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 자연어처리, 정보검색, 텍스트마이닝

송 사 광(Sa-kwang Song)

정회원

- 1999년 : 충남대학교 대학원 컴퓨터과학 석사
- 2011년 : KAIST 대학원 전산학 박사
- 2005년 ~ 2010년 : 한국전자통신연구원 연구원

- 2010년 ~ 현재 : 과학기술정보연구원 선임연구원

<관심분야> : 텍스트 마이닝, 자연어처리, 정보검색

정 한 민(Hannin Jung)

정회원

- 1994년 : POSTECH 전자계산학 석사
- 2003년 : POSTECH 컴퓨터공학 박사
- 2004년 ~ 현재 : 한국과학기술정보연구원 책임연구원/실장

- 2009년 1월 ~ 현재 : 한국콘텐츠학회 이사

<관심분야> : 시맨틱웹, HCI, 자연어처리