생의학분야 PLOT 및 관계추출을 위한 테스트컬렉션 구축

Construction of Test Collection for Extraction of Biomedical PLOT & Relations

최윤수, 최성필, 정창후
한국과학기술정보연구원

Yun-Soo Choi, Sung-Phl Choi , Chang-Hoo Jeong
Korea Institute of Science and Technology Information

요약

대용량 문서에서 정보를 추출하는 작업은 크게 개체명 인식, 전문용어 인식, 관계추출 작업으로 구성된다. 이들 각각의 기술들은 지금까지 독립적으로 연구되어 왔기 때문에, 이와 관련된 기계학습모델을 위한 테스트컬렉션 또한 독립적으로 구축되어 왔다. 과학기술문헌의 경우 개체명과 전문용어가 혼재되어 있는 형태로 구성된 문서가 많아, 기존의 연구결과를 이용하여 접근한다면 결과물 통합과정의 불편함과 처리속도에 많은 제약이 따르므로, 개체명과 전문용어를 동시에 추출 할 수 있는 기계학습 모델을 위한 테스트컬렉션이 필요하다. 본 연구에서는 생의학 분야 과학기술문헌에 대한 개체명, 전문용어를 통합한 PLOT(Person, Location, Organization, Terminology)과, PLOT 간의 관계추출을 위한 테스트컬렉션을 구축한다.

Abstract

Large-scaled information extraction consists of named-entity recognition, terminology extraction and relation extraction. Since all the elementary technologies have been studied independently so far, test collections for related machine learning models also have been constructed independently. As a result, it is difficult to handle scientific documents to extract both named-entities and technical terms at once. In this study, we integrate named-entities and terminologies with PLOT(Person, Location, Organization, Terminology) in a biomedical domain and construct a test collection of PLOT and relations between PLOTs.

Ⅰ. 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 recognition), (2)대용어 참조해소(coreference resolution), (3)관계추출(relation extraction)의 세 가지 요소기술로 세분화 된다.[1]

개체명인식은 신문기사, 뉴스 등의 매체에서 인명(Person), 지명(Location), 기관명(Organization)을 추출하는 일반 개체명인식과 과학기술 데이터, 특히 생의학분야에서 유전자(Gene), 단백질(Protein)등을 추출하는 전문용어 인식으로 구분할 수 있다.

일반개체명 인식 및 관계인식은 MUC-6에서 출발하여 CoNLL, ACE 등을 거쳐 많은 연구가 되었고, 전문용어 인식과 전문용어들 간의 관계인식은 주로 생의학분야에서 BioNLP, BioCreative 등을 통해 독립적으로 연구되어 왔다.[2][3][4][5]

그러나, 과학기술문헌의 경우 일반개체명과 전문용어가 혼재되어 있는 형태로 구성된 문서가 많아, 기존의 연구결과를 이용하여 접근한다면 결과물 통합과정의 불편함과 처리속도에도 많은 제약이 따른다.[6]

본 연구에서는 독립적으로 연구되어 온 일반개체명과 생의학분야 전문용어를 통합하여 PLOT이라 정의하고, PLOT의 세부 분야를 확정한 뒤, PLOT을 인식하고 PLOT간의 관계를 인식하는 기계학습모델의 학습 및 테

스트 데이터를 위한 테스트컬렉션을 구축한다.

Ⅱ. 테스트컬렉션 구축

PLOT은 [표 1]과 같은 클래스로 구분된다. 일반개체명은 PLO로 세분화되고, 전문용어은 기술명과 분야특화명으로 구분된다. 분야특화명은 대상 과학기술문헌의 분야에 적합한 세분화된 전문용어이고, 기술명은 과학기술문헌에서 사용되는 일반적인 전문용어이다.

PLOT간의 관계추출을 위해 1)활용/적용하다 2)변화/변경하다 등 39개의 관계를 정의하였고, 개체태깅과 관계태깅을 위한 DTD를 제작하였다.

표 1. PLOT 구성

No	상위클래스	하위클래스	비 고
1	일반개체명	Person	인명
2		Location	지명
3		Organization	기관명
4	기술명	TechTerm	장비/수술명 등
5		Others	기타
6	분야특화명	Gene	유전자명
7		Protein	단백질명
8		Disease	질병명
9		Organism	유기체명
10		Drug	약명

테스트컬렉션 구축을 위한 대상데이터는 한국과학기술정보연구원이 보유하고 있는 해외학술지에서 선별하여 10,261건(2000년~2008년), 과학기술 신문기사[1] 11,185건(2000년~2009년)을 수집하였다.

테스트컬렉션 구축 시 오류를 최소화 하고 작업속도를 증진시키기 위하여, 테스트컬렉션 구축지원도구를 개발하여 사용하였다. 구축지원도구는 구축된 테스트컬렉션에 대한 오류검증, 구축통계 등의 부가적인 작업을 또한 지원하도록 설계되었다.

표 2. KEEC/KREC 2009 구축 현황

통 계 정 보	건 수	비 고
문서 수	354	전체 구축 건수
문장 수	8,303	태깅된 문장
PLOT 수	21,142	태깅된 PLOT
PLOT포함 문장 수	7,032	PLOT을 포함
관계 수	3,494	태깅된 관계 수
관계포함 문장 수	2,107	관계를 포함

전문가 4인이 2인1조로 구축지원도구를 이용하여 KEEC/KREC 2009를 구축 및 검증작업을 수행하였다. PLOT간의 관계태깅은 한 문장안에서의 관계로 한정하였고, 대용대명사에 대한 참조는 전체문서를 대상으로 하였다. 구축된 테스트컬렉션 현황은 [표 2]와 같다.

Ⅲ. 결론 및 향후연구

테스트컬렉션은 기계학습모델의 학습 및 테스트를 위하여 제작된다. 기계학습모델의 특성상 학습을 위한 테스트컬렉션의 품질 및 구축량이 매우 중요한 요소이다. 본 연구에서 구축된 KEEC/KREC[2]에 대한 품질향상을 위한 구축방법 및 구축결과에 대한 검증이 필요하고, 구축량의 증진을 위해 구축도구의 기능을 개선할 여지가 있다.

본 연구에서 테스트컬렉션을 위해 사용된 DTD는 용어에 대한 엘레멘트를 이용하여 태깅하는 방식으로 구성되어 있다. 이 때문에 병렬구조로 연결되어 있는 용어(예, FOG and GATA proteins)를 처리할 수 없다. 향후 연구에서는 이러한 용어들을 수용할 수 있는 DTD 제작이 필요하다.

본 연구의 결과물은 정보추출 연구자 및 개발자에게 공개될 예정이며, 생의학분야 PLOT추출 및 PLOT간 관계추출을 목적으로 하는 학술대회의 연구결과 발표 및 제품 비교 등에 활용될 예정이다.

■ 참 고 문 헌 ■

[1] Bunescu, R.C., Mooney, R.J., "A Shortest Path

1) EurekAlert [http://www.eurekalert.org]

2) KEEC/KREC : KISTI Entity/Relation Extraction Collection,

Dependency Kernel for Relation Extraction", Proceedings of the Human Language Technology Conference, Vancouver, B.C., pp.724-731, 2005.

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

[3] NIST, 2009, Automatic Content Extraction Evaluation [www.itl.nist.gov/iad/mig//tests/ace/]

[4] BioCreative, [http://biocreative.sourceforge.net]

[5] BioNLP, [http://compbio.uchsc.edu/BioNLP2009]

[6]] 최윤수,정창후,최성필,류범종,김재훈, "대용량 자원기반 과학기술 핵심개체 탐지를 위한 정보추출기술 통합에 관한 연구",정보관리연구 pp.1-22, 2009.

[7] 최윤수,최성필,정창후,윤화묵,류범종, "과학기술 분야 용어 간 관계추출 시스템의 평가를 위한 테스트컬렉션 구축", 한국콘텐츠학회 2009 춘계종합학술대회 pp.754-758