바이오 분야 학술 문헌에서의 분야별 관계 추출 데이터셋 반자동 구축에 관한 연구*

- 알츠하이머병 유관 유전자 간 상호 작용 중심으로 -

A Study on the Semiautomatic Construction of Domain-Specific Relation Extraction Datasets from Biomedical Abstracts

- Mainly Focusing on a Genic Interaction Dataset in Alzheimer's Disease Domain -

최 성 필(Sung-Pil Choi)** · 유 석 종(Suk-Jong Yoo)***
조 현 양(Hyun-Yang Cho)****

〈목 차〉

Ⅰ. 서론
Ⅱ. 관련연구
Ⅲ. 관계 추출 학습 집합 반자동 구축 프로세스
Ⅳ. 알츠하이머병 분야 유전자간 상호 작용 추출 학습 집합 구축
1. 알츠하이머병 분야 유전자간 상호 작용 정보 생성
2. PubMed 데이터베이스를 활용한 유전자 간 상호 작용 포함 문장 수집
Ⅴ. 결론 및 향후 연구 방향

초 록

본 논문에서는 생의학 분야의 특정 세부 분야에 특화된 관계 추출 학습 말뭉치를 효율적으로 구축할 수 있는 시스템을 소개한다. 이 시스템은 대상 분야에 해당하는 용어집(유전자, 단백질, 질환 명칭 등)을 입력하면, 대용량 상호 작용 데이터베이스를 통해서 이들 용어 간의 연관 관계를 1차적으로 생성하고 생성된 연관 관계 집합을 다시 학술 데이터베이스에서 검색하여 최종적으로 연관 관계 포함 문장을 추출하는 형태로 수행된다. 개발된 시스템의 유용성 검증을 위해서 알츠하이머병 분야에서의 유전자 간 상호 작용 학습 말뭉치를 구축하는데 본 시스템을 적용하였고, 140개의 유전자 집합을 입력하여 이 분야에 특화된 학습 집합인 유전자 쌍 및 상호 작용 포함 문장 3,510 건을 추출하였다. 본 논문에서 제안한 시스템을 활용함으로써 기존에 완전 수작업으로 수행되던 연관 관계 추출용 학습 말뭉치 구축의 효율성을 높일 수 있고 다양한 세부 분야에 적합한 학습 말뭉치 구축에 도움을 줄 수 있다.

키워드: 관계 추출, 학습 집합 구축, 유전자 간 상호 작용, 기계 학습, 텍스트 마이닝

ABSTRACT

This paper introduces a software system and process model for constructing domain-specific relation extraction datasets semi-automatically. The system uses a set of terms such as genes, proteins diseases and so forth as inputs and then by exploiting massive biological interaction database, generates a set of term pairs which are utilized as queries for retrieving sentences containing the pairs from scientific databases. To assess the usefulness of the proposed system, this paper applies it into constructing a genic interaction dataset related to Alzheimer's disease domain, which extracts 3,510 interaction-related sentences by using 140 gene names in the area. In conclusion, the resulting outputs of the case study performed in this paper indicate the fact that the system and process could highly boost the efficiency of the dataset construction in various subfields of biomedical research.

Keywords: Relation extraction, Dataset construction, Genic interactions, Machine learning, Text mining

* 본 연구는 한국과학기술정보연구원 주요사업 "초고성능컴퓨팅 기반 건강한 고령사회 대응 빅데이터 기술개발" 과제의 연구비 지원으로 수행되었음(K-16-L03-C02-S02).

** 경기대학교 문헌정보학과 조교수(spchoi@kgu.ac.kr) (제1저자)
*** 한국과학기술정보연구원 생명의료융합기술연구실 실장(codegen@kisti.re.kr) (공동저자)
**** 경기대학교 문헌정보학과 교수(hycho@kgu.ac.kr) (교신저자)

•논문접수: 2016년 11월 20일 •최초심사: 2016년 11월 28일 •게재확정: 2016년 12월 22일
•한국도서관·정보학회지 47(4), 289-307, 2016. [http://dx.doi.org/10.16981/kliss.47.201612.289]

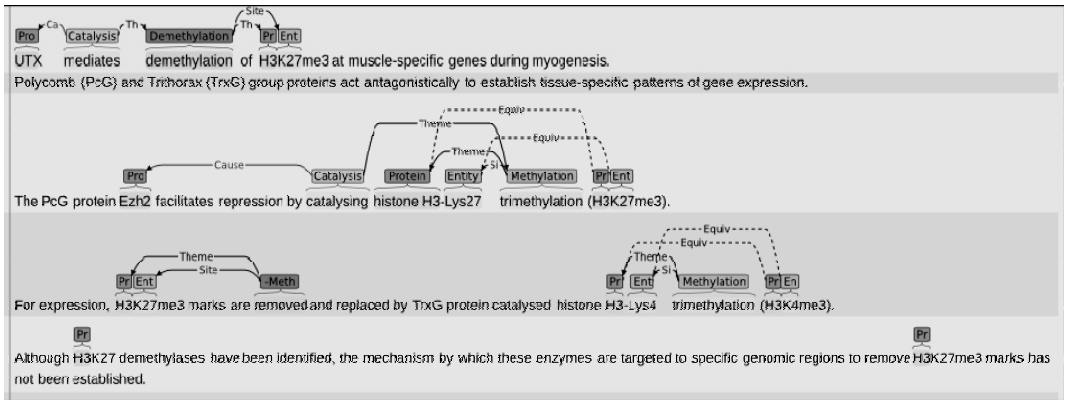
Ⅰ. 서론

기계 학습(Machine Learning) 기반의 자연어 처리(Natural Language Processing) 및 정보 추출(Information Extraction) 기술이 급속도로 발달함에 따라서 이에 필요한 대량의 학습 집합(Training Set)에 대한 요구 사항도 지속적으로 높아지고 있다. 일반적으로 기계 학습 방법은 크게 지도 학습(Supervised Learning)과 비지도 학습(Unsupervised Learning)으로 구분할 수 있다(Hastie, Tibshirani, and Friedman 2009). 이 중, 지도 학습은 기계 학습 모델을 유추하기 위해서 충분한 규모의 학습 집합을 요구한다. 학습 집합이란 이미 학습된 기계 학습 모델의 입력에 대해서 그에 해당하는 출력, 즉 정답을 미리 기술해 놓은 자료로서, 이를 구축하는데 분야 전문가들을 활용한 고도의 수작업이 요구된다(Ivanović and Budimac 2014; Rubin, Shah, and Noy 2008). 특히 신문 기사나 소셜 데이터 등과 같은 일반적인 문헌을 바탕으로 구축되는 학습 집합과는 달리 생의학 분야에서의 학습 집합은 특정 분야의 학술 정보를 기반으로 구축되어야 하므로 의학 및 생물학을 전공한 전문가가 없이는 구축이 거의 불가능하다(Alex et al. 2008).

새롭게 발견되거나 고안된 대부분의 학술 지식들이 학술 문헌 내에서 자연어 형태로 표현되어 있기 때문에, 현재 기하 급수적으로 증가하고 있는 생의학 분야의 학술 문헌에 대한 정형화 및 지식화를 위해서 가장 많이 활용되고 있는 기술은 정보 추출(Information Extraction) 기술이다(Saffer and Burnett 2014). 정보 추출은 비정형적인 텍스트를 정형화하기 위한 텍스트 마이닝 기법의 한 종류로서, 크게 텍스트 내의 주요 용어를 자동으로 식별 및 분류하는 개체명 인식(Named-Entity Recognition) 기술과 이들 용어 간의 관계를 인식하고 분류하는 관계 추출 기술로 구성된다(Blaschke, Hirschman, and Valencia 2002). 아래 <그림 1>은 비정형 문헌에 대한 정보 추출의 결과를 Annotation 도구의 일종인 Brat[1]을 이용해서 가시화한 결과를 보여주고 있다.

<그림 1>에서 보는 바와 같이 학술 문헌에 대한 정보 추출 과정에서 우선 문헌에 나타난 “UTX”, “H3K27”, “Ezh2” 등과 같은 주요 용어를 식별하고, 이들 간의 의미적 연관 관계를 인식 및 분류하게 된다(“histone H3”와 “H3”는 동등 관계 등). 기계 학습 기반 정보 추출 시스템을 개발하기 위해서는 위와 같이 구성된 “표지 부착된 텍스트 말뭉치(annotated text corpus)”가 필요하다. 이를 기반으로 기계 학습 모듈은 추출할 용어 주변 문맥 정보 및 용어 간 관계 표현 문구 등을 추출하여 자질로 활용함으로써 새로운 문장에 포함된 주요 정보들을

1) “brat”은 일본 동경대학에서 개발한 말뭉치 구축 도구임. (http://brat.nlplab.org/index.html)



<그림 1> 생의학 논문 초록에 대한 정보 추출 결과 (Kim et al., 2011)

추출할 수 있게 된다.

본 논문에서는 기계 학습 모듈의 학습 집합으로 활용될 수 있는 이 표지 부착된 텍스트 말뭉치(이하 표지 부착 말뭉치로 명시함)를 학술 문헌 집합에서 효율적으로 생성할 수 있는 방법론을 제시한다. 이를 위해서 세부 분야별 용어 사전 만을 가지고도 최종적으로 용어 사전 내의 표제어는 물론 표제어 간 연관 관계가 표현되어 있는 문장 집합을 자동으로 수집할 수 있는 분야별 관계 추출 학습 집합 반자동 구축 시스템을 소개한다. 여기서 “반자동”이라는 표현을 사용한 이유는 궁극적으로는 시스템의 결과물을 분야 전문가가 최종적으로 확인할 필요가 있음을 시사하지만, 기존의 학습 집합 구축 방법론(Alnazzawi, Thompson, and Ananiadou 2014; Hirschman, Yeh, Blaschke, and Valencia 2005; Thompson, Iqbal, McNaught, and Ananiadou 2009)들과 비교하면, 대상 문헌 확보, 수동 문장 분리, 용어 식별 등과 같은 필수적으로 수반되는 수작업이 불필요하기 때문에 높은 효율성을 나타냄을 알 수 있다. 본 논문에서 개발된 시스템의 유용성을 확인하기 위해서 의학 분야에서 높은 관심과 함께 많은 연구가 진행되고 있는 퇴행성 뇌질환의 일종인 알츠하이머병 분야와 유관한 유전자들 간의 상호 작용 추출을 위한 관계 추출 학습 집합 구축에 본 시스템을 적용한 결과를 보인다.

본 논문의 구성은 다음과 같다. 우선 2장에서는 관계 추출 학습 집합 구축과 관련된 기존 연구들을 살펴본다. 이어서 3장에서 본 논문에서 개발된 관계 추출 학습 집합 반자동 구축 모델에 대해서 설명하고, 4장과 5장에서는 이 시스템을 활용하여 알츠하이머병과 직접적으로 관련이 있는 유전자 간의 상호 작용 추출을 위한 학습 집합 구축 과정을 세부적으로 살펴본다. 마지막으로 6장에서 연구 결과에 대한 논의와 함께 향후 연구 방향을 기술한다.

Ⅱ. 관련 연구

지금까지 생의학 분야 학술 정보 지식화를 위해서 개발된 많은 시스템들의 성능 평가는 물론 기계 학습 기반 엔진들의 학습 집합으로 활용될 수 있는 다양한 종류의 표지 부착 말뭉치들(Annotated Corpora)이 구축되어 왔다. 이들 대부분의 말뭉치들은 특정 학술 대회에서의 경쟁 평가 대회(Community Challenges)를 통해서 구축되고 공개되었으며, 해당 대회가 끝난 후에도 지속적으로 활용되어 많은 연구 성과를 도출하는데 도움을 주고 있다(Huang and Lu 2016). 여기서는 지금까지 구축된 다양한 말뭉치 중에서 본 논문에서 중점을 두고 있는 생의학 분야 관계 추출 말뭉치 및 그 구축 방법에 대해서 한정하여 설명한다.

생의학 분야에서의 개체 간 관계 추출(entity-entity relation extraction)의 종류에는 약물 간 상호작용(Drug Drug Interaction, DDI) 추출(Segura Bedmar, Martnez, and Sánchez Cisneros 2011), 단백질 간 상호작용(Protein Protein Interaction, PPI) 추출(Krallinger, Leitner, Rodriguez-Penagos, and Valencia 2008), 단백질 잔류물 관계(Protein Residue Relation) 추출(Ravikumar, Liu, Cohn, Wall, and Verspoor 2012), 유전자 간 관계(Genic Interaction) 추출(Nédellec 2005), 진료 기록 문헌에서의 개념 간 관계 추출(Uzuner, South, Shen, and DuVall 2011) 등이 있으며 대부분 해당 평가 학술 대회를 통해서 섭외된 분야 전문가들에 의해 표지 부착 말뭉치들이 수동 구축되어 공개되었다. 다음 <표 1>은 Huang and Lu(2016)에서 정리한 바이오 분야 경쟁 평가 학술 대회 및 그 대회를 통해서 구축된 말뭉치에 대한 자료 중에서 관계 추출에 해당되는 부분만 발췌하여 개별적으로 정리한 도표이다.

<표 1> 에서 보는 바와 같이 구축 연도에 따라 개별 말뭉치가 서로 다른 버전이라고 고려한다면 총 10 종의 말뭉치가 현재 존재하며 이들 모두는 다수의 전문가에 의한 수작업으로 구축되었다. 위에서 열거한 경쟁 평가 학술 대회(Community Challenge) 중에서 현재까지도 지속적으로 개최되고 있는 학회는 ACL(Association for Computational Linguistics) 산하의 BioNLP-ST와 미국의 NIH가 후원하고 있는 NCBC(National Center for Biomedical Computing)가 개최하고 있는 i2b2가 있다. 특히 i2b2는 기존의 학술 문헌에서 벗어나 의료 정보 즉, 진료 기록 정보, 임상 정보, 의학 특허 등의 정보를 기반으로 지속적으로 말뭉치를 구축하고 공유한다.

다양한 종류의 생의학 지식 중에서 단백질 및 유전자 간 상호 작용 정보는 연구를 수행함에 있어서 가장 핵심적이고 중요한 정보이다. 이러한 이유로 위에서 기술된 말뭉치 외에도 많은 말뭉치들이 구축되었다. AIMed(Bunescu et al. 2005)는 1,000 건의 Medline 초록에 나타

<표 1> 바이오 분야 경쟁 평가 학술대회 및 구축된 관계 추출 말뭉치

경쟁 평가 학술 대회 (Challenges) (웹 사이트 URL)	구축된 표지 부착 말뭉치 종류	구축 연도
BioCreative (http://www.biocreative.org/)	단백질 간 상호작용 추출 말뭉치 (Protein Protein Interactions Extraction)	2007, 2009, 2010
LLL (http://www.cs.york.ac.uk/aig/lll/lll-05/)	유전자 간 상호작용 추출 말뭉치 (Genic Interactions Extraction)	2005
DDIExtraction (http://www.mavir.net/conf/137-ddiextraction2013)	약물 간 상호작용 추출 말뭉치 (Drug Drug Interactions Extraction)	2011, 2013
BioNLP-ST (http://2011.bionlp-st.org, http://2013.bionlp-st.org)	박테리아 비오톱(biotope)과 유전자 상호작용 추출 말뭉치 (Bacteria, BB)	2011, 2013
	유전자 간 상호작용 추출 말뭉치 (Genic Interactions Extraction, GE)	2011
i2b2 (https://www.i2b2.org/NLP/)	진료 기록에서 의학 개념 간의 관계 추출 말뭉치 (Clinical Relation Extraction)	2010

나는 유전자 및 단백질 명칭들을 대상으로 단일 문헌(초록) 내에서 그들 간의 상호 작용을 나타내는 표현이 사용되었다면 이를 수동으로 표시함으로써 구축된 말뭉치이다. 그런데 이 말뭉치의 특징이자 단점은 상호 작용 표현의 범위를 단일 문장으로 국한시키지 않았다는 점이다. 예를 들어, 한 쌍의 단백질 명칭 P1과 P2가 각기 다른 문장에 존재하더라도 전체적인 의미상 서로 상호 작용을 가진다고 판단되면 이를 표시하였다. 그런데 지금까지 연구된 대부분의 관계 추출 및 상호 작용 추출 모델은 단일 문장 내에서의 한 쌍의 개체 간 관계 유무 판단 및 관계 분류를 수행하고 있으므로 이를 활용하기 위해서는 세심한 전처리 작업이 필요하다. BioInfer(Pyysalo et al. 2007)는 세심하게 선정된 1,100 건의 문장 집합에 대해서 개별 문장에 존재하는 유전자, 단백질, RNA 명칭들과 같은 개체명들을 표시하고 이들 간의 상호 작용 정보를 표현한 표지 부착 말뭉치이다. 이 말뭉치의 가장 큰 특징은 단순히 개체명 간의 상호작용 유무 외에 세부 관계 유형까지도 파악하여 이를 작성해 놓았다는 점이다. 그 외에도 HPRD50(Fundel, K□ffner, and Zimmer 2007), IEPA(Ding, Berleant, Nettleton, and Wurtele 2002), LLL(N□dellec 2005) 등이 많이 활용되고 있는 생의학 분야 유전자 및 단백질 간 상호 작용 추출 말뭉치들이다.

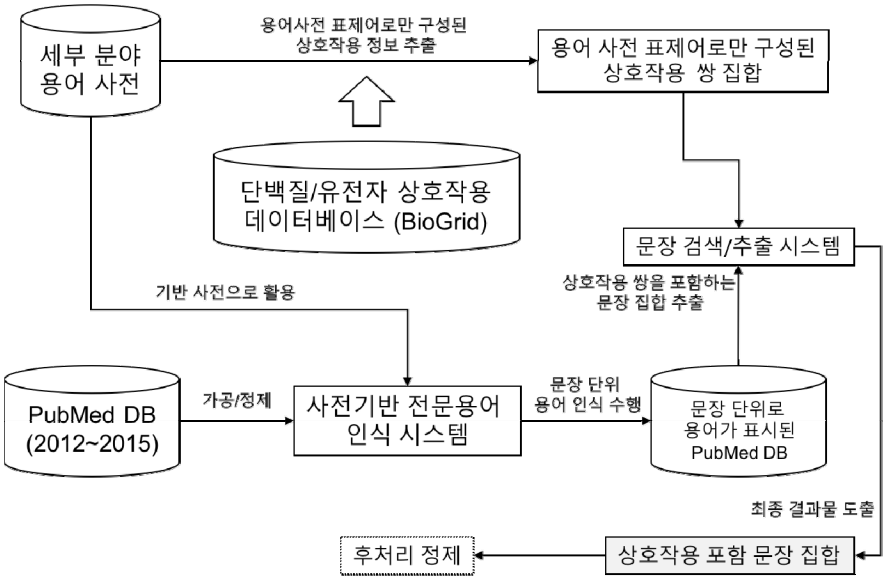
위에서 제시한 다양한 학습 데이터를 기반으로 국내에서도 생의학 분야 텍스트 마이닝(박경미, 황규백 2011; 허고은, 송민 2014) 및 정보 추출(정창후 외 2010; 최성필 2016)에 관한 연구가 다양하게 수행되었다. 그러나 대부분 분야 전문가에 의해 수작업으로 구축된 한정된 학습 집합 혹은 기반 자원을 이용하였으며, 본 논문에서 제안하는 생의학 분야 관계 추출을 위한 학습 집합의 효율적 구축 방안에 대한 심층적인 연구는 이루어지지 않았다.

지금까지의 관련 연구를 살펴본 결과, 거의 대부분의 관계 추출 말뭉치들이 수작업으로 구축되었으며 현재까지도 분야 전문가를 활용한 수작업 구축이 주를 이루고 있다. 일부 연구에

서 기존의 말뭉치들을 확장하거나 보완하기 위한 방법론들(Haddow and Alex 2008; Ravikumar et al. 2012)을 제시하기도 하였으나 처음부터 새로운 세부 분야의 관계 추출 말뭉치를 구축하는 체계적인 방법론에 대한 연구는 아직 부족한 실정이다.

Ⅲ. 관계 추출 학습 집합 반자동 구축 프로세스

본 논문에서 제안하는 학습 집합 반자동 구축 프로세스는 기본적으로 한 쌍의 개체명이 단일 문장 혹은 특정 문맥 범위 내에 존재하면 서로 연관성이 있다는 가정하에서 출발한다(Mintz, Bills, Snow, and Jurafsky 2009). 우선 구축하고자 하는 해당 분야에 속하는 용어 집합을 활용하여 2개 이상의 용어가 존재하는 문장을 대용량 데이터베이스에서 추출하여 이를 후처리 검증함으로써 최종적인 학습 집합이 구축된다(<그림 2>).



<그림 2> 관계 추출 학습 집합 반자동 구축 프로세스 도식화

<그림 2>에서 보듯이, 우선 추출 대상 학술 논문 초록 데이터베이스를 가공하고 정제하여 전체 문장 집합을 추출한다. 이렇게 문장 분리된 데이터에 “사전 기반 전문용어 인식 시스템”을 적용하여 개별 문장 내에 존재하는 용어를 식별하고 이를 데이터베이스화한다. 이때 “세부 분야 용어 사전”은 기존에 존재하는 체계적으로 정제된 분야 사전이 활용 가능하다면 그대로 적용할 수 있다. 그러나 4장에서 보는 바와 같이, 본 논문에서는 기반 텍스트 자원이 충분하지 않은 “알츠하이머병” 분야를 다루기 때문에 일반적인 표제어 사전이 아닌 온톨로지 형태의

자원을 변환/가공하여 용어집을 구축하였다. 위 그림의 상단에서 보는 바와 같이, 특정 분야에 해당하는 용어집이 선정되면 이를 바탕으로 용어 간의 상호작용을 생성하기 위해서 단백질 및 유전자 상호작용 데이터베이스를 검색하여 모든 용어 간 상호작용 쌍을 생성한다. 본 논문에서 사용한 상호작용 데이터베이스는 생의학 분야 연구 과정에서 규명되거나 발견된 유전자 및 단백질 간의 상호 작용 정보를 수동으로 구축해 놓은 대용량 공개 자료로서 대표적인 데이터베이스로는 BIND(Bader, Betel, and Hogue 2003), BioGRID(Stark et al. 2006), DIP (Database of Interacting Proteins)(Xenarios et al. 2000), IntAct(Hermjakob et al. 2004), MINT(Molecular INTeraction database)(Chatr-aryamontri et al. 2007) 등이 있다. 본 논문에서는 이들 중에서 BioGRID(Stark et al. 2006)[2]를 사용한다. BioGRID는 현재 버전 3.4.136까지 출시되어 있고 총 1,066,335개의 유전자 및 단백질 상호 작용 정보를 포함하고 있다.

<그림 2>에서의 "문장 검색/추출 시스템"은 용어가 식별된 문장 집합을 대상으로 추출된 용어 간 상호작용 쌍을 검색하여, 해당 상호작용 쌍이 존재하는 문장을 선별하게 된다. 최종적으로 상호작용 쌍이 존재하는 모든 문장들을 추출하여 후처리 정제 작업을 거치게 된다.

상호작용 정보를 포함하는 실제 문장들을 수집하기 위해서 세계 최대의 의학 분야 학술 데이터베이스인 PubMed[3] 자료 중에서 2012년부터 2015년까지 총 5,009,821 건의 논문 학술 초록을 이용하였으며, 이를 분석하여 추출된 문장 집합의 규모는 총 39,045,561 건이다. 이들 문장 집합에 대해서 "사전 기반 전문용어 인식 시스템"을 활용하여 문장 내에 존재하는 유전자 명칭에 대한 자동 식별 작업을 진행하였다. 개발된 전문 용어 인식 시스템은 특정 문장에 대한 품사 태깅을 통해서 문장 내에 존재하는 명사구("형용사 + 명사" 혹은 복합 명사)를 용어 후보로 추출하고 이들 집합을 대상으로 세부 분야 사전에 이들 용어 후보가 존재하는지를 검색함으로써 해당 명사구가 실제 용어인지를 판별한다. 이 과정에서 개별 문장에 대한 품사 태깅을 위해서 본 논문에서는 Stanford CoreNLP(Manning et al. 2014) 라이브러리를 사용하였다. 최종적으로 이전에 분야 용어 사전과 BioGRID를 이용하여 추출된 세부 분야 상호작용 정보를 질의로 하여 문장 검색을 수행한 후, 상호 작용 정보가 포함된 문장 집합들을 추출하게 된다.

Ⅳ. 알츠하이머병 분야 유전자간 상호 작용 추출 학습 집합 구축

이 장에서는 3장에서 소개한 반자동 학습 집합 구축 시스템을 이용하여 알츠하이며병 분야 유전자간 상호 작용 추출 학습 말뭉치를 추출하는 과정을 상세히 기술한다. 전체 과정은 2단

2) http://thebiogrid.org/
3) http://www.ncbi.nlm.nih.gov/pubmed

계로 나뉘는데 우선 BioGRID를 이용하여 해당 분야 유전자간 상호 작용 정보(유전자 쌍 집합)를 생성하는 과정과 생성된 상호 작용 정보를 바탕으로 최종적으로 상호 작용 포함 문장을 추출하는 과정으로 나뉜다.

1. 알츠하이머병 분야 유전자간 상호 작용 정보 생성

우선 알츠하이머병과 직접적으로 관련이 있는 유전자 리스트(용어사전)를 확보하기 위해서 본 연구에서 활용한 용어집은 Alzheimer's Disease Ontology (이하 ADO)[4] (Malhotra et al. 2014) 이다. 이 온톨로지는 세계 최초로 구축된 알츠하이머병 분야 공개 온톨로지로서 동 분야에 관련된 다양한 유형의 전문용어 및 개념들을 포함하고 있다. ADO에 대한 세부적인 통계 자료는 다음 표와 같다.

<표 2> Alzheimer's Disease Ontology(ADO) 구축 현황 통계

특성 항목	개수
NUMBER OF CLASSES	1564
NUMBER OF INDIVIDUALS	0
NUMBER OF PROPERTIES	12
MAXIMUM DEPTH	11
MAXIMUM NUMBER OF CHILDREN	147
AVERAGE NUMBER OF CHILDREN	6
CLASSES WITH A SINGLE CHILD	57
CLASSES WITH MORE THAN 25 CHILDREN	8
CLASSES WITH NO DEFINITION	752

개별 클래스들은 일반적으로 우리가 생각하는 개체명(named-entity) 혹은 용어일 수도 있고 상위 분류(upper class) 정보나 용어 타입일 수도 있다. 일반적인 온톨로지와는 달리 속성(Property) 정보가 매우 부족하고, 단지 용어 사전(terminology)으로서의 역할만 수행할 수 있다. 개별 클래스에는 상위 클래스(parent class)가 지정되어 있으므로 이들 상위 클래스를 개체명의 유형으로 간주해도 된다. 알츠하이머병에 직접적인 연관성이 있는 클래스들도 존재하지만 매우 일반적인 클래스들도 많이 존재한다. 따라서 이 온톨로지 전체를 활용하는 것은 세부 분야별 학습 집합을 구축하는 데는 별로 도움이 되지 않는다. 위 표에서 보듯이 이 온톨로지에는 "Individual"이 존재하지 않는다. 다시 말해서 클래스의 인스턴스가 없다는 뜻인데, 이는 모든 클래스, 복합 개념, 용어들이 클래스로 정의되어 있음을 의미한다. 아래 표에 ADO 온톨로지의 일부를 예시로써 보여주고 있다.

4) http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics.html

<표 3> ADO 온톨로지 내용 예시

Class ID (URI)	용어	부모 클래스 (유형)
http://scai.fraunhofer.de/AlzheimerOntology#hepatitis	hepatitis	http://scai.fraunhofer.de/AlzheimerOntology#active_infectious_disease
http://scai.fraunhofer.de/AlzheimerOntology#Lymphocyte	Lymphocyte	http://scai.fraunhofer.de/AlzheimerOntology#Agranulocyte
http://scai.fraunhofer.de/AlzheimerOntology#Monocyte	Monocyte	http://scai.fraunhofer.de/AlzheimerOntology#Agranulocyte
http://scai.fraunhofer.de/AlzheimerOntology#wine	wine	http://scai.fraunhofer.de/AlzheimerOntology#alcohol
http://scai.fraunhofer.de/AlzheimerOntology#NOS1	NOS1	http://scai.fraunhofer.de/AlzheimerOntology#Alzheimer_risk_factor_gene
http://scai.fraunhofer.de/AlzheimerOntology#ACE	ACE	http://scai.fraunhofer.de/AlzheimerOntology#Alzheimer_risk_factor_gene
http://scai.fraunhofer.de/AlzheimerOntology#MAPK3	MAPK3	http://scai.fraunhofer.de/AlzheimerOntology#Alzheimer_risk_factor_gene
http://scai.fraunhofer.de/AlzheimerOntology#DPYSL2	DPYSL2	http://scai.fraunhofer.de/AlzheimerOntology#Alzheimer_risk_factor_gene
http://scai.fraunhofer.de/AlzheimerOntology#TGM2	TGM2	http://scai.fraunhofer.de/AlzheimerOntology#Alzheimer_risk_factor_gene

위 표에서 보는 바와 같이 개별 클래스의 식별자를 나타내는 “Class ID”는 URI와 Name으로 구성되어 있으며, 이 “Name”이 일반적인 용어로 생각될 수 있다. 또한 이와 더불어 “Preferred Label”에는 이 “Name”이 단독으로 명기되어 있다. 마지막으로 “Parents”에는 현재 클래스(용어)의 부모 클래스(타입)가 지정되어 있다. 따라서 이러한 정보를 활용하여 본 연구에서는 이 온톨로지를 아래 표와 같은 형태의 용어집으로 변환하였다.

<표 4> 변환된 알츠하이머병 관련 용어집 예시

개체명 (용어)	개체명 타입 (분류)
A2M	Alzheimer_risk_factor_gene
Abeta 40	amyloid_beta_protein
Abeta 42	amyloid_beta_protein
ACE	Alzheimer_risk_factor_gene
acetophenazine	drug_used_in_treatment
acetylcholinesterase	enzyme
Action potential	processes_related_to_neurons
Adenosinergic neuron	Neurons
Adrenergic neuron	Neurons
AGER	Alzheimer_risk_factor_gene
AKT1	Alzheimer_risk_factor_gene
ALDH2	Alzheimer_risk_factor_gene
Alexy	Symptoms
alpha-ketoglutarate dehydrogenase complex	enzyme
alpha-secretase gene	Alzheimer_risk_factor_gene
Alpha2-Macroglobulin	protein
Alzheimer disease	neurodegenerative_disease
ambenonium	drug_used_in_treatment
amitriptyline	drug_used_in_treatment
amoxapine	drug_used_in_treatment

이제 더 이상 ADO의 클래스는 클래스 자체가 아닌 용어로서의 역할을 수행할 수 있으며, 개별 용어에 그 용어에 대한 타입(클래스)이 지정되어 있다. 일반적으로 중요도가 떨어지거나 개체명 인식 및 관계 추출 작업과 별로 상관이 없다고 판단되는 항목은 제거하고, 전체 데이터 중에서 511개 규모의 용어 집합을 선정했으며, 이들에 대한 유형별 개수는 아래 표와 같다.

〈표 5〉 변환 구축된 ADO 기반 용어 집합 유형별 개수

용어 유형	유형 설명	용어 수
Alzheimer_risk_factor_gene	알츠하이머 유발 유전자	140
Amyloid_beta_protein	아밀로이드 베타 단백질	2
Brain_regions	뇌 영역	18
Chemokine	케모카인	18
Cortex	대뇌 및 소내 피질	10
Cytokine	사이토카인	13
Disorder	알츠하이머 관련 장애	17
Drug_used_in_treatment	알츠하이머 치료에 사용되는 약물	136
Enzyme	효소	19
Mental_disorder	알츠하이머 관련 정신 장애	5
Neurodegenerative_disease	신경병성 질환	8
Neuron_as_cellular_entity	세포 개체로서의 뉴런	17
Neurons	알츠하이머 관련 뉴런	17
Processes_related_to_neurons	알츠하이머병 진행과 관련된 뉴런	19
Protein	알츠하이머 관련 단백질	35
Symptoms	알츠하이머 관련 증상	37
합계		511

위의 표에서 알 수 있듯이 알츠하이머병에 위험 인자가 될 수 있는 유전자(140) 명칭과 알츠하이머병 치료에 사용되는 약물(136) 명칭이 가장 많은 수를 차지하고 있다. 그런데 유형에서 보는 바와 같이 일부 의미가 중복되는 경우도 발생한다. 예를 들어, “disorder”와 “mental disorder”는 상하위 관계이다. 또한 “neurons”와 “neuron_as_cellular_entity”도 역시 마찬가지이다. 이는 계층 구조로 되어 있는 ADO의 클래스 구성을 일괄적으로 평탄화한 결과로 야기되는 문제이다. 이러한 부분은 향후 이 사전 자체를 확장하면서 지속적으로 보강해야한다.

본 논문에서 위와 같이 구성된 알츠하이머병 용어 사전(Alzheimer’s Disease Dictionary, ADD) 항목 중에서 “알츠하이머 유발 유전자(Alzheimer_risk_factor_gene)”에 주목하였다. 그 이유는 BioGRID 내에 존재하는 유전자 간 상호작용 정보와 이 유전자 리스트를 비교하여 매핑시킬 수 있기 때문이다. 본 연구에서는 BioGRID 내의 812,281 건의 인간 유전자 간 상호작용(Genetic Interaction) 정보를 가지고 ADO 내의 140개 유전자 명칭과 대조시켜서 알츠하이머병 관련 유전자 간 상호작용 정보를 추출하였다. 우선 활용한 BioGRID 내의 유전자 상호작용 정보의 예를 아래 표에서 보여준다.

<표 6> BioGRID 내의 호모 사피엔스 섹션에 존재하는 Genetic Interactions (일부)

유전자1	유전자2	유전자1의 동의어	유전자2의 동의어
INSL3	MTERF4	RLF|RLNL|ley-I-L	MTERFD2
IQCF3	NDFIP1	-	N4WBP5
KLRF1	GAPDHS	CLEC5C|NKp80	GAPD2|GAPDH-2|GAPDS|HSD-35
KLRF1	TOMM40	CLEC5C|NKp80	C19orf1|D19S1177E|PER-EC1|PEREC1|TOM40
KLRF1	REEP5	CLEC5C|NKp80	C5orf18|D5S346|DP1|TB2|YOP1
KLRF1	CBWD1	CLEC5C|NKp80	COBP
KLRF1	REEP6	CLEC5C|NKp80	C19orf32|DP1L1|TB2L1
LCN15	MYO9B	PRO6093|UNQ2541	CELIAC4|MYR5
LCN15	ALKBH2	PRO6093|UNQ2541	ABH2
LCN15	DIS3L	PRO6093|UNQ2541	DIS3L1
LCN15	NUDT15	PRO6093|UNQ2541	MTH2
LCN15	SETX	PRO6093|UNQ2541	ALS4|AOA2|SCAR1|bA479K20.2
LCN15	ATG4A	PRO6093|UNQ2541	APG4A|AUTL2

총 249,075 개의 상호 작용 항목 각각에는 서로 상호 작용을 일으키는 유전자 쌍과 개별 유전자 명칭에 대한 동의어 리스트가 포함되어 있다. 이들 중에서 한 쌍의 유전자 모두가 앞에서 구축된 ADD의 유전자 엔트리에 포함되어 있다면 이들을 추출하는 작업을 수행하였으며, 그 결과 아래 표와 같이 총 493 건의 상호 작용 정보가 추출되었다.

<표 7> ADO의 유전자 표제어 항목으로 구성된 BioGRID Genic Interactions (일부)

유전자1	유전자2	유전자1에 대한 동의어	유전자2에 대한 동의어
TP53	TP53	BCC7|LFS1|P53|TRP53|	BCC7|LFS1|P53|TRP53|
APEH	APEH	AARE|ACPH|APH|D3F15S2|D3S48E|DNF15S2|OPH|	AARE|ACPH|APH|D3F15S2|D3S48E|DNF15S2|OPH|
CRMP1	CRMP1	CRMP-1|DPYSL1|DRP-1|DRP1|ULIP-3|	CRMP-1|DPYSL1|DRP-1|DRP1|ULIP-3|
CRMP1	DPYSL3	CRMP-1|DPYSL1|DRP-1|DRP1|ULIP-3|	CRMP-4|CRMP4|DRP-3|DRP3|LCRMP|ULIP|ULIP-1|
DPYSL2	DPYSL2	CRMP-2|CRMP2|DHPRP2|DRP-2|DRP2|N2A3|ULIP-2|ULIP2|	CRMP-2|CRMP2|DHPRP2|DRP-2|DRP2|N2A3|ULIP-2|ULIP2|
DPYSL2	DPYSL3	CRMP-2|CRMP2|DHPRP2|DRP-2|DRP2|N2A3|ULIP-2|ULIP2|	CRMP-4|CRMP4|DRP-3|DRP3|LCRMP|ULIP|ULIP-1|
DPYSL2	DPYSL5	CRMP-2|CRMP2|DHPRP2|DRP-2|DRP2|N2A3|ULIP-2|ULIP2|	CRAM|CRMP-5|CRMP5|Ulip6|
FTL	FTL	LFTD|NBIA3|	LFTD|NBIA3|
ATXN1	ATXN1	ATX1|D6S504E|SCA1|	ATX1|D6S504E|SCA1|
BAG2	BAG2	BAG-2|dJ417I1.2|	BAG-2|dJ417I1.2|
NUPL1	NUP62	PRO2463|	IBSN|SNDI|p62|

위 표를 보면 일부 항목들이 동일한 유전자 명칭을 가지고 있음을 알 수 있다(“TP53”, “DPYSL2” 등). 이는 작업 과정에서의 오류가 아니라 BioGRID 데이터베이스 내에 원래부터

동일한 유전자 쌍에 대한 상호 작용이 존재하기 때문이다. 추출된 493개의 유전자 쌍 정보를 수집된 PubMED 데이터베이스의 2012년 전체, 2013년 전체, 2014년 전체, 2105년도 3월까지의 초록 내에 존재하는 문장을 대상으로 검색을 수행하여 최종적으로 상호 작용 포함 문장 집합을 추출하였으며, 그 결과는 다음 절에서 설명한다.

2. PubMed 데이터베이스를 활용한 유전자 간 상호 작용 포함 문장 수집

앞에서도 잠시 언급하였으나 본 논문에서는 서로 상호 작용이 존재하는 한 쌍의 개체가 하나의 문장에 표현되어 있다면 그 문장은 해당 상호 작용을 설명하고 있다는 Distant Supervision 가정(Mintz et al. 2009)을 활용한다. 물론 이 가정은 항상 올바르지는 않으나 특정 분야에서의 관계 추출 학습 집합이 존재하지 않을 경우, 두 개체 간의 관계를 설명하는 문장 인스턴스를 수집하기 위한 효과적인 방법론을 제공하고 있기 때문에 널리 활용되고 있다. 따라서 다음과 같이 두 개의 유전자가 서로 “bind” 상호 작용을 나타내고 있음을 표현하는 문장을 추출할 수 있다.

Interleukin-8 (IL8) receptors IL8RA and **IL8RB** on neutrophil membranes **bind** to **IL8** and direct neutrophil recruitment to sites of inflammation , including acutely injured arteries. (**bind 관계**)

〈그림 3〉 두 유전자 “IL8RB”와 “IL8”을 “bind” 상호 작용으로 표현하는 문장

유전자 간 상호 작용 포함 문장 추출의 세부 과정은 다음과 같다. 우선 대상 문장에 대한 품사 태깅(Part-Of-Speech Tagging)을 수행한 후에 명사구 추출(Noun-Phrase Chunking)을 통해서 용어 후보(Term Candidate)가 될 수 있는 명사구를 모두 추출한다. 추출된 용어 후보가 사전(여기서는 앞에서 추출한 알츠하이머병 관련 유전자 명 사전)에 존재하는지를 검사하고 만일 이 사전에 존재하는 후보라고 판명되면 이 정보를 저장한다. 만일 한 문장에서 두 개의 알츠하이머병 관련 유전자가 존재하고 이들 유전자 한 쌍이 앞에서 추출한 493개의 유전자 쌍 집합에 존재하면 그 문장을 인스턴스 후보로 선택한다. 최종적으로, 본 연구에서 각 연도별로 수집된 알츠하이머병 관련 유전자 상호 작용 포함 문장 집합에 대한 추출 통계 정보는 다음 표와 같다.

〈표 8〉 연도별 유전자 간 상호 작용 포함 문장 추출 통계

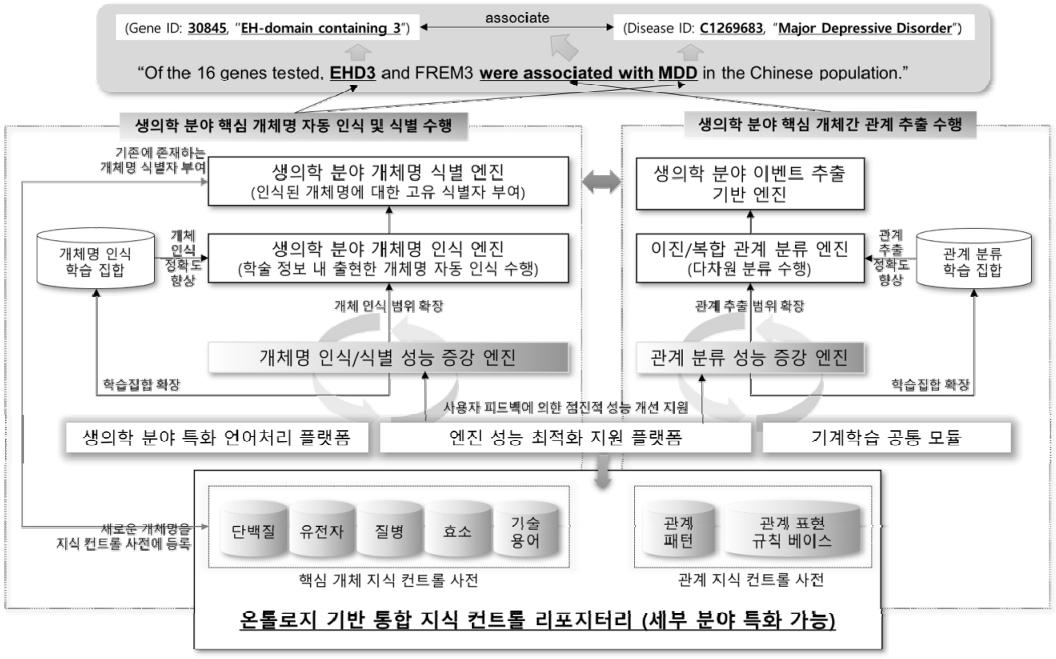
	총 문장 개수	추출된 문장 개수	태깅된 유전자명 개수 (중복 허용)
2012년도	11,257,037	993	2,290
2013년도	12,147,669	1,006	2,298
2014년도	12,782,557	1,242	2,830
2015년도	2,858,298	269	594
전체	39,045,561	**3,510**	**8,012**

일단 추출 대상 문장 수에 비해서 추출된 문장의 개수가 3,510건으로 매우 적음을 알 수 있다. 이는 문장 추출을 위한 기반 유전자 사전의 수가 140개로 매우 적은 것에 기인하며 향후 유전자는 물론 다른 유형(예: 단백질, 증상, 약물 등)들도 포함시키면 그 수가 많이 늘어날 것으로 예상된다. 다시 말해서, 작업의 출발점인 세부 분야 용어 사전의 규모 및 품질이 최종적으로 도출되는 상호 작용 포함 문장 집합의 규모 및 품질을 결정짓는 주요 요인이라고 볼 수 있다. <표 9>에서는 실제 추출된 학습 집합을 정형화된 테이블 형태로 출력한 예를 보인다.

<표 9> 유전자 쌍 및 해당 상호 작용 포함 문장 예시

유전자 1	유전자 2	상호 작용 포함 문장
AhR	CYP1A1	4-ClBQ-induced increase CYP1A1 expression was associated with an increase in the nuclear translocation of **AhR** protein as well as an increase in the luciferase-reporter activity of a human **CYP1A1** xenobiotic response element (XRE) .
PON1	C4a	The combination of C4a, FGA, CP and **PON1** improved slightly the predictive ability of **C4a** alone (AUROC 0.81) .
androgen receptor	TP53	New recurrent alterations have been identified in PCa (e.g., TMPRSS2-ERG translocation, SPOP and CHD1 mutations, and chromoplexy), and many previous ones in well-established pathways have been validated (e.g., **androgen receptor** overexpression and mutations ; PTEN, RB1, and **TP53** loss/mutations) .
AR	CYP17A1	**AR** was examined by targeted sequencing in metastatic tumor biopsies from 18 patients with CRPC who were progressing on a **CYP17A1** inhibitor (17 on abiraterone, 1 on ketoconazole) , alone or in combination with dutasteride, and by whole-exome sequencing in residual tumor in one patient treated with neoadjuvant leuprolide plus abiraterone .
IGF1R	GC	In addition, we identified **IGF1R** as a regulatory target of miR-133a in **GC**.

<표 9>에서 보는 바와 같이, 추출된 문장들은 많은 경우 두 유전자 간의 상호 작용 정보에 대한 표현을 담고 있음을 알 수 있다. 그러나 최종적으로 추출된 알츠하이머병 분야 유전자 간 상호 작용 추출 학습 집합은 수작업 구축 결과에 비해서 100% 정확성을 보장하지는 않으므로 후처리 수동 검증 작업이 필요하다. 특히, 현재까지 개발된 단백질 간 상호작용 자동 추출 시스템들(Choi & Myaeng 2010; Lee, Kim, Lee, Lee, & Kang 2012; Li, Guo, Jiang, & Huang 2014)의 성능이 우수하므로 이를 이용한 반자동 검증(Semi-automatic verification)도 가능하다. 본 논문에서 수행한 연구 결과를 토대로 <그림 4>와 같은 분야별 학습 집합 반자동 구축 지원 플랫폼을 제안한다.



<그림 4> 생의학 분야 관계 추출 학습 집합 반자동 구축 플랫폼 구조

위 <그림 4>에서 보는 바와 같이 기존에 개발된 생의학 분야 개체명 인식 및 관계 추출 엔진을 기반으로 신규로 입력되는 학술 문헌 데이터에 대한 분석이 가능하고 그 결과 특정 문장 내에 존재하는 개체명(단백질, 유전자, 질병, 효소명 등)과 그들 사이의 의미적 연관 관계를 식별할 수 있다. 분야 전문가는 자동 처리 엔진에 의해서 분석 완료된 문장 집합을 검증하고 오류를 수정함으로써 정제된 학습 집합을 효율적으로 구축할 수 있다. 마지막으로 정제된 신규 데이터는 기존 엔진의 성능 개선을 위해서 추가적인 학습 집합으로 활용되어 점진적으로 그 성능을 개선시킨다. <그림 4>에서 보여지는 "개체명 인식/식별 성능 증강 엔진"이나 "관계 분류 성능 증강 엔진"은 위에서 기술한 기능을 수행하는 핵심적인 모듈로서 1장에서 제시한 Annotation 도구를 포함하고 있으며 이 도구를 바탕으로 분야 전문가는 완전히 새로운 문장을 기반으로 학습 집합 구축을 수행하는 것이 아니라 전처리 분석이 완료된 문장들을 검증하고 수정하는 형태로 작업을 수행할 수 있다.

Ⅴ. 결론 및 향후 연구 방향

본 논문에서는 생의학 분야의 특정 세부 분야에 특화된 관계 추출 학습 말뭉치를 효율적으로 구축할 수 있는 시스템을 개발하였다. 개발된 시스템은 대상 분야에 해당하는 용어집(유전

자, 단백질, 질환 명칭 등)을 입력하면, 대용량 상호 작용 데이터베이스를 통해서 이들 용어 간의 연관 관계를 1차적으로 생성하고 생성된 연관 관계 집합을 다시 학술 데이터베이스에서 검색하여 최종적으로 연관 관계 포함 문장을 추출하는 형태로 수행된다. 개발된 시스템의 유효성 검증을 위해서 알츠하이머 병 분야의 유전자 사전을 입력하여 이 분야에 특화된 학습 집합인 유전자 쌍 및 상호 작용 포함 문장 3,510 건을 추출하였다.

향후 연구 과제로서 우선 본 논문에서 반자동으로 구축된 학습 집합에 대한 후처리 검증을 통해서 완성된 형태의 학습 집합을 구성하고 이를 바이오 분야 정보 추출 연구자들에게 공개하여 각 연구자들이 개발한 관계 추출 엔진에 대한 객관적인 비교 평가를 할 수 있도록 할 예정이다. 또한 전체 시스템의 성능 개선을 통해서 알츠하이머 병 분야 외에도 다양한 분야에 대한 학습 집합 구축을 통해서 기계학습 기반 생의학 분야 지식화 연구의 활성화에 지속적으로 기여하고자 한다.

참고 문헌

Alex, B., Grover, C., Haddow, B., Kabadjor, M., Klein, E., Matthews, M., Wang, X. 2008. Assisted Curation: Does Text Mining Really Help?. In *Pacific Symposium on Biocomputing* (Vol. 13, pp. 556-567).

Alnazzawi, N., Thompson, P., & Ananiadou, S. 2014. Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL* (pp. 69-74).

Bader, G. D., Betel, D., & Hogue, C. W. V. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research,* 31(1): 248-250.

Blaschke, C., Hirschman, L., & Valencia, A. 2002. Information extraction in molecular biology. *Briefings in Bioinformatics,* 3(2): 154-165.

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., & Wong, Y. W. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine,* 33(2): 139-155.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., & Cesareni, G. 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Research,* 35(Database issue), D572-D574. https://doi.org/10.1093/nar/gkl950

Choi, S.-P., & Myaeng, S.-H. 2010. Simplicity is Better: Revisiting Single Kernel PPI Extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 206-214). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ding, J., Berleant, D., Nettleton, D., & Wurtele, E. 2002. Mining MEDLINE: abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* 326-337.

Fundel, K., Küffner, R., & Zimmer, R. 2007. RelEx—Relation extraction using dependency parse trees. *Bioinformatics,* 23(3): 365-371. https://doi.org/10.1093/bioinformatics/btl616

Haddow, B., & Alex, B. 2008. Exploiting Multiply Annotated Corpora in Biomedical Information Extraction Tasks. In D. T. Nicoletta Calzolari (Conference Chair) Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Ed.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC' 08).* Marrakech, Morocco: European Language Resources Association (ELRA).

Hastie, T., Tibshirani, R., & Friedman, J. 2009. *The Elements of Statistical Learning.* New York, NY: Springer New York.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Apweiler, R. 2004. IntAct: an open source molecular interaction database. *Nucleic Acids Research, 32*(Database issue), D452-D455. https://doi.org/10.1093/nar/gkh052

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics,* 6(Suppl 1), S1.

Huang, C.-C., & Lu, Z. 2016. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in Bioinformatics,* 17(1): 132-144.

Ivanović, M., & Budimac, Z. 2014. An overview of ontologies and data resources in medical domains. *Expert Systems with Applications,* 41(11), 5158-5166.

Kim, J.-D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N., & Tsujii, J. ichi. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared*

Task 2011 Workshop (pp. 1-6). Stroudsburg, PA, USA: Association for Computational Linguistics.

Krallinger, M., Leitner, F., Rodriguez-Penagos, C., & Valencia, A. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology,* 9(Suppl 2), S4. https://doi.org/10.1186/gb-2008-9-s2-s4

Lee, J., Kim, S., Lee, S., Lee, K., & Kang, J. 2012. High Precision Rule Based PPI Extraction and Per-pair Basis Performance Evaluation. In *Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics* (pp. 69-76). New York, NY, USA: ACM.

Li, L., Guo, R., Jiang, Z., & Huang, D. 2014. Improving Kernel-based protein-protein interaction extraction by unsupervised word representation. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on* (pp. 379-384). IEEE.

Malhotra, A., Younesi, E., G□ndel, M., M□ller, B., Heneka, M. T., & Hofmann-Apitius, M. 2014. ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association,* 10(2), 238-246.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (pp. 55-60).

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. 2009. Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume* 2 (pp. 1003-1011). Stroudsburg, PA, USA: Association for Computational Linguistics.

N□dellec, C. 2005. Learning language in logic-genic interaction extraction challenge. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)* (Vol. 7). Citeseer.

Pyysalo, S., Ginter, F., Heimonen, J., Bj□rne, J., Boberg, J., J□rvinen, J., & Salakoski, T. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics,* 8(1): 50.

Ravikumar, K., Liu, H., Cohn, J. D., Wall, M. E., & Verspoor, K. 2012. Literature mining

of protein-residue associations with graph rules learned through distant supervision. *Journal of Biomedical Semantics,* 3 Suppl 3, S2.

Rubin, D. L., Shah, N. H., & Noy, N. F. 2008. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics,* 9(1): 75-90.

Saffer, J. D., & Burnett, V. L. 2014. Introduction to Biomedical Literature Text Mining: Context and Objectives. In *Biomedical Literature Mining* (pp. 1-7). Springer.

Segura Bedmar, I., Mart□nez, P., & S□nchez Cisneros, D. 2011. The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research,* 34(Database issue), D535-539.

Thompson, P., Iqbal, S. A., McNaught, J., & Ananiadou, S. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics,* 10(1): 349.

Uzuner, □., South, B. R., Shen, S., & DuVall, S. L. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association: JAMIA,* 18(5): 552-556.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., & Eisenberg, D. 2000. DIP: the Database of Interacting Proteins. *Nucleic Acids Research,* 28(1), 289-291.

박경미, 황규백. 2011. 자연어처리 기반 바이오 텍스트 마이닝 시스템.『정보과학회논문지 : 컴퓨팅의 실제 및 레터』, 17(4).

정창후, 최성필, 이민호, 최윤수. 2010. 기술용어 간 관계추출의 성능평가를 위한 반자동 테스트 컬렉션 구축 프레임워크 개발.『한국콘텐츠학회논문지』, 10(2).

최성필. 2016. 기계 학습을 이용한 바이오 분야 학술 문헌에서의 관계 추출에 대한 실험적 연구.『한국문헌정보학회지』, 50(2).

허고은, 송민. 2014. 텍스트 마이닝 기반의 그래프 모델을 이용한 미발견 공공 지식 추론.『정보관리학회지』, 31(1).

국한문 참고문헌의 영문 표기
(English translation / Romanization of reference originally written in Korean)

Park, Kyung-Mi, Kyu-Baek Hwang. 2011, A Bio-Text Mining System Based on Natural Language Processing. *KIISE Transactions on Computing Practices,* 17(4).

Jeong, Chang-Hoo, Sung-Pil Choi, Min-Ho Lee, Yun-Soo Choi. 2010. *The Journal of the Korea Contents Association.* 10(2).

Choi, Sung-Pil. 2016. An Experimental Study on the Relation Extraction from Biomedical Abstracts using Machine Learning. *Journal of the Korean Society for Library and Information Science.* 50(2).

Heo, Go Eun, Min Song. 2014. Inferring Undiscovered Public Knowledge by Using Text Mining-driven Graph Model. *Journal of the Korean Society for Information Management.* 31(1).