생의학 분야 학술 논문에서의 개체명 인식 및 관계 추출을 위한 언어 자원 수집 및 통합적 구조화 방안 연구*

A Study on Collecting and Structuring Language Resource for Named Entity Recognition and Relation Extraction from Biomedical Abstracts

강 슬 기 (Seul-Ki Kang)**
최 윤 수 (Yun-Soo Choi)***
최 성 필 (Sung-Pil Choi)****

목 차

초 록

본 논문에서는 급격히 증가하는 생의학 분야 비정형 텍스트에서 핵심적 내용을 추출할 수 있는 기계학습 기반 정보 추출 시스템을 구축하기 위한 언어자원 수집 및 통합적 구조화 방안을 제안한다. 제안된 방법은 정보 추출 시스템을 크게 개체명 인식과 개체명 간 관계 추출 시스템으로 구분하고, 각각의 시스템에 적합한 학습데이터를 구성하기 위해 생의학 분야 개체명 사전과 학습 집합을 수집한다. 그리고 수집된 해당 자원들의 특성을 분석하여 개체 구별을 위해 필수적으로 포함시켜야 할 항목들을 도출하고 이를 통해 시스템 학습과정에서 사용될 학습 데이터를 구성하기 위한 항목을 선정한다. 이와 같이 선정된 학습데이터의 구성 내용에 따라 수집된 자원들을 가공하여 학습 데이터를 구축한다. 본 연구에서는 생의학 분야의 하위 분야인 유전자, 단백질, 질병, 약물 4개 분야에 대한 개체명 사전과 학습 집합을 수집하여 각각을 학습 데이터로 구축하였으며, 개체명 사전을 통해 구축된 개체명 인식용 학습 데이터를 대상으로 개체명 수용 범위를 측정하기 위한 검증 과정을 수행하였다.

ABSTRACT

This paper introduces an integrated model for systematically constructing a linguistic resource database that can be used by machine learning-based biomedical information extraction systems. The proposed method suggests an orderly process of collecting and constructing dictionaries and training sets for both named-entity recognition and relation extraction. Multiple heterogeneous structures for the resources which are collected from diverse sources are analyzed to derive essential items and fields for constructing the integrated database. All the collected resources are converted and refined to build an integrated linguistic resource storage. In this paper, we constructed entity dictionaries of gene, protein, disease and drug, which are considered core linguistic elements or core named entities in the biomedical domains and conducted verification tests to measure their acceptability.

키워드: 정보 추출, 개체명 인식, 관계 추출, 바이오 텍스트 마이닝, 학습 집합
Information Extraction, Named-Entity Recognition, Relation Extraction, Bio-text Mining, Training Set

---

* 본 연구는 한국과학기술정보연구원 주요사업 "초고성능컴퓨팅 기반 건강한 고령사회 대응 빅데이터 기술개발" 과제의 연구비 지원으로 수행되었음(K-17-L03-C02).
** 경기대학교 일반대학원 문헌정보학과 석사과정(rkdtmfrl1007@kyonggi.ac.kr) (제1저자)
*** 한국과학기술정보연구원 생명의료융합기술연구실 책임연구원(armian@kisti.re.kr) (공동저자)
**** 경기대학교 문헌정보학과 조교수(spchoi@kgu.ac.kr) (교신저자)
논문접수일자: 2017년 10월 23일 최초심사일자: 2017년 10월 23일 게재확정일자: 2017년 11월 18일
한국문헌정보학회지, 51(4): 227-248, 2017. [http://dx.doi.org/10.4275/KSLIS.2017.51.4.227]

1. 서 론

최근 생의학 분야 연구가 급격히 진보하면서 생의학 분야의 다양한 개념 및 이론들이 등장하였다. 이에 따라 학술 정보 및 기술 문헌 등 생의학 분야 비정형 텍스트의 양이 빠른 속도로 증가하고 있다(박경미, 황규백 2011). 그러나 비정형 학술 정보를 정형화하여 체계적으로 제시할 수 있는 학술 지식 서비스가 부족하여 해당 분야 연구자들이 선행 연구 및 연구 동태 파악에 많은 시간을 소비하고 있다.

이러한 문제를 개선하기 위해서는 비정형 텍스트 자원을 지식화할 수 있는 생의학 분야 핵심 개체 및 관계 추출 시스템이 필요하다(Ananiadou et al. 2006). 비정형 텍스트에서 핵심 개체를 인식하고 추출하기 위한 자연어 처리 기술은 주로 규칙 기반 접근법과 통계 기반 접근법을 통해 이루어져 왔으며, 최근에는 기계학습 방법 중 하나인 딥 러닝 기술에 기반 한 자연어 처리 연구가 활발히 진행되고 있다(이혜진, 김재웅 2017).

이러한 기계학습 기반의 자연어 처리 성능을 안정적으로 향상시키기 위해서는 정형화된 다량의 언어자원(임재현 2016) 즉, 학습데이터를 통해 모델이 어휘적, 구문적, 의미적 자질들을 학습하는 과정이 필요하다(박경미, 황규백 2011). 그러나 현재 언어자원을 수집하고 관리하기 위한 일련의 과정 및 체계가 존재하지 않으며, 이에 대한 연구 또한 부족한 상황이다.

따라서 본 논문에서는 기계학습 기반의 생의학 분야 개체명 인식 및 개체 간 관계 추출 시스템을 위한 언어자원의 수집 과정과 데이터 정형화를 위한 체계를 제시함으로써 이를 통해 범용 주제 분야 기계학습 기반 자연어 처리 시스템에 적용할 수 있는 언어자원의 통합적 저장・관리 체계를 구축하기 위한 방안을 제시하고자 한다.

2. 관련 연구

현재 외국에서는 생의학 분야 심층 지식의 중요도와 그 활용성을 인지하고 세부 분야별 지식베이스를 효율적으로 구축하기 위한 다차원의 시도를 진행 중에 있다(Huang and Lu 2016). 생의학 분야 문헌을 대상으로 핵심 정보를 추출하는 바이오텍스트 마이닝을 전문으로 하고 있는 기관들은 보다 다양한 모습의 자동화 도구들을 개발 및 제작하고 있으며, 오랜 연구 끝에 일정 수준 이상의 성능을 보여주는 기반 엔진들이 개발되고 있는 상황이다(Choi 2016).

대부분의 연구는 주로 BioNLP 분야의 다양한 대회들을 통해 연구 성과가 발표되고 있으며(Kim et al. 2013), 대표적인 대회로는 ACE Meta-Knowledge, Anatomy Corpora, BioCause 등을 들 수 있다. 이러한 대회들은 시스템 평가를 위한 컬렉션들을 지속적으로 구축하고 배포하고 있으며(The National Centre for Text Mining 2016), 생의학 분야 연구에 있어 기반 자원들로 활용되어 높은 수준의 연구 성과들이 보고되고 있다.

국내에서도 범용 분야의 언어처리 분야에서 딥 러닝 기법을 활용하는 사례들이 증가하고 있다(박성배 2005). 대표적인 예로 한의약 분야에서는 한의학 학술 문헌에 대한 딥 러닝 기반 자연어 처리 과정을 도입하여 한약진흥재단에

서 인공지능을 활용한 한의임상정보학 포럼을 갖는 등 언어처리에 대한 연구가 활발히 진행되고 있다. 그러나 생의학 분야의 경우, 사회적 관심과 정부 주도의 연구 개발 투자에도 불구하고 바이오 분야 심층 지식베이스의 구축 및 연계, 활용 분야에 적극적인 지원 및 투자가 이루어지지 않고 있다. 따라서 유관분야 연구는 KAIST의 자연어 처리 연구실에서 소규모로 진행되고 있는 상황이다.

특히 생의학 분야에서는 개체명과 개체명 간의 상호관계를 통해 유관 연구를 위한 새로운 가설을 발견하고 예측할 수 있기 때문에 이러한 유기체간 상관관계를 파악하는 것이 매우 중요하다(허고은, 송민 2014). 따라서 핵심 개체를 추출하고 추출된 핵심 개체 간의 관계를 파악하는 정보 추출 시스템이 개발된다면 생의학적 비정형 텍스트의 핵심적 내용을 보다 효과적으로 제시할 수 있게 된다(Jensen et al. 2006, 119-129). 이러한 시스템을 위해서는 우선적으로 핵심 개체명을 인식하고 개체명 간 관계 추출하기 위한 시스템 구축에 기반이 될 수 있는 학습 데이터의 구성이 필요하다. 학습데이터는 기계 학습에서 모델의 성능을 향상시키기 위한 모델의 학습 과정에 필요한 데이터 집합으로 각각의 데이터는 일관된 양식으로 작성되어야 한다.

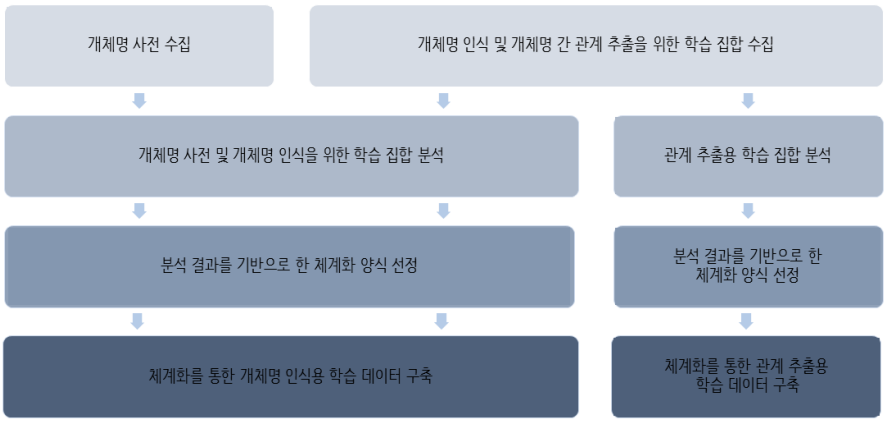
개체명 인식에서의 학습 데이터 구성에 대한 중요성은 지속적으로 언급되어 왔다. 송영길은 크게 규칙기반 방법과 확률 기반 방법으로 연구되고 있는 개체명 인식에서 학습 데이터로 사용되는 개체명 사전이 성능향상에 중요한 역할을 한다고 언급하면서, 위키피디아를 기반으로 개체명 사전을 반자동으로 구축하는 방법을 제안했다(송영길 외 2015). 신성호 외 6인은 개체명 인식에서 학습 모델과 함께 또 다른 한 축을 이루는 요소인 학습 집합의 품질이 개체명 인식의 정확도에 매우 많은 영향을 미친다고 언급하며 학습 집합의 중요성을 강조했다(신성호 외 2014).

그러나 현재 학습 데이터를 구성하기 위한 언어자원의 수집 과정 및 학습 데이터 정형화를 위한 체계가 부족한 상황이다. 따라서 유관분야 연구에 있어 학습 데이터의 구축을 위한 시간 및 비용이 많이 소요되고 있으며, 연구자들이 보다 높은 수준의 연구를 진행하는데 어려움을 겪고 있다. 따라서 본 논문에서는 정보추출 시스템을 위한 학습 데이터를 구축하는 과정을 체계화하기 위한 방안을 제안한다. 제안된 방법은 학습 데이터 구축을 위한 체계화된 통합적 저장・관리 방안으로써 언어자원을 수집하고 해당 자원들의 특성을 분석하여 정보추출 시스템에 적합한 형태를 지니는 학습 데이터로 구축하는 모든 과정을 포함한다.

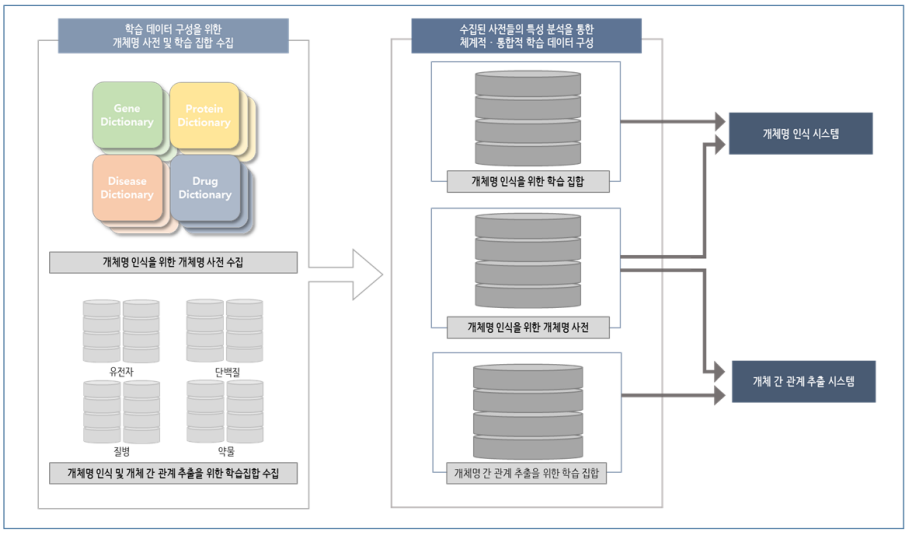
3. 정보추출을 위한 언어 자원 구조화 모형

본 논문에서는 생의학 분야 핵심 개체명 인식 및 개체명 간 관계 추출 시스템 구축에 필수적인 학습 데이터 구축 과정을 체계화하기 위한 유관 분야 언어 자원의 수집 및 통합적 저장・관리 방안을 제안한다.

개체명 인식 및 개체명 간 관계 추출 시스템은 최근 언어처리에 있어 성능향상의 가능성을 인정받은 기계학습 기법을 기반으로 이루어진



〈그림 1〉 학습 데이터 구축을 위한 생의학 분야 언어 자원의 수집 및 통합적 저장 · 관리 체계



〈그림 2〉 생의학 분야 핵심 개체 추출 및 개체 간 관계 추출 시스템 개요

다. 앞서 언급한 바와 같이 기계 학습 기반의 정보 추출 성능을 안정적으로 향상시키기 위해서는 다량의 언어자원을 일관된 형식으로 정형화한 학습데이터를 통한 모델 학습이 필수적이다.

따라서 본 논문에서는 생의학 분야 개체명 인식 시스템과 개체명 간 관계 추출 시스템을 위해 다량의 유관분야 언어자원을 학습 데이터로 구축하는 과정을 통합적으로 체계화하는 방안을 제안한다. 또한 제안된 방법을 통해 실제로 학습 데이터를 구축하는 과정과 구축된 학습 데이터의 일부를 검증한 결과를 함께 제시한다. 학습 데이터는 각각의 시스템에 따라 적합한 체계화 양식을 선정하고, 이에 따라 언어자원들을 체계화 하는 과정을 통해 구축된다.

그러나 구축된 각각의 학습 데이터는 전처리 모듈의 수정을 통해 실제적 시스템 학습 과정에서 개체명 인식 시스템을 위한 학습 데이터가 개체명 간 관계 추출 시스템의 학습에 사용될 수 있으며, 관계 추출용 학습 데이터 역시 개체명 인식 시스템의 학습에 사용될 수 있다.

4. 연구 방법

본 논문에서는 기계학습 기반의 생의학 분야 개체명 인식 및 개체명 간 관계 추출 시스템을 구축하기 위한 학습데이터를 체계적으로 정형화 할 수 있는 통합적 저장・관리 방안을 제안한다. 제안된 방법은 학습 데이터 구성과정을 언어 자원의 수집, 수집된 자원의 분석, 분석을 통한 체계화 양식 선정, 자원의 체계화 및 검증 단계로 구분한다.

먼저, 언어 자원의 수집 단계는 시스템이 학습에 사용할 학습데이터를 구성하는데 기반이 될 언어자원들을 수집하는 단계이다. 수집된 언어자원들은 일련의 과정을 통해 정형화되어 학습 데이터로 구축되고 시스템 학습에 활용된다. 본 연구에서는 개체명 인식 시스템을 위한 학습데이터 구성을 위해 다량의 개체명을 가장 명확하게 표현하고 있는 개체명 사전과 개체명 인식 학습 집합을 언어자원으로 선정하여 수집하였다. 또한, 개체명 간 관계 추출 시스템을 위한 학습데이터를 구성하기 위해 개체명 간의 관계를 표현하는 언어자원인 개체명 간 관계 정의 코퍼스를 선정하여 수집하였다. 개체명 간 관계 정의 코퍼스는 개체1과 개체2를 각각 기재하고 개체들 간의 관계를 TRUE/FALSE 혹은 Positive/Negative의 형태로 표현하는 언어자원이다.

다음으로 수집된 자원의 분석 과정은 앞서 수집된 언어자원들을 대상으로 각각의 항목과 내용을 분석하는 단계이다. 해당 단계에서는 항목 및 내용 분석을 통해 개체명을 식별하는데 반드시 포함되어야 하는 항목과 여러 자원들에서 공통적으로 포함하고 있는 항목들을 도출한다. 만약 분석 결과가 잘못되거나 주요 항목들이 누락될 경우 학습 데이터의 내용이 불완전해지거나 신뢰할 수 없는 데이터가 될 수 있으므로 철저한 분석이 필요하다.

분석을 통한 체계화 양식 선정 단계는 이전 단계에서 도출된 중요 항목들을 통해 학습 데이터의 구성 양식을 선정하는 단계이다. 해당 단계에서는 각각의 개체명 및 개체명 간 관계들을 식별하기 위한 식별자 부여 방법과 각 항목을 구별하는 방법, 항목의 내용을 구성하고 표현하는 양식을 선정하는 방법 등 학습 데이터를 구성하는데 필요한 모든 양식들을 선정한다.

마지막으로 자원의 체계화 및 검증 과정을 통해 수집된 언어자원들을 선정된 체계화 양식에 따라 가공하여 학습 데이터를 구축한다. 구축된 학습 데이터는 해당 분야의 개체명 및 개체명 간 관계를 얼마나 수용하고 있는지를 확인하기 위한 검증 단계를 거친다. 본 연구에서 구축된 학습 데이터는 구축과정에서 활용된 수집 자원의 유형에 따라 개체명 사전을 통해 구축한 학습 데이터와 코퍼스 자원을 통해 구축한 학습 데이터로 구분할 수 있다. 코퍼스 자원을 통해 구축한 학습 데이터의 경우 수집된 코퍼스 자원이 구축 기관의 검증 과정을 거쳐 배

포되고 있고 이미 다양한 연구에서 활용되고 있기 때문에 데이터 검증 대상에서 제외하고, 개체명 사전을 통해 구축한 학습 데이터만을 대상으로 검증 과정을 실시하였다. 데이터 검증은 유관 분야의 논문 초록을 수집하고, 수집된 데이터를 대상으로 학습 데이터에 존재하는 개체명을 검색하는 개체명 완전 일치 방법을 통해 수행되었다.

5. 데이터 구축 및 분석

본 연구에서는 기계학습 기반의 생의학 분야 정보 추출 시스템을 위한 학습 데이터를 구축하는 과정을 체계적으로 정형화할 수 있는 통합적 저장・관리 방안을 제안한다.

학습 데이터 구성을 위한 언어자원의 수집은 생의학 분야의 특화 분야로 다양한 BioNLP 대회에서 관계 추출 성능 경쟁 및 평가의 대상이 되며, 생의학 분야 개체명 인식 및 개체 간 관계 추출이 가장 활발하게 진행되고 있는 Gene(유전자), Protein(단백질), Disease(질병), Drug(약물) 4개의 분야를 대상으로 하였다. 본 연구에서 학습 데이터 구성을 위해 수집된 언어 자원은 자원 형식에 따라 개체명 사전과 코퍼스 자원으로 구분되며 코퍼스 자원은 다시 개체명 인식용 학습 집합과 관계 추출용 학습 집합으로 구분된다. 제안된 학습 데이터의 구축 과정을 통해 수집된 개체명 사전과 개체명 인식용 학습 집합은 개체명 인식 시스템을 위한 학습 데이터로, 관계 추출용 학습 집합은 개체명 간 관계 추출 시스템을 위한 학습 데이터로 구축되었다.

5.1 생의학 분야 개체명 사전을 통한 학습 데이터 구축

개체명 인식 시스템을 위한 학습 데이터 구축을 위해 수집한 유관 분야 사전데이터는 총 5개이다. 수집된 사전 데이터는 모두 자원 분석 단계, 체계화 양식 선정 단계, 체계화 및 검증 단계를 통해 학습 데이터로 구축하였다.

5.1.1 개체명 사전 자원 수집

개체명 추출 모델을 위한 통합적 기반 확보를 위한 생의학 분야 개체명 사전 수집 결과는 〈표 1〉과 같이 유전자 분야에서 2개, 단백질, 질병, 약물 분야에서 각각 1개의 개체명 사전이 수집되었다.

5.1.2 개체명 사전 자원의 분석

본 연구를 위해 수집된 사전들은 구축한 기관 및 단체의 기준에 따라 생성된 데이터로 각각 내용의 구성 및 포함 범위, 데이터 형식 등이 모두 다르다. 따라서 학습데이터로써의 통일성을 확보하고, 실제 모델 학습 단계에서 전처리 모듈에 대한 용이성을 확보하기 위해 학습 데이터의 양식을 통일하는 가공 과정이 필요하다.

따라서 수집된 개체명 사전들을 대상으로 데이터 항목 및 내용에 대한 세부 분석을 수행하였으며, 데이터 항목 분석을 통해 개체명 사전에서 공통적으로 포함되는 항목과 개체명 식별을 위해 반드시 필요한 항목들을 도출한 결과는 〈표 2〉와 같다.

또한 데이터 내용 분석 결과 유전자 사전과 단백질 사전 내에서 개체명간 중복이 발생하는

〈표 1〉 수집된 생의학 분야 개체명 사전 데이터

분야	개체명 사전명	사전 정보	개체명 수
유전자	HGNC dataset	인간 유전자 명세에 대한 표준을 지정하는 유전자 명명 위원회(HUGO)가 제공하는 유전자 기호 및 이름에 대한 리스트(HUGO Gene Nomenclature Committee 2017).	42,077
	*CTD_genes vocabulary	MDI Biological Laboratory & NC State University에서 구축한 유전자 및 질병 어휘 데이터베이스(Comparative Toxicogenomics Database 2017).	457,511
단백질	**SwissProt	Universal Protein Resource(UniProt)에서 구축한 단백질과 관련된 모든 정보들을 포함하는 데이터베이스(UniProt 2017).	554,241
질병	Disease vocabulary (MEDIC)	CTD(Comparative Toxicogenomics Database)에서 제작한 미국 국립의학 도서관의 의학분야 주제명 표목(MeSH) 중 "Disease" 분야의 하위 항목과, Online Mendelian Inheritance in Man(OMIM) 중 유전병에 대한 데이터(Comparative Toxicogenomics Database 2017).	11,865
약물	DrugBank database	캐나다 정부의 지원을 받는 최첨단 대사 연구 시설인 'TMID(Metabolomics Innovation Center)'에서 구축한 오픈 소스 데이터(DrugBank 2017).	8,283

* 2017년 2월 업데이트 된 데이터를 수집
** 2017년 6월 29일 업데이트 된 데이터를 수집

〈표 2〉 수집된 개체명 사전 분석을 통한 공통 항목 도출

	유전자		단백질	질병	약물
	HGNC dataset	CTD_genes vocabulary	SwissProt	Disease vocabulary (MEDIC)	DrugBank database
Name	symbol	GeneSymbol	DE-RecName	DiseaseName	Common name
Synonyms	name	GeneName	DE-Short	Synonyms	Synonyms
	alias_name	Synonyms	DE-AltName	-	-
	-	-	DE-Short	-	
ID_정보원명	hgnc_id	GeneID	ID	-	DrugBank ID
		AltGeneIDs			
ID_기관명	uniprot_ids	BioGRIDIDs	-	DiseaseID	CAS
	-	PharmGKBIDs	-	AltDiseaseIDs	UNII
	-	UniProtIDs	-	ParentIDs	Standard InChI Key
	-	-	-	TreeNumbers	
Reference	-	-	GN-Name	-	-
	-	-	GN-OS(common)	-	-

것을 확인하였으며, 개체명 간 중복이 발생하는 원인을 파악하기 위해 각 분야에서의 개체명 부여 방법을 추가적으로 조사하였다. 유전자 사전에서 발생하는 개체명간 중복은 약 6만 건으로 이는 Gene Name에 유전자군을 총칭하는 용어인 tRNA(Beuning and Musier-Forsyth 1999)[1] 와

1) tRNA는 단백질 합성 시 상보적인 안티코돈을 가지고 있어 mRNA에 해당 아미노산을 운반해 주는 RNA를 총칭하는 개념.

ncRNA(Tripathi et al. 2010)[2]가 Gene Name에 포함된 결과이다. 따라서 해당 중복을 줄여 개체명 식별을 보다 용이하게 하기 위하여 국제유전학회의 유전자 명명법 가이드라인에 따라 유전자에 부여되는 학술적 측면의 유전자 이름인 유전자 기호(Gene Symbol)를 개체명으로 활용하는 방안을 마련하였다.

또한, 단백질 사전에서 확인된 개체명 중복은 단백질 개체가 발현되는 유기체에 따라 구별되며, 발현되는 유기체가 다를 경우 같은 이름의 개체라도 다른 개체로 구별되는 단백질 분야의 특성에 따른 결과이다. 따라서 유기체를 기준으로 2차 세부 내용 분석을 실시하였으며, 분석 결과 발현 유기체가 동일한 경우에는 단백질 개체명에 대한 중복이 없는 것을 확인하였다. 이러한 단백질 분야의 특성을 반영하고, 발현 유기체에 따른 단백질 개체를 구별하기 위하여, 참고사항 필드를 추가하여 단백질의 발현 유기체를 표현하는 방안을 마련하였다.

5.1.3 개체명 사전의 체계화 양식 선정

개체명 사전 자원의 항목 분석 결과 수집된 개체명 사전에서 개체명, 개체명의 동의어, 사전을 구축한 정보원에서 부여한 ID, 기타기관에서 부여한 개체 ID등을 공통 항목으로 도출하였으며, 데이터 내용의 분석을 통해 개체명 간 중복을 확인하고 개체명 식별을 보다 용이하게 할 수 있는 중복 제거 방안을 마련하였다. 이러한 과정을 통해 도출된 항목 및 주제 분야별 특성을 반영하여 선정한 개체명 사전의 체계화 양식은 〈그림 3〉과 같다. 〈그림 3〉의 '\t'는 아스키 문자 집합에서 정의된 제어부호인 이스케이프 문자로, 탭을 의미하며 학습 데이터에서 항목 간 구분을 위해 사용하였다.

첫 번째 항목인 'ID_ICL'은 사전 데이터 가공 시 부여한 자체식별자로 필수적으로 부여해야 하는 항목이다. 항목을 필수적으로 부여한다는 것은 기재할 내용이 없을지라도 항목을 구성하고 그 내용을 빈칸으로 둔다는 것을 의미한다. 자체식별자는 선정된 가공 양식에 따라 사전 데이터를 가공하는 과정에서 다양한 주제 분야 개체명 사전을 통합적으로 관리하기 위한 일관성을 확보하고 각각의 개체명 식별을 용이하게 하기 위해 〈표 3〉과 같이 부여하였다.

식별자의 첫 글자인 D는 사전(Dictionary)을 의미하는 것이며, 두 번째로 나타나는 GE, PR, DI, DR은 각 개체명이 포함되는 주제를 쉽게 파악할 수 있도록 각 주제의 영문명 2글자를 부여한 결과이다. 마지막으로 기재되는 숫자는 사전에서 개체가 등장하는 순서에 따라 부여하였다.

두 번째 항목인 'Name'은 개체명을 의미하며 필수적으로 부여해야 한다. 특히 본 연구에서 구축하는 학습 데이터는 개체명 인식 시스템이 개체명을 학습하는데 기반이 되는 데이터

ID_ICL \t Name \t Synonyms_1 | Synonyms_2 … \t ID_정보원명 \t ID_기관명1 \t ID_기관명2 \t … Reference

〈그림 3〉 개체명 사전 데이터의 체계화 양식

2) ncRNA는 비번역 RNA로 생체에서 발현한는 RNA 중에서 단백질을 번역하지 못하는 RNA를 총칭하는 개념.

〈표 3〉 개체명 사전의 자체 식별자 부여 양식

분야	식별자 부여 양식
유전자	D_GE1 ~ D_GExxxx (개체 끝 번호)
단백질	D_PR1 ~ D_PRxxxx (개체 끝 번호)
질병	D_DI1 ~ D_DIxxxx (개체 끝 번호)
약물	D_DR1 ~ D_DRxxxx (개체 끝 번호)

이므로 가장 중요한 항목이라고 할 수 있다. 'Synonyms'는 개체명에 대한 동의어 및 관련어, 약어를 기재하는 항목으로 역시 필수적으로 부여해야 한다. 'Name'과 함께 개체명으로 사용될 수 있는 내용이기 때문에 수집 사전에 존재하는 동의어 및 관련어, 약어를 모두 기재한다. 해당 내용이 1개 이상 존재할 경우 'Synonyms' 항목 내에 모드 기재하되 각 내용을 구별하기 위해 내용 사이에 '|'를 기재한다. 다음 항목인 'ID_정보원명'은 개체명 사전을 구축한 기관 및 단체에서 부여한 ID이며, 선택적으로 부여할 수 있다. 만약 1개 이상의 ID가 있을 경우에는 1개 이상의 'ID_정보원명' 항목이 부여된다. 'ID_기관명'은 기타 기관에서 부여한 개체의 ID로 일반적으로 유관분에서 통용되는 ID를 의미한다. 'ID_기관명'은 선택적으로 작성할 수 있으며, 'ID_정보원명'과 마찬가지로 여러 번 부여되고 작성될 수 있다. 'Reference'는 각 주제 분야 및 사전에 따라 존재하는 부가적인 참고사항을 의미하며 선택적으로 부여할 수 있다.

5.1.4 개체명 사전의 체계화를 통한 학습 데이터 구축

선정된 체계화 양식에 따라 수집된 개체명 사전을 체계화 한 결과 구축된 학습 데이터의 구성 양식 및 포함하고 있는 개체명 수는 〈표 4〉와 같다.

사전 데이터의 가공 결과 유전자, 단백질, 질병, 약물 분야에서 각각 9개, 7개, 6개, 6개의 필드가 구성되었으며, 개체명 수는 유전자 499,507개, 단백질 554,241개, 질병 11,865개, 약물 8,283개를 포함한다.

5.2 생의학 분야 정보 추출용 학습 집합 수집 및 구축

본 연구에서 수집한 생의학 분야 정보 추출용 학습 집합은 크게 개체명 인식용 학습 집합과 개체명 간 관계 추출용 학습 집합으로 구분된다. 여기서 학습 집합은 특정 문장에서 하나의 개체명을 추출하여 개체의 특성을 나타내거나, 두 개의 개체명을 추출하여 두 개체명 간의 관계를 나타내는 언어자원인 코퍼스(Corpus, 말뭉치) 자원을 의미한다. 코퍼스 자원은 각 구축 기관에서 데이터 검증 과정을 거쳐 구축하고, 배포하고 있으며 다양한 연구에서 이미 활용되고 있는 언어자원이다. 수집된 학습 집합 중에서 개체의 특성을 나타내는 학습 집합을 개체명 인식 코퍼스라고 하며, 개체명 간 관계를 나타내는 학습 집합을 관계 정의 코퍼스라고 한다. 해당 코퍼스들은 각각 개체명 인식 시스템을 위한 학습 데이터와 개체명 간 관계 추출 시스템을 위한 학습 데이터로 구축하는데 활용되었다.

〈표 4〉 개체명 사전 체계화를 통해 구축된 학습 데이터의 구성 양식 및 개체명 개수

주제	사전의 필드명	개체명 수
유전자	ID_ICL \t Name \t Synonyms \t ID_HGNC \t ID_CTD \t ID_AltCTD \t ID_BioBRID \t ID_PharmGKB \t ID_Uniprot	499,507
단백질	ID_ICL \t Name \t Synonyms \t ID_SwissProt \t ID_EntryName \t GeneName (Reference 필드) \t Organism(Reference 필드)	554,241
질병	ID_ICL \t Name \t Synonyms \t ID_MEDIC \t ID_MESH \t ID_OMIM	11,865
약물	ID_ICL \t Name \t Synonyms \t ID_DrugBank \t ID_CAS \t IT_UNII \t ID_InChi	8,283

5.2.1 개체명 인식을 위한 학습 데이터 구축

개체명 인식 시스템을 위한 학습 데이터는 수집된 개체명 인식 코퍼스들을 기반으로 구축되었으며, 수집된 개체명 인식 코퍼스 자원의 수는 총 3개이다. 수집된 자원들은 모두 자원 분석, 체계화 양식 선정, 체계화 단계를 통해 학습 데이터로 구축되었다.

(1) 개체명 인식 코퍼스 자원 수집

개체명 인식용 학습 데이터 구축을 위해 수집된 개체명 인식 코퍼스는 유전자 분야에서 1개, 질병분야에서 2개이며 각 코퍼스에 대한 정보 및 포함하고 있는 개체명 개수는 〈표 5〉와 같다.

(2) 개체명 인식 코퍼스 자원의 분석

수집된 3개의 개체명 인식 코퍼스를 대상으로 데이터 항목 및 내용에 대한 세부 분석을 진행하였다. 데이터 항목 분석 결과 개체명 인식 코퍼스는 일반적으로 개체명과 개체명의 동의어, 개체의 유형 정보, 개체를 추출한 문장, 개체를 추출한 문서, 개체를 추출한 문장 및 문서의 PubMed ID와 개체의 위치 정보 등을 포함하고 있는 것을 확인하였다.

또한 데이터 내용 분석을 통해 수집된 코퍼스들이 각각 문장, 문서, 문단 등 서로 다른 기준에 따른 위치(Offset)정보를 제공하는 것을 〈표 6〉과 같이 확인하였다. 따라서 위치 정보에

〈표 5〉 수집된 생의학 분야 개체명 인식 코퍼스 자원

분야	코퍼스명	코퍼스 정보	개체명 수
유전자	Genia Term Annotation collection	BioNLP 분야 대회를 통해 공개된 시스템 평가를 위한 컬렉션으로 실제 문장 정보와 문장내 유전자 개체명 정보를 태깅한 XML 데이터(GENIA: The BioNLP Shared Task 2016 2017)	62,137
질병	NCBI_Corpus	미국 국립 생물공학 정보센터에서 PubMed 초록에 등장하는 질병 명에 수동으로 주석을 달아둔 데이터(National Center for Biotechnology Information 2017).	6,881
	Arizona Disease Corpus (AZDC)	애리조나 주립 대학의 디에고(DIEGO) 연구실에서 PubMed 초록에 등장하는 질병 명에 수동으로 주석을 달아둔 데이터(Biomedical Informatics Lab at ASU 2017).	4,522

〈표 6〉 개체명 인식 코퍼스의 Offset 제공 기준

분야	정보원명	제공되는 offset 기준
질병	NCBI	문서
	AZDC	문장

대한 표현 양식을 통일하기 위하여 위치정보를 추출하는 기준을 문장단위로 선정하였으며, 이에 따라 분석 과정에서 위치정보를 재추출하는 과정이 추가로 수행되었다.

(3) 개체명 인식 코퍼스 자원의 체계화 양식 선정

개체명 인식 코퍼스 자원을 대상으로 항목을 분석하여 도출한 공통 항목들은 개체명, 개체명의 동의어, 개체의 유형, 개체를 추출한 문장, 문장에 대한 PubMed ID, 문장 내 개체의 위치이다. 따라서 이를 기반으로 〈그림 4〉와 같이 개체명 인식 코퍼스 자원의 체계화 양식을 선정하였다.

첫 번째 항목인 'EntityName'에는 개체명을, 'Synonyms'에는 개체명에 대한 동의어 및 관련어, 약어 등을 기재한다. 동의어 및 관련어, 약어 등이 다수 존재할 경우 항목 내에 모두 기재하되 '|'를 통해 구별한다. 'Entity_Type'에는 개체의 유형을 기재하며, 'Offset'에는 분석과정에서 재추출한 문장단위 기준의 위치정보를 기재한다. 'PMID' 항목에는 다음 항목인 'Sentence'에 기재된 문장을 포함하고 있는 문서의 PubMed ID를 기재한다. 'Sentence' 항목은 개체가 추출된 문장을 의미한다. 'ID_ICL' 항목은 개체명 인식용 학습데이터를 가공하는 과정에서 부여한 자체 식별자이다. 다양한 주제 분야 코퍼스의 통합 및 일관성 확보를 위해 데이터 가공과정에서 자체적으로 부여하였으며, 식별자 부여 양식은 〈표 7〉과 같다. 식별자의 첫 글자인 C는 코퍼스(Corpus)를 의미하며, 두 번째로 나타나는 GE, DI는 각 개체명이 포함되는 주제의 영문명 2글자이다. 다음으로 E는 개체(Entity)를 의미하며, 마지막으로 숫자는 코퍼스 데이터 내에 개체가 등장하는 순서에 따라 부여하였다.

'ID_ICL' 뒤에 나타나는 'ID_기관명' 항목은 여러 기관들에서 부여한 개체의 ID를 의미한다. 만약 ID를 부여한 기관이 1개 이상 존재할 경우 'ID_기관명' 항목이 여러 개 부여될 수 있다. 부여된 타기관 ID가 없는 경우에는 'ID_기관명' 항목을 부여하지 않아도 되지만 이외의 모든 항목들은 반드시 부여되어야 한다.

(4) 개체명 인식 코퍼스의 체계화를 통한 학습 데이터 구축

선정된 체계화 양식에 따라 수집된 개체명 인식 코퍼스를 체계화 한 결과로 구축된 개체명 인식용 학습 데이터의 구성 양식 및 포함하고 있는 개체명 수는 〈표 8〉과 같다.

EntityName \t Synonyms \t Entity_Type \t Offset \t PMID \t Sentence \t ID_ICL \t ID_기관명

〈그림 4〉 개체명 인식용 학습데이터의 체계화 양식

〈표 7〉 개체명 인식 코퍼스의 개체 식별자 부여 양식

분야	개체 식별자 부여 양식
유전자	C_GE_E1 - C_GE_Exxxx (개체 끝 번호)
질병	C_DI_E1 - C_DI_Exxxx (개체 끝 번호)

〈표 8〉 개체명 인식용 학습 데이터의 구성 양식 및 개체명 개수

주제	코퍼스명	구성 양식	개체명 수
유전자	Genia Term Annotation collection	EntityName \t Synonyms \t EntityType \t Offset \t PMID \t Sentence	24,010
질병	NCBI_Corpus	EntityName \t Synonyms \t EntityType \t Offset \t PMID \t Sentence \t ID_ICL \t ID_UMLS	6,337
	Arizona Disease Corpus(AZDC)	EntityName \t Synonyms \t EntityType \t Offset \t PMID \t Sentence \t ID_ICL \t ID_MESH \t ID_OMIM	2,680

5.2.2 개체명 간 관계추출을 위한 학습 데이터 구축

개체명 간 관계 추출 시스템을 위한 학습 데이터는 수집된 관계 정의 코퍼스를 기반으로 구축되었으며, 수집된 관계 정의 코퍼스 자원의 수는 총 4개이다. 수집된 자원들을 대상은 모두 자원 분석, 체계화 양식 선정, 체계화 단계를 통해 학습 데이터로 구축되었다.

(1) 관계 정의 코퍼스 자원 수집

개체명 간 관계 추출용 학습 데이터 구축을 위해 수집된 관계 정의 코퍼스는 유전자 분야에서 1개, 단백질 분야에서 2개, 약물 분야에서 1개이다. 각 코퍼스에 대한 정보 및 포함하고 있는 개체명 수는 〈표 9〉와 같다.

(2) 관계 정의 코퍼스 자원의 분석

수집된 4개의 관계 정의 코퍼스를 대상으로 데이터 항목 및 내용에 대한 세부 분석을 진행하였다. 데이터 항목 분석 결과 관계 정의 코퍼스는 일반적으로 관계를 구성하는 2개의 개체명과 개체명의 동의어, 각 개체에 대한 ID, 관

〈표 9〉 수집된 생의학 분야 개체 관계 정의 코퍼스 자원

분야	코퍼스명	코퍼스 정보	개체명 수
유전자	Genetic Association Database(GAD)	약사 및 약물 정보 전문가에 의해 수동으로 구축된 FDA-approved drug package inserts(PIs) 주석 데이터(National Institutes of Health 2017).	5,330
단백질	LocText Corpus	뮌헨 기술 대학교 정보학 박사 과정 연구 센터(CeDoSIA)가 구축한 단백질, 세포 내 지방화, 유기체, 유기체간의 관계에 대해 수동으로 주석을 달아둔 데이터(tagtog 2017)	550
	Protein-Protein and Drug-Drug interaction Silver Standard Corpora	Philippe Thomas and Tamara Bobić and Ulf Leser and Martin Hofmann-Apitius and Roman Klinger의 논문(Thomas et al. 2012)에 의해 생성된 데이터(Fraunhofer Institute for Algorithms and Scientific Computing SCAI 2017).	200,000
약물	PK DDI	캐나다 정부의 지원을 받는 최첨단 대사 연구 시설인 'TMID (Metabolomics Innovation Center)'에서 구축한 오픈 소스 데이터(The University of Pittsburgh Pharmacokinetic Drug-drug Interation(PK DDI) Package Insert Corpus 2017).	1,893

계 식별자(ID), 개체명과 그 관계를 추출한 문장 및 문서, 개체명 및 관계를 추출한 문장 또는 문서의 PubMed ID와 개체의 위치정보 등을 포함하는 것을 확인하였다.

또한 데이터 내용 분석을 통해 수집된 코퍼스들이 각각 문장, 문서, 문단 등 서로 다른 기준에 따른 위치(Offset)정보를 제공하는 것을 〈표 10〉과 같이 확인하였다. 따라서 위치 정보에 대한 표현 양식을 통일하기 위하여 위치정보를 추출하는 기준을 문장단위로 선정하였으며, 이에 따라 분석 과정에서 위치정보를 재추출하는 과정이 추가로 수행되었다.

(3) 관계 정의 코퍼스 자원의 체계화 양식 선정

관계 정의 코퍼스 자원을 대상으로 항목 및 내용을 분석하여 도출한 공통 항목들은 Relation ID, Relation 유형, 개체 ID, 개체명, 개체명의 동의어, 개체를 추출한 문장, 문장 내 개체 위치 정보이다. 따라서 이를 기반으로 〈그림 5〉와 같이 관계 정의 코퍼스 자원의 체계화 양식을 선정하였다.

관계 추출용 학습 데이터를 구성하는 항목들은 크게 관계정보와 관련된 항목, 개체명 1과 관련된 항목, 개체명 2와 관련된 항목, 부가정보에 대한 항목으로 구분할 수 있으며 각각에 해당하는 항목명은 〈표 11〉과 같다.

〈표 10〉 개체 관계 정의 코퍼스들의 Offset 제공 기준

분야	정보원명	제공되는 offset 기준
유전자	GAD	문장
단백질	LocText	문서
	SilverPPICorpus	문장
약물	PK DDI	문서

ID_ICL_Relation \t Relation_Type \t ID_ICL_Entity1 \t Entity1_Name \t Entity1_Synonyms \t Entity1_Offset \t Entity1_Type \t ID_ICL_Entity2 \t Entity2_Name \t Entity2_Synonyms \t Entity2_Offset \t Entity2_Type \t Sentence \t ID_기관명_Relation \t ID_기관명_Entity1 \t ID_기관명_Entity2

〈그림 5〉 관계 추출용 학습 데이터의 체계화 양식

〈표 11〉 관계 추출용 학습 데이터의 구성 항목 구분

구분	항목명
관계정보와 관련된 항목	ID_ICL_Relation, Relation_Type, ID_기관명_Relation
개체명1과 관련된 항목	ID_ICL_Entity1, Entity1_Name, Entity1_Synonyms, Entity1_Offset, Entity1_Type, ID_기관명_Entity1
개체명 2와 관련된 항목	ID_ICL_Entity2, Entity2_Name, Entity2_Synonyms, Entity2_Offset, Entiy2_Type, ID_기관명_Entity2
부가 정보에 대한 항목	Sentence

관계 정보와 관련된 항목은 'ID_ICL_Relation'과 'Relation_Type'이다. 'ID_ICL_Relation' 항목은 관계 추출용 학습 데이터 구축 과정에서 개체명 간 관계에 부여한 관계 식별자이다. 관계 식별자는 다양한 주제 분야 코퍼스의 통합 및 일관성 확보를 위해 데이터 가공과정에서 자체적으로 부여하였으며, 식별자 부여 양식은 〈표 12〉와 같다. 식별자의 첫 글자인 C는 코퍼스(Corpus)를 의미하며, 두 번째로 나타나는 GE, PR, DR은 각각 개체명이 포함되는 주제의 영문명 2글자이다. 다음으로 R은 관계(Relation)를 의미하며, 마지막으로 숫자는 코퍼스 데이터 내에 개체가 등장하는 순서에 따라 부여하였다.

'Relation_Type' 항목은 수집된 관계 정의 코퍼스에 기재된 관계 유형을 기재하는 항목이고, 'ID_기관명_Relation' 항목은 코퍼스를 구축한 기관에서 부여한 관계 식별자를 기재하는 항목이다.

다음으로 개체명 1과 관련된 항목은 'ID_ICL_Entity1', 'Entity1_Name', 'Entity1_Synonyms', 'Entity1_Offset', 'Entity1_Type'이다. 'ID_ICL_Entity1'은 관계 추출용 학습 데이터 구축 과정에서 개체명 간 관계에 부여한 관계 식별자이며, 식별자 부여 양식은 '2.1의 (3)개체명 인식 코퍼스 자원의 체계화 양식 선정'에서 부여한 관계 식별자 양식과 같다. 'Entity1_Name' 항목에는 개체명을 기재하고, 'Entity1_Synonyms'에는 개체명의 동의어 및 약어 등을 기재한다. 'Entity1_Offset'에는 개체명을 추출한 문장 내에서의 개체명1의 위치정보를 기재하고 'Entity1_Type'에는 개체명의 유형을 기재한다. 'ID_기관명_Entity1'은 코퍼스를 구축한 기관에서 부여한 개체 식별자를 기재하는 항목이다.

개체명 2와 관련된 항목들은 'ID_ICL_Entity2', 'Entity2_Name', 'Entity2_Synonyms', 'Entity2_Offset', 'Entity2_Type'이며, 개체명 1과 관련된 항목들을 기재하는 방법과 동일한 방법으로 기재한다.

마지막으로 부가정보에 대한 항목인 'Sentence'에는 관계를 구성하는 두 개의 개체명을 추출한 문장을 기재한다. 관계 추출용 학습 데이터를 구성하는 모든 항목들은 필수적으로 부여해야 한다.

(4) 가공 결과

선정된 체계화 양식에 따라 수집된 개체명 관계 정의 코퍼스를 체계화 한 결과로 구축된 개체명 간 관계 추출용 학습 데이터의 구성 양식 및 포함하고 있는 개체명 수는 〈표 13〉과 같다.

〈표 12〉 관계 정의 코퍼스의 관계 식별자 부여 양식

분야	관계 식별자 부여 양식
유전자	C_GE_R1 ~ C_GE_Rxxxx (개체 끝 번호)
단백질	C_PR_R1 ~ C_PR_Rxxxx (개체 끝 번호)
약물	C_DR_R1 ~ C_DR_Rxxxx (개체 끝 번호)

〈표 13〉 관계 추출용 코퍼스의 체계화 양식 및 개체명 개수

주제	코퍼스명	코퍼스의 필드명	개체명 수
유전자	Genetic Association Database(GAD)	ID_ICL_Relation \t Relation_Type \t ID_ICL_Entity1 \t Entity1_Name \t Entity1_Synonyms \t Entity1_Offset \t Entity1_Type \t ID_ICL_Entity2 \t Entity2_Name \t Entity2_Synonyms \t Entity2_Offset \t Entity2_Type \t Sentence \t ID_GAD_Relation \t ID_GAD_Entity1 \t ID_GAD_Entity2	5,330
단백질	Protein-Protein Silver Standard Corpora	ID_ICL_Relation \t Relation_Type \t D_ICL_Entity1 \t Entity1_Name \t Entity1_Synonyms \t Entity1_Offset \t Entity1_Type \t D_ICL_Entity2 \t Entity2_Name \t Entity2_Synonyms \t Entity2_Offset \t Entity2_Type \t Sentence \t ID_SilverPPI_Relation \t ID_SilverPPI_Entity1 \t ID_SilverPPI_Entity2	200,000
	LocText_Corpus	ID_ICL_Relation \t Relation_Type \t ID_ICL_Entity1 \t Entity1_Name \t Entity1_Synonyms \t Entity1_Offset \t Entity1_Type \t ID_ICL_Entity2 \t Entity2_Name \t Entity2_Synonyms \t Entity2_Offset \t Entity2_Type \t Sentence \t ID_LocText_Relation \t ID_UniProt_Entity1 \t ID_GO_Entity2	320
약물	PK DDI	ID_ICL_Relation \t Relation_Type \t ID_ICL_Entity1 \t Entity1_Name \t Entity1_Synonyms \t Entity1_Offset \t Entity1_Type \t ID_ICL_Entity2 \t Entity2_Name \t Entity2_Synonyms \t Entity2_Offset \t Entity2_Type \t Sentence\t ID_PKDDI_Relation \t ID_PKDDI_Entity1 \t ID_PKDDI_Entity2	1,893

6. 구축 결과 분석

본 논문에서는 생의학 분야 학술 논문에서의 개체명 인식 및 개체 간 관계 추출 시스템을 위한 학습 데이터를 구축하기 위하여 유관분야 개체명 사전과 학습 집합을 수집하였다. 또한 수집된 자원들을 대상으로 특성을 분석하고, 분석을 통해 체계화 양식을 선정하였으며, 결과적으로 수집 자원들을 학습 데이터로 체계화하였다. 그 중에서 개체명 사전을 통해 구축된 학습 데이터를 대상으로 학습데이터의 생의학 분야 개체명 수용 범위를 파악하기 위한 검증을 실시하였다. 검증 방법은 유관 분야의 텍스트 문서에 학습 데이터에 존재하는 개체명 즉, 유전자명, 단백질명, 질병명, 약물명을 검색하는 일치도 평가이다. 본 연구에서 일치도 평가를 위해 선정한 유관분야 텍스트 문서는 미국 국립 의학 도서관과 미국 국립 보건원에서 운영하는 대표적인 생의학 분야 데이터베이스인 PubMed에서 수록하고 있는 모든 초록을 다운로드한 데이터이다. 그러나 해당 데이터는 양이 매우 방대하여 일치도 검사 과정에서 데이터를 처리하는 데 있어 많은 문제들이 발생하였다. 따라서 현재 생의학 분야에서 가장 활발

한 연구가 진행되어 빠른 속도로 비정형 텍스트의 양이 증가하고 있는 알츠하이머 분야로 그 범위를 축소하여 PubMed 알츠하이머 초록을 대상으로 개체명 사전의 일치도를 평가하였다. PubMed 알츠하이머 초록은 PubMed에 'Alzheimer'를 검색(2017.05.26.)하고, 조건을 'Abstract'로 부여한 결과를 대상으로 수집하였다.

〈표 14〉 PubMed 알츠하이머 초록 데이터에 대한 통계정보

	건수	문장 수	어절 수
PubMed 알츠하이머 초록	91,730	132,440	102,491,804

일치도 평가를 위한 검색 방법은 크게 개체명 단순 비교를 통한 일치도 검사와 NLTK (Natural Language Toolkit) 파서 적용을 통한 일치도 검사로 구분된다. NLTK 파서는 자연어 처리 및 문서 분석을 위한 도구로 품사 태깅, 구문분석, 형태소 분석 등의 기능을 제공한다(Natural Language Toolkit 2017).

일치도 평가를 진행하기에 앞서 보다 정확한 검색을 위하여 사전의 개체명과 PubMed 알츠하이머 초록 데이터를 모두 소문자화 하는 전처리 과정이 선행되었다.

6.1 개체명 단순 비교를 통한 일치도 검사

개체명 단순 비교를 통한 일치도 검사는 가공 사전의 'Name' 필드에 해당하는 개체명을 PubMed 알츠하이머 초록 데이터에 검색하는 방식이다. 개체명이 PubMed 알츠하이머 초록에 존재하는지 확인하기 위한 완전 일치 검사이다.

그러나 해당 검사 방법은 개체명이 속한 모든 단어에서 검색되는 문제점을 지닌다. 예를 들어 개체명이 질병 중 하나인 '틱증'을 나타내는 'Tics'일 경우 전처리 과정을 통해 소문자 되어 'tics'가 되고, PubMed 알츠하이머 초록에 존재하는 모든 'Tics'를 찾아낸다. 따라서 개체명으로 사용된 'Tics'뿐만 아니라 'biotics(생명역학)', 'genetics(유전학)'에서도 개체명이 검색되어 올바른 검색 수치를 파악할 수 없다.

또한 개체명의 약어가 소문자화 되어 영어에서 일반적으로 사용되는 전치사나 조사, 동사와 같은 단어가 되는 문제점이 발생하였다. 예를 들어 'Ataxia Telangiectasia'는 '모세혈관 확장 운동실조'라는 질병명이다. 해당 질병명의 약어는 'AT'이며, 이를 소문자화 할 경우 'at'가 되어 영어의 전치사 'at'과 같은 단어가 된다. 따라서 PubMed 알츠하이머 초록의 모든 전치사와 일치하여 정확한 검색 수치를 파악하는 것이 불가능하였다.

6.2 NLTK 파서 적용을 통한 일치도 검사

NLTK 파서 적용을 통한 일치도 검사는 개체명 단순 비교를 통한 일치도 검사에서 발생하는 문제점들을 확인하고, 이를 개선하기 위해 진행된 검사이다.

해당 검사를 위해 먼저 PubMed 알츠하이머 초록 데이터에 NLTK 파서를 적용하여 단어 단위의 POS(part of speech, 품사) 정보를 태깅하였다. 그리고 개체명으로 활용된 단어만을 확인하기 위하여, 해당되는 품사정보인 일반명사(NN), 일반명사의 복수형태(NNS), 고유명

사(NNP), 고유명사의 복수형태(NNPS) 태그가 달린 단어들만을 검색 횟수와 함께 따로 추출하였다. 추출된 단어와 검색 횟수들을 통해 1차 조사 결과를 수정하였으며 이에 대한 결과는 〈표 15〉와 같다.

유전자 통합 사전에 존재하는 개체명의 개수는 183,603개이며 전체 개체명의 약 2%에 해당하는 3,128개의 개체가 PubMed 알츠하이머 초록에서 검색되었다. 단백질 사전에서 검색된 개체는 전체 개체명의 약 2%에 해당하는 1,883개이며, 질병사전에서는 전체 개체명의 10%에 해당하는 1,173개이다. 약물 사전에서 검색된 개체명은 전체 개체명의 약 13%에 해당하는 1,082개로 가장 높은 일치도 검사결과를 보였다. 이는 PubMed 알츠하이머 초록에 알츠하이머를 치료하는 특정 약물 개체명이 자주 등장하기 때문으로 파악된다.

일치도 검사를 위한 검색을 통해 확인한 PubMed 알츠하이머 초록에 존재하는 각 분야별 개체명 등장 순위는 〈표 16〉과 같다.

유전자 분야에서 가장 많이 검색된 개체명은 알츠하이머 병(alheimer disease)으로 15,500회 검색되었다. 두 번째와 세 번째로 많이 검색된 개체명은 아밀로이드전구체 단백질(amyloid precursor protein)과 인슐린(insulin)으로 각각 5,993회, 2,720회 검색되었다.

단백질 분야에서 가장 많이 검색된 개체명은 아밀로이드전구체 단백질(Amyloid precursor protein)로 5,993회 검색되었다. 두 번째와 세 번째로 많이 검색된 개체명은 인슐린(insulin)과 아세틸콜린에스테라제(acetylcholinesterase)로 각각 2,720회, 2,333회 검색되었다. 이외에도 콜린에스테라아제(cholinesterase), 프로테아제(protease) 등 알츠하이머와 관련된 개체명이 검색되었다.

질병 분야에서 가장 많이 검색된 개체명은 알츠하이머 병(alzheimer's disease)이며 72,666회 검색되었다. 두 번째 및 세 번째로 많이 검색된 개체명은 치매(dementia)와 알츠하이머 병(alzheimer disease)으로 각각 57,571회, 15,500회 검색되었으며 모두 알츠하이머와 관련 있는 개체명이라 할 수 있다. 이외에도 인지장애(cognitive impairment), 경도인지장애(mild cognitive impairment), 신경퇴행성 질병(neurodegenerative disease) 등 알츠하이머와 관련된 개체명이 검색된 것을 확인할 수 있다.

약물 분야에서는 알츠하이머를 치료하는 약물들이 주로 검색되었다. 약물 분야에서 가장 많이 검색된 개체명은 도네페질(donepezil)로 2,977회 검색되었다. 두 번째와 세 번째로 많이

〈표 15〉 개체명 일치도 평가 결과

	유전자	단백질	질병	약물
기존 사전에 존재하는 개체명(Name) 수	183,603	124,204	11,865	8,283
PubMed 알츠하이머 초록에 존재하지 않는 개체명의 수	180,475	122,321	10,692	7,201
PubMed 알츠하이머 초록에 존재하는 개체명의 수	3,128 (2%)	1,883 (2%)	1,173 (10%)	1,082 (13%)

〈표 16〉 알츠하이머 초록에 존재하는 각 분야별 개체명의 순위

사전 / 순위	유전자	단백질	질병	약물
	CTD & HGNC	SwissProt	MEDIC	DrugBank
1	alzheimer disease (15,500)	amyloid precursor protein (5,993)	alzheimer's disease (72,666)	donepezil (2,977)
2	amyloid precursor protein (5,993)	insulin (2,720)	dementia (57,571)	memantine (2,253)
3	insulin (2,720)	acetylcholinesterase (2,333)	alzheimer disease (15,500)	rivastigmine (1,771)
4	acetylcholinesterase (2,333)	cholinesterase (2,177)	cognitive impairment (13,546)	galantamine (1,553)
5	diabetes (2,251)	protease (1,309)	mild cognitive impairment (6,314)	dopamin (1,550)
6	cholinesterase (2,177)	neurotrophic factor (810)	neurodegenerative disease (5,778)	acetylcholine (1,516)
7	presenilin (1,769)	microtubule-associated protein (801)	neurodegenerative diseases (4,229)	tacrine (1,044)
8	growth factor (1,666)	beta-amyloid precursor protein (784)	senile plaque (4,053)	nitric oxide (888)
9	protease (1,309)	acetyltransferase (750)	neurodegenerative disorder (4,024)	choline (877)
10	amyloid protein (1,091)	phosphatase (669)	amyloid plaque (3,897)	melatonin (799)

검색된 개체명은 메만틴(memantine)과 리바스티그민(rivastigmine)으로 각각 2,253회, 1,771회 검색되었다. 이외에도 갈란타민(galantamine), 도파민(dopamin), 아세틸콜린(acetylcholine) 등이 검색되었다.

개체명 일치도 평가 결과가 비교적 낮게 나타나는 것은 검증 대상으로 선정한 PubMed 알츠하이머 초록 데이터의 규모에 비해 구축된 학습 데이터의 규모가 작아서 나타나는 결과이다. 본 연구는 기계학습을 위한 학습 데이터 구축을 체계화 하는 방안을 마련하기 위한 초기 연구로써 제안된 방법을 통해 구축된 학습데이터가 실제 기계 학습에 사용될 수 있는지에 대한 가능성을 확인하기 위한 절차로 데이터 검증을 실시하였다. 검증 결과 나타난 일치도를 통해 그 가능성을 확인할 수 있으며, 개체명이 검색된 횟수를 통해 실제로 유관분야 비정형 텍스트에 자주 등장하는 개체명들이 학습 데이터에

다수 포함되어 있는 것을 확인할 수 있다. 따라서 후속 연구를 통해 자원의 수집 대상 및 텍스트 범위를 확대하여 보다 광범위한 학습 데이터를 구축하고, 이에 대한 일치도를 지속적으로 확인할 예정이다.

7. 결론 및 제언

본 연구에서는 급격히 증가하는 비정형 텍스트를 효율적으로 파악하고 이를 통해 핵심 개체명 및 개체명 간 관계를 추출하기 위하여 기존재하는 유관분야의 언어자원을 수집하고 이를 통해 학습 데이터를 구축하는 과정을 체계화하기 위한 방안을 제안하였다.

제안된 방법은 먼저 학습 데이터 구성에 기반이 될 유관분야 언어 자원을 수집한다. 수집된 자원은 유관분야 개체명 사전과 개체명 인식용 학습 집합, 관계 추출용 학습 집합이다. 그리고 수집된 다양한 형태의 자원들을 대상으로 항목 및 내용에 대한 특성 분석을 수행하여, 개체 식별에 반드시 필요한 항목들을 도출한다. 도출된 항목들을 학습 데이터의 구성 양식을 선정하며, 선정된 양식에 따라 수집된 언어자원들을 체계화된 학습 데이터로 변환하는 가공 과정을 수행한다.

실제로 해당 과정을 통해 5개의 개체명 사전과 3개의 개체명 인식용 학습 집합이 개체명 인식 시스템을 위한 학습 데이터로 구축되었으며, 4개의 관계 추출용 학습 집합이 개체명 간 관계 추출 시스템을 위한 학습 데이터로 구축되었다.

또한 본 연구에서는 제안된 방법을 통해 구축한 개체명 사전 기반 학습 데이터를 대상으로 학습 데이터의 개체명 수용범위를 검증하기 위한 데이터 검증을 실시하였다. 데이터 검증에는 최근 생의학 분야에서 가장 활발히 연구되어 비정형 텍스트의 양이 급격히 증가하고 있는 알츠하이머 분야 PubMed 초록 데이터를 사용하였으며, 개체명 사전의 데이터 검증 결과 유전자 분야와 단백질 분야에서 각각 2%, 질병 분야에서 10%, 약물 분야에서 13%의 데이터 일치도를 확인하였다.

본 연구는 기계 학습 기반의 생의학 분야 개체명 인식 및 개체명 간 관계 추출 시스템을 위한 학습 데이터 구축 과정을 체계화하기 위한 방안을 마련하기 위한 초기 연구로서, 자원의 수집 대상 및 텍스트 범위를 최소화 하여 학습 데이터를 구축하였다. 따라서 일치도 검사를 위한 검색 대상인 PubMed 초록 데이터의 규모에 비해 학습 데이터의 규모가 현저히 작아 일치도 검사 결과가 낮게 측정되는 결과를 보였다. 하지만 해당 수치는 제안된 방법의 가능성을 확인하기 위한 도구이며, 일치도와 함께 나타난 개체명 검색 결과를 통해 구축된 학습 데이터가 유관분야에서 자주 사용되는 개체명을 수록하고 있다는 것을 확인할 수 있었다. 따라서 후속 연구를 통해 언어 자원의 수집 분야 및 텍스트 범위를 확대하여 보다 방대한 광범위한 학습 데이터를 구축하고 이에 대한 검증을 실시하여 개선된 일치도를 확인할 예정이다.

참 고 문 헌

[1] 박성배. 2005. 기계학습/텍스트마이닝과 생명과학. 『정보과학회지』, 23(5): 32-40.

[2] 박경미, 황규백. 2011. 자연어처리 기반 바이오 텍스트 마이닝 시스템. 『정보과학회논문지: 컴퓨팅의 실제 및 레터』, 17(4): 205-213.

[3] 송영길, 정석원, 김학수. 2015. 위키피디아 기반 개체명 사전 반자동 구축 방법. 『정보과학회논문지』, 42(11): 1397-1403.

[4] 신성호 외. 2014. 개체명 인식 향상을 위한 학습 집합 및 개체명 인식 모델 구축. 『정보과학회논문지: 컴퓨팅의 실제 및 레터』, 20(7): 425-429.

[5] 이혜진, 김재웅. 2017. 자연어 처리 기술 현황 및 표준화 동향에 관한 연구. 『한국통신학회 학술대회 논문집』, 2017년 6월 21일, 제주: 라마다 프라자 제주 호텔: 876-877.

[6] 허고은, 송민. 2014. 텍스트 마이닝 기반의 그래프 모델을 이용한 미발견 공공 지식 추론. 『정보관리학회지』, 31(1): 231-250.

[7] Ananiadou, S., Kell, D. B. and Tsujii, J. 2006. "Text Mining and Its Potential Applications in Systems Biology." *Trends in Biotechnology*, 24(12): 571-579.

[8] Beuning, P. and Musier-Forsyth, K. 1999. "Transfer RNA Recognition by Aminoacyl-tRNA Synthetases." *Biopolymers*, 52(1): 1-28.

[9] Biomedical Informatics Lab at ASU. 2017. *Arizona Disease Corpus*. [online] [cited 2017. 6. 1.] 〈http://diego.asu.edu〉

[10] Choi, S. 2016. "Extraction of Protein-Protein Interactions (PPIs) from the Literature by Deep Convolutional Neural Networks with Various Feature Embeddings." *Sage Journals*, 2016.

[11] Comparative Toxicogenomics Database. 2017. *Gene vocabulary*. [online] [cited 2017. 4. 27.] 〈http://ctdbase.org/;jsessionid=0868DE4D459374D22AB222F9CC3ECA43〉

[12] DrugBank. 2017. *COMPLETE DATABASE: All drugs*. [online] [cited 2017. 4. 27.] 〈https://www.drugbank.ca/〉

[13] Fraunhofer Institute for Algorithms and Scientific Computing SCAI. 2017. *Silver Standard Corpus for Protein Protein and Drug Drug Interaction*. [online] [cited 2017. 6. 2.] 〈https://www.scai.fraunhofer.de/en.html〉

[14] GENIA: The BioNLP Shared Task 2016. 2017. *The BioNLP Shared Task*. [online] [cited 2017. 10. 9.] 〈http://2016.bionlp-st.org/〉

[15] Huang, C. and Lu, Z. 2016. "Community Challenges in Biomedical Text Mining over 10 years: Success, Failure and the Future." *Briefings in Bioinformatics*, 17(1): 132-144.

[16] HUGO Gene Nomenclature Committee. 2017. *Complete HGNC Dataset*. [online] [cited 2017. 4. 27.] 〈https://www.genenames.org/〉

[17] Jensen, L. J., Saric, J. and Bork, P. 2006. "Literature Mining for the Biologist: from Information Retrieval to Biological Discovery." *Nature Reviews Genetics*, 7(2): 119-129.

[18] Kim, J., Wang, Y. and Yasunori, Y. 2013. "The Genia Event Extraction Shared Task, 2013 Edition-Overview." In *Proceedings of the BioNLP Shared Task 2013 Workshop*, August 9, 2013, Sofia: Association for Computational Linguistics.

[19] National Center for Biotechnology Information. 2017. *PubMed*. [online] [cited 2017. 6. 11.] 〈https://www.ncbi.nlm.nih.gov/〉

[20] National Institutes of Health. 2017. *Genetic Association Database*. [online] [cited 2017. 6. 1.] 〈https://www.nih.gov/〉

[21] Natural Language Toolkit. 2017. *Natural Language Processing with Python*. [online] [cited 2017. 7. 29.] 〈http://www.nltk.org/〉

[22] The National Centre for Text Mining. 2016. *Text Mining Resources*. [online] [cited 2017. 9. 17.] 〈http://www.nactem.ac.uk/resources.php〉

[23] The University of Pittsburgh Pharmacokinetic Drug-drug Interation (PK DDI) Package Insert Corpus. 2017. *Download the PK-DDI corpus with consensus annotations*. [online] [cited 2017. 6. 20.] 〈https://dbmi-icode-01.dbmi.pitt.edu/dikb-evidence/package-insert-DDI-NLP-corpus.html〉

[24] tagtog. 2017. *LocText*. [online] [cited 2017. 6. 2.] 〈https://www.tagtog.net/〉

[25] Thomas, P et al. 2012. "Weakly labeled corpora as silver standard for drug-drug and protein-protein interaction." In *Proceedings of the Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM) on Language Resources and Evaluation Conference (LREC)*, 2012. Istanbul, Turkey.

[26] Tripathi, V. et al. 2010. "The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation." *Molecular Cell*, 39(6): 925-938.

[27] UniProt. 2017. *Uniprot data*. [online] [cited 2017. 4. 27.] 〈http://www.uniprot.org/〉

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Park, S. 2005. "Machine Learning/Text Mining and Life Science." *Journal of KIISE*, 23(5):

32-40.

[2] Park, K. and Hwang, k. 2011. "A Bio-Text Mining System Based on Natural Language Processing." *KIISE Transactions on Computing Practices*, 17(4): 205-213.

[3] Song, Y., Jeong, S. and Kim, H. 2015. "(A)Semi-automatic Construction method of a Named Entity Dictionary Based on Wikipedia." *Journal of KIISE*, 42(11): 1397-1403.

[4] Shin, S. et al. 2014. "Construction of Tagged Corpus and a Statistical Model for Improvement of Named Entity Recognition." *KIISE Transactions on Computing Practices*, 20(7): 425-429.

[5] Lee, H. and Kim, J. 2017. "A Study on the Natural Language Processing(NLP) Technical and Standardization Trend." *Proceedings of Symposium of the Korean Institute of communications and Information Sciences*, June 21, Jeju: Ramada Plaza Jeju Hotel: 876-877.

[6] Heo, G. and Song, M. 2014. "Inferring Undiscovered Public Knowledge by Using Text Mining-driven Graph Model." *Journal of the Korean society for Information Management*, 31(1): 231-250.