토픽 모델링 기반 과학적 지식의 불확실성의 흐름에 관한 연구

The Stream of Uncertainty in Scientific Knowledge using Topic Modeling

허고은 (Go Eun Heo)*

초 록

과학적 지식을 얻는 과정은 연구자의 연구를 통해 이루어진다. 연구자들은 과학의 불확실성을 다루고 과학적 지식의 확실성을 구축해나간다. 즉, 과학적 지식을 얻기 위해서 불확실성은 반드시 거쳐가야 하는 필수적인 단계로 인식되고 있다. 현존하는 불확실성의 특성을 파악하는 연구는 언어학적 접근의 hedging 연구를 통해 소개되었으며 컴퓨터 언어학에서 수작업 기반으로 불확실성 단어 코퍼스를 구축해왔다. 기존의 연구들은 불확실성 단어의 단순 출현 빈도를 기반으로 특정 학문 영역의 불확실성의 특성을 파악해오는데 그쳤다. 따라서 본 연구에서는 문장 내 생의학적 주장이 중요한 역할을 하는 생의학 문헌을 대상으로 불확실성 단어 기반 과학적 지식의 패턴을 시간의 흐름에 따라 살펴보고자 한다. 이를 위해 생의학 온톨로지인 UMLS에서 제공하는 의미적 술어를 기반으로 생의학 명제를 분석하였으며, 학문 분야의 패턴을 파악하는데 용이한 DMR 토픽 모델링을 적용하여 생의학 개체의 불확실성 기반 토픽의 동향을 종합적으로 파악하였다. 시간이 흐름에 따라 과학적 지식의 표현은 불확실성이 감소하는 패턴으로 연구의 발전이 이루어지고 있음을 확인하였다.

ABSTRACT

The process of obtaining scientific knowledge is conducted through research. Researchers deal with the uncertainty of science and establish certainty of scientific knowledge. In other words, in order to obtain scientific knowledge, uncertainty is an essential step that must be performed. The existing studies were predominantly performed through a hedging study of linguistic approaches and constructed corpus with uncertainty word manually in computational linguistics. They have only been able to identify characteristics of uncertainty in a particular research field based on the simple frequency. Therefore, in this study, we examine pattern of scientific knowledge based on uncertainty word according to the passage of time in biomedical literature where biomedical claims in sentences play an important role. For this purpose, biomedical propositions are analyzed based on semantic predications provided by UMLS and DMR topic modeling which is useful method to identify patterns in disciplines is applied to understand the trend of entity based topic with uncertainty. As time goes by, the development of research has been confirmed that uncertainty in scientific knowledge is moving toward a decreasing pattern.

키워드: 텍스트 마이닝, 불확실성, DMR 토픽 모델링, 의미적 술어, 트렌드 분석
text mining, uncertainty, DMR topic modeling, semantic predication, trend analysis

* 연세대학교 문헌정보학과 연구교수(goeun.heo@yonsei.ac.kr)

■ 논문접수일자 : 2019년 2월 18일 ■ 최초심사일자 : 2019년 3월 13일 ■ 게재확정일자 : 2019년 3월 27일
■ 정보관리학회지, 36(1), 191-213, 2019. 〔http://dx.doi.org/10.3743/KOSIM.2019.36.1.191〕

1. 서 론

과학은 지식창조의 중요한 생산물이며 과학적 지식을 얻는 과정은 연구를 통해 이루어진다. 연구는 현존하는 한계점이나 과학적 불확실성(uncertainty)을 제거하기 위한 과정(Jensen, 2008)으로 연구자들은 불확실성을 다루고 불확실성을 해결하기 위해 노력한다. 연구자들은 기존에 알려져 있지 않은 과학적 진술에 대해 지속적으로 탐구하여 불확실성의 원천을 정교화하고 확실성을 구축해나간다(Ravetz, 1973; Zehr, 1999; Jensen, 2008). 즉, 과학적 지식과 학술적인 연구 결과(findings)를 얻기 위해 불확실성은 반드시 거쳐가야 하는 필수적인 단계로 인식되며 과학적 지식을 이해하기 위해 지식의 상태(epistemic status)의 흐름을 파악할 필요가 있다.

새로운 지식의 발견은 연구자의 연구 결과물인 학술 문헌을 통해 표현되어 왔으며 이러한 학술 문헌의 양은 시간이 흐름에 따라 기하급수적으로 증가하며 학문 영역이 발전하고 있다. 이처럼 시간의 흐름은 지식의 불확실성의 패턴을 발견하는데 중요한 요인이 될 수 있음에도 불구하고 기존의 연구들(Friedman, Alderson, Austin, Cimino, & Johnson, 1994; Hyland, 1998; Falahati, 2006; Rizomilioti, 2006; Vold, 2006)은 불확실성 단어(uncertainty word)의 단순 출현 빈도를 기반으로 특정 학문 영역의 불확실성의 특성을 파악해오는데 그쳤다.

특히, 생의학 학술 문헌에서 과학적 지식을 발견하는 과정은 불확실한 개체간의 관계성을 확실한 지식으로 연결해나가는 연구의 일환으로 학술 문헌에 표현된 연구자의 가설과 추측에 대한 주장이 중요한 역할을 하는 생의학 문헌의 특성에 따라 생의학 영역에서의 불확실성 연구의 필요성은 이미 많은 연구를 통해 인식되어 왔다(Friedman et al., 1994; Farkas, Vincze, Móra, Csirik, & Szarvas, 2010; Kilicoglu, Rosemblat, & Rindflesch, 2017; Malhotra, Younefsi, Gurulingappa, & Hofmann-Apitius, 2013; Szarvas, Vincze, Farkas, & Csirik, 2008; Vincze, Szarvas, Farkas, Móra, & Csirik, 2008; Wilbur, Rzhetsky, & Shatkay, 2006; Zerva, Batista-Navarro, Day, & Ananiadou, 2017).

따라서 본 연구에서는 문장 내 명제가 중요한 생의학적 의미를 지니는 생의학 문헌에서 과학적 지식의 불확실성의 특성을 시간의 흐름에 따라 파악하고자 한다. 이를 위해 Chen, Song, Heo(2018)의 연구에서 구축한 불확실성 단어를 이용하여 텍스트 마이닝 기반으로 생의학 문헌을 처리하고 생의학 명제를 표현하는데 적합한 SemMedDB(Semantic Medline Database)(Kilicoglu, Shin, Fiszman, Rosemblat, & Rindflesch, 2012)의 의미적 술어(semantic predications)를 추출하여 시간의 흐름에 따른 불확실성 단어의 패턴을 살펴본다. 더불어 시계열적 주제별 동향을 파악하는데 용이한 텍스트 마이닝 기법인 DMR 토픽 모델링(Mimno & McCallum, 2012)을 생의학 학술 문헌에서 중요한 역할을 하는 생의학 개체들을 기반으로 수행하여 불확실성 단어를 포함한 생의학 문장 내 개체들의 토픽과 각 토픽의 연도별 토픽 분포를 확인한다.

2. 이론적 배경

2.1 Hedging과 불확실성 연구

현존하는 불확실성을 해결하기 위한 학술적 연구로서 불확실성의 특성을 이해하는 연구들이 시도되었다. 불확실성에 대한 언어학적 접근은 hedging 연구를 통해 소개되었다. Lakoff (1972)는 hedging 단어를 처음으로 주장한 연구자로, hedging을 "words whose job is to make things fuzzier or less fuzzy"로 표현했다. 즉, 단어의 정확한 의미를 모호하게 만드는 단어들로 정의된다. Hyland(1998)는 학술 문헌의 hedging의 의미에 초점을 두고 과학적 담화에서 불확실성의 역할을 이해하기 위해 불확실성의 유형을 확인했다. Vincze et al.(2008)은 BioScope 코퍼스의 초록 내 문장의 17.70%와 전문(full-papers) 내 문장의 19.44%가 hedge 단어에 포함된다고 밝혔으며, 생의학 학술 문헌을 포함한 데이터베이스인 MEDLINE에서는 약 11%가 추측성 문장으로 구성된다고 보고했다(Light, Qiu, & Srinivasan, 2004).

또한 컴퓨터 언어학(computational linguistics)의 연구에서 불확실성을 자동적으로 추출하고 불확실성 텍스트의 범위를 인식하는 CONLL (Computational Natural Language Learning) 학회(Farkas et al., 2010)를 통해 활발한 연구들이 수행되어 왔다. 생의학 문헌의 데이터를 이용하여 불확실성을 다루기 위한 주석 지침(annotation guideline)을 개발하거나(Wilbur, Rzhetsky, & Shatkay, 2006), BioScope 코퍼스를 구축하고(Szarvas, Vincze, Farkas, & Csirik, 2008; Vincze et al., 2008) Hedging을 기반으로 추측성 문장과 비추측성 문장의 텍스트 분류 문제(Light, Qiu, & Srinivasan, 2004; Medlock & Briscoe, 2007; Szarvas, 2008; Malhotra, Younesi, Gurulingappa, & Hofmann-Apitius, 2013; Zerva, Batista-Navarro, Day, & Ananiadou, 2017; Kilicoglu, Rosemblat, & Rindflesch, 2017)를 규칙 기반 또는 기계 학습 기법으로 제안했다.

이처럼 기존의 연구들은 학술 문헌에서 불확실성을 파악하고, 특정 주장이나 연구 결과의 추측성과 불확실성을 효율적으로 인식할 수 있도록 지침을 개발하고 코퍼스를 구축해왔다. 최근 연구로 Chen, Song, Heo(2018)는 불확실성 단어 리스트를 정의하고 초기 단어 리스트를 기반으로 인공 신경망 기법으로 널리 알려진 Word2Vec(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)을 적용하여 의미적 연관 단어를 자동적으로 추출한 연구를 수행했다. 이는 기존 연구들의 한계점인 수작업을 보완하였을 뿐만 아니라 기존 연구에서 시도하지 못했던 학술 영역간의 일반성과 확장성을 고려하여 불확실성 단어를 구축한 연구로 의의를 지닌다.

또한 특정 학문 영역에서의 불확실성에 대한 연구나 단어의 출현 빈도를 기반으로 학문 영역간 특성을 비교한 연구들이 수행되었다. 의학 영역에서 Friedman et al.(1994)은 방사선학(radiology) 보고서에 포함된 hedging과 불확실성의 의미를 밝혔으며 연구 결과를 추출하기 위해 확실성을 다음과 같이 5가지 단계로 구분했다: 확실성이 없음(no), 낮은 확실성(low-certainty), 중간 확실성(moderate certainty), 높은 확실성(high certainty), 평가 불가(can not evaluate),

Rizomilioti(2006)는 세 가지 다른 학문 영역인 고고학, 문학비평, 생물학에서 지식의 양상(epistemic modality)을 표현하는 언어학적 도구를 분석하였다. 고고학 학술 논문이 더 많은 불확실성 단어를 포함하고 있으며 문학비평 논문에서는 가장 적은 불확실성 단어가 출현했다. Hyland(1998)는 인문학 학술 문헌이 과학 학술 문헌에 비해 더 많은 hedging 장치가 표현된다고 밝혔다. Falahati(2006)는 의학, 화학, 심리학 연구에서의 hedging 분포를 살펴보았고, 심리학 학술 문헌이 가장 많은 hedge를 포함했다. Vold(2006)는 세 가지 언어인 영어, 프랑스어, 노르웨이어와 두 다른 학문 영역인 언어학과 의학에서 hedge의 사용에 대한 비교를 수행했다. 영어와 노르웨이어가 프랑스어에 비해 더 많은 hedge를 사용하였고, 두 학문 영역 간의 차이는 특별히 나타나지 않았다.

새로운 지식이 발견되고 연구가 발전해오는 과정에서 시간의 흐름은 지식의 불확실성의 패턴을 발견하는데 중요한 요인이 될 수 있다. 그럼에도 불구하고 기존의 연구들은 불확실성 단어의 단순 출현 빈도를 기반으로 특정 학문 영역내에서 또는 학문 영역 간의 불확실성의 특성을 파악해왔다. 즉, 불확실성 단어를 기반으로 시간적 흐름에 따른 학문 분야의 특성과 패턴을 발견하는 연구가 시도되지 않았다.

2.2 텍스트 마이닝 기반 토픽 패턴 분석 연구

계량정보학은 학문분야의 연구 동향과 연구 생산성 및 학술적 영향력을 파악하기 위한 양적 접근법으로 잘 알려져 있는 학문이다. 이는 문헌정보학(White & McCain, 1998; Uzun, 2002; Zhao & Strotmann, 2008; Liu, Hu, & Wang, 2011; Åström, 2007; Zhao & Zhang, 2011), 생물정보학(An & Wu, 2011; Cambrosio, Limoges, Courtial, & Laville, 1993; Rip & Courtial, 1984; Song, Kim, Zhang, Ding, & Chambers, 2014) 등과 같은 특정 학문 분야의 토픽의 변화를 파악하거나 연구 영역을 분석하는 기본적인 접근법으로 동시 출현 단어 분석(Callon, Law, & Rip, 1986; Ding, Chowdhury, & Foo, 2001; Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011; Milojević, Sugimoto, Yan, & Ding, 2011), 동시 인용 분석(Culnan, 1986, 1987; Chen, 2006; Pilkington & Meredith, 2009), 저자 동시 인용 분석(White & McCain, 1998; Acedo & Casillas, 2005; Nerur, Rasheed, & Natarajan, 2008; Zhao & Strotmann, 2008; Chen & Guan, 2011), 공저자 분석(Peters & Van Raan, 1991; Åström, 2007), 또는 이러한 기법들을 동시에 적용하여 특정 학문 영역의 동향을 파악한 연구들이 있었다(Malin & Carley, 2007; Chen & Guan, 2011; Chang & Huang, 2012; Heo & Song, 2013).

이처럼 학문 분야의 연구 동향과 생산성을 파악한 연구들은 시대를 망라하여 중요한 연구 분야로 인식되어 왔으며 최근에는 전통적인 계량정보학 기법에 텍스트 마이닝 기법을 결합하여 토픽 모델링 기반으로 학문의 패턴을 분석하는 연구들이 다수 시도되어 왔다(Griffiths & Steyvers, 2004; Newman & Block, 2006; Jin, Heo, Jeong, & Song, 2013; Song, Heo, & Lee, 2015; Liu, Omar, Liou, Chi, & Hsu, 2015; Jeong, Heo, Kang, Yoon, & Song, 2016; Heo,

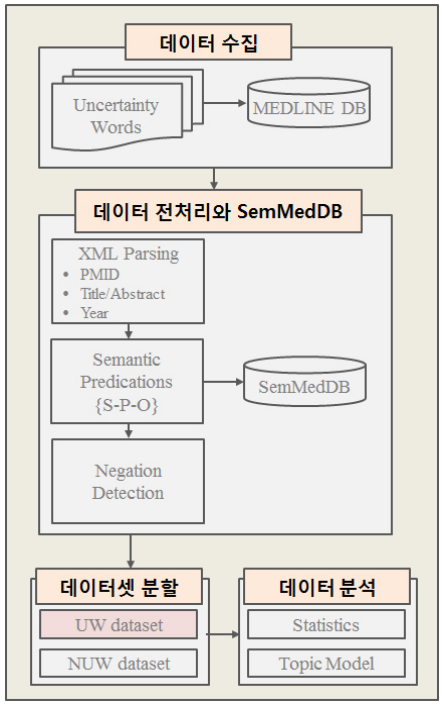
Kang, & Song, 2017).

토픽 모델링은 레이블되지 않은 대량의 텍스트를 간단하게 분석하는 방법으로 각 토픽은 함께 출현한 단어들의 클러스터로 구성된다(Steyvers & Griffiths, 2007). 초기의 대표적인 토픽 모델링 기법은 Blei, Ng, Jordan(2003)의 LDA(Latent Dirichlet Allocation) 토픽 모델링이다. 이 후 LDA에 결과 모델을 생성하는 방법을 좌우하는 하이퍼 파라미터(hyper parameter) α가 문헌의 메타데이터(연도, 저자, 국가 등)에 따라 달라질 수 있다고 가정하여 메타데이터를 기반으로 변화되는 주제 분포(dirichlet distribution)를 파악하기 위한 DMR(Dirichlet-Multinomial Regression) 토픽 모델링이 Mimno와 McCallum 연구자에 의해 제안되었다(Mimno & McCallum, 2012). 최근 연구들에서는 학문 분야의 동향을 파악하기 위해 연도 메타데이터를 포함한 연구들이 수행되었다(Song et al., 2014; Song, Heo, & Lee, 2015; Jeong et al., 2016).

따라서, 본 연구에서는 기존의 불확실성 관련 연구에서 시도되지 않았던 시간의 흐름에 따른 불확실성 연구의 패턴을 파악하고자 한다. 이를 위해 학술 문헌에 표현된 연구자의 가설과 추측 관련 주장이 중요한 역할을 하는 생의학 학술 문헌을 대상으로 선정하였고 생의학 개체들 간의 관계성을 효과적으로 해석(Hristovski, Friedman, Rindflesch, & Peterlin, 2006; Sebastian, Siew, & Orimaye, 2017)하는 UMLS에서 제공하는 의미적 술어를 적용하였다. 이는 기존의 전통적인 동시 출현 단어 기법에서 요구되는 수작업 절차를 생략할 수 있을 뿐만 아니라 문헌 내의 복잡하고 숨겨진 정보를 정규화, 일반화하여 다양한 컨셉 간의 관계성을 명확하게 활용할 수 있다는 장점을 지닌다. 특히, 불확실성 단어가 포함된 문장의 생의학 개체들이 어떠한 주제들을 포함하고 있는지 특성을 파악하고 시간의 흐름에 따른 토픽의 변화 추이를 살펴보기 위해 기존의 시계열적 연구 동향과 패턴 분석에 널리 활용되어 오고 있는 기법인 DMR 토픽 모델링(Mimno & McCallum, 2012)을 적용하였다.

3. 연구 설계

본 연구의 개요는 〈그림 1〉과 같다. 각 단계에 대한 설명은 해당 절에서 상세히 기술한다.



〈그림 1〉 연구 개요

3.1 데이터 수집

본 연구에서는 생명과학과 생의학 주제의 학술 문헌을 포함하고 있는 검색엔진인 PubMed에서 196개 불확실성 단어(Chen, Song, & Heo, 2018)를 질의어로 하여 제목과 초록에 해당 단어가 포함되어 있는 문헌 데이터를 수집했다. 해당 연도의 특성을 확인하고 분석하기 위해 충분한 데이터가 확보될 수 있는 1990년부터 2016년까지 총 27년으로 출판연도를 한정하였다. 한 예로 불확실성 단어 'contradictory'에 대한 쿼리는 "(contradictory[Title/Abstract]) AND ("1990" [Date - Publication]: "2016" [Date - Publication])"과 같다. 총 문헌 데이터는 2,489,466건이며, 데이터 크기는 26.1GB이다. 'Suspected' 단어가 135,705건으로 가장 높은 빈도수를 차지했으며, 'misconstrues' 단어가 19건으로 가장 낮은 빈도수를 차지했다.

3.2 XML 파싱

PubMed로부터 수집한 XML형식의 문헌 데이터는 문헌의 일관성을 유지하기 위해 데이터 구조를 마크업 문서에 대한 논리 구조로 정의한 문서형 정의(Document Type Definition, DTD)를 기준으로 구성된다. 대량의 데이터 처리에 유용한 기법인 SAX(Simple API for XML) 파서를 적용하여 XML 파싱을 수행하였다. 2017년 1월 1일 자의 MEDLINE/PubMed DTD를 기반으로 총 45개의 요소(elements)중 본 연구에서 필요한 4가지 요소인 〈PMID〉, 〈PubDate〉, 〈ArticleTitle〉, 〈Abstract〉, 〈AbstractText〉의 태그명, 속성명, 속성값 및 요소 내용을 추출했다. 〈PMID〉는 PubMed 문헌의 고유 식별자이며, 〈PubDate〉는 저널 출판 연도, 〈ArticleTitle〉은 문헌의 제목, 〈Abstract〉와 〈AbstractText〉는 문헌의 초록을 의미한다.

3.3 Semantic MEDLINE DB

SemMedDB는 규칙 기반의 의미 해석 프로그램(semantic interpreter)인 SemRep(Rindflesch & Fiszman, 2003)을 이용하여 PubMed의 제목과 초록으로부터 의미적 술어를 추출한 데이터를 보유하고 있는 데이터베이스이다(Kilicoglu et al., 2012). 의미적 술어는 미국 국립의학도서관(National Library of Medicine, NLM)에서 제공하는 생의학 분야의 각종 용어를 통합 개념체계로 구성한 생의학 온톨로지인 UMLS (Unified Medical Language System)의 지식 정보원(UMLS knowledge sources) (Bodenreider, 2004)으로부터 추출된다. UMLS의 지식 정보원 중 하나인 의미망(semantics network)에서 해당 개체에 대한 의미 유형(semantic types)과 개체들 간의 관계성(relationships)을 정의하였고, 메타시소러스(metathesaurus)에서 해당 개체에 대한 대표 개념인 CUI(Concept Unique Identifier)가 정의되어 있다. 의미적 술어는 주어-서술어-목적어(S-P-O)의 트리플(triple) 구조로 구성되어 있다. 주어와 목적어 쌍은 메타시소러스 개념(metathesaurus concepts)으로 구성되며, 술어는 의미망의 관계 유형(relation type)과 일치한다.

SemMedDB는 생의학 영역의 지식 발견을 수행하는데 도움을 주는 지식 자원(knowledge resource)으로 알려져 있으며, PubMed(26,700,000

건의 문헌, 2016년 12월 31일 기준)로부터 약 89,200,000건의 의미적 술어에 대한 정보를 소장하고 있다.

본 연구에서는 가장 최신 버전인 SemmedVER30을 적용하였으며, 테이블은 총 5개로 CITATIONS, GENERIC_CONCEPT, PREDICATION, PREDICATION_AUX, SENTENCE로 구성되어 있다. 이 중 본 연구에서 주로 참조한 테이블은 PREDICATION으로 총 12개의 필드로 구성되어 있다.

즉, 문헌의 PMID를 기반으로 SemMedDB의 PREDICATION 테이블에 저장된 정보들을 추출했다. PREDICATION에 존재하는 트리플 구조를 가지는 고유 문헌 수는 총 1,768,757건으로 수집한 문장의 71% 비율을 차지했다. 고유 문장 수는 7,023,380이며, 전체적으로 11,520,923건의 의미적 술어 매칭 결과가 나타났다. 본 연구의 분석에 필요한 7가지 정보인 SUBJECT_NAME, SUBJECT_SEMTYPE, PREDICATE, OBJECT_NAME, OBJECT_SEMTYPE, PMID, SENTENCE_ID의 정보를 추출했다.

SemMedDB 매칭 결과 중에서도 불확실성 단어를 포함하는 문장의 특성을 확인하기 위해 SENTENCE_ID를 기반으로 SENTENCE 테이블의 문장을 모두 추출한 후 불확실성 단어를 포함하는 문장 749,120건에 대한 SemMedDB의 의미적 술어 결과를 추출하였다. 의미적 술어는 총 1,242,704건이며 고유 문헌 수는 643,103건이다. 전체 데이터 집합의 의미적 술어 결과 중 불확실성 단어를 포함하는 의미적 술어 결과가 차지하는 비율은 10.79%이다.

3.4 Negation Detection

분석 대상이 되는 불확실성 단어 데이터 집합을 정교하게 만들기 위해 문장 내 불확실성 단어를 기준으로 부정 표현(negation)의 존재 여부를 확인하였다. 규칙 기반의 부정 표현 발견(negation detection) 기법 중 임상(clinical) 텍스트로부터 부정 표현을 확인하기 위해 고안된 대표적인 알고리즘인 NegEx(Chapman et al., 2001) 알고리즘을 적용했다.

본 연구에서는 부정 표현의 상태뿐만 아니라 문장 내의 문맥 정보인 시간성(temporality)과 경험자(experiencer)도 주석이 가능하도록 구현된 Solti, Cooke, Xia, Wurfel(2009)의 GenNegEx 알고리즘을 적용했다. GenNegEx v1.2에서 2,376건의 테스트 문장(negated: 491, affirmed: 1,885)에 대해 성능 평가를 수행한 결과 F척도 94%의 높은 성능을 보였다.

불확실성 단어가 포함된 문장 총 749,120건의 부정 표현을 확인하기 위해 초기에 설정된 NegEx 트리거 단어를 이용하였고, 부정 표현 발견 결과로 5,441건의 문장이 발견되었다. 해당 문장들은 9,021건의 의미적 술어로 구성되어 있었다. 이 문장들은 불확실성 단어가 포함되어있지만 불확실성 단어를 기준으로 부정 표현이 함께 존재하므로 결과적으로 불확실성을 의미하지 않기에 분석 대상에서 제외되어 기본 데이터 집합으로 이동하였다.

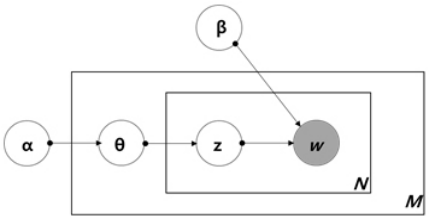
최종 분석 대상 데이터 집합은 〈표 1〉과 같다. 분석 데이터 집합의 총 문헌 수는 1,768,757건이다. 이 중 불확실성 단어를 포함한 데이터 집합의 문헌 수는 640,232건이고 문장은 743,679건이며, 이에 대한 의미적 술어는 총 1,233,683건이다.

〈표 1〉 최종 데이터 집합

	문헌	문장	의미적 술어
데이터 집합	1,768,757	7,023,380	11,520,923
불확실성 데이터 집합	640,232	743,679	1,233,683
기본 데이터 집합	1,701,959	6,279,701	10,287,240

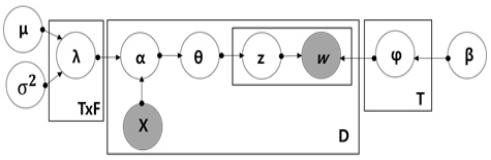
3.5 DMR 토픽 모델

〈그림 2〉는 LDA 모델(Blei, Ng, & Jordan, 2003)을 그래픽화한 것이다. LDA의 기본 가정은 단어들의 분포로 표현되는 각 토픽에서 토픽들 간의 랜덤 혼합(random mixture)으로 문헌들이 구성되는 것으로 본다. N차원의 벡터에서 확률 분포를 계산할 때 깁스 샘플링(gibbs sampling)을 적용하며, 샘플링의 기준에 따라 세 가지의 레벨을 가진다. 파라미터 α와 β는 코퍼스 레벨의 파라미터들이며 코퍼스를 생성하는 과정에서 한 번 샘플된다. Θ는 문헌 레벨의 변수로 각 문헌마다 한 번씩 샘플된다. 마지막으로 z와 w는 단어 레벨의 변수들로 각 문헌 내의 각 단어마다 한 번씩 샘플된다. M은 문헌의 개수이며, N은 문헌에 속한 단어의 개수이다. Θ는 문헌의 주제 분포이며, z는 해당 단어가 속한 주제를 의미한다. α는 문헌의 주제 분포를 조절하는 파라미터이며, β는 주제의 단어 분포를 조절하는 파라미터로 보통 각 0.1과 0.001로 설정한다. 이 값이 1에 가까워질수록 문헌에 많은 주제가 포함되고, 주제에 많은 단어가 포함되는 것이다. 문헌 내 단어 w를 관측하여 z의 값을 정하고, 주제 분포를 업데이트하여 가장 적합한 z값을 찾아내어 잠재된(latent) 문헌별 주제 분포와 주제별 단어 분포를 추론해낸다.



〈그림 2〉 LDA 모델

DMR 토픽 모델(Mimno & McCallum, 2012)의 그래픽 표현은 〈그림 3〉과 같다. 〈그림 2〉에서 문헌 밖에 존재했던 파라미터 α는 문헌 안으로 포함되었고, 두 파라미터 x와 λ로부터 영향을 받는다. x는 메타데이터이며, λ는 평균 μ와 표준편차 σ인 정규분포를 따르는 메타데이터별 파라미터로 주제 개수 T와 메타데이터 개수 F에 따라 T x F 차원의 행렬로 표현할 수 있다. φ는 단어차원의 벡터로 파라미터 β로부터 영향을 받는다. 문헌 생성 과정은 두 단계로 이루어진다. 우선 주제별로 메타데이터의 분포와 단어 분포를 추출한다. 다음은 모든 주제에 대해 문헌의 주제 분포를 결정하는 파라미터 α를 구하고, 모든 단어에 대해 주제 분포에서 특정 주제를 추출하여 주제 내에서 특정 단어를 추출하는 식으로 문헌을 생성한다.



〈그림 3〉 DMR 모델

즉, LDA와의 차이점은 λ가 주제 분포에 영향을 미치고 또 단어의 주제 z에 영향을 미친다. 주제 z에서 단어 w가 추출된다. 관찰 가능

한 w와 w에 임의로 배정된 z를 통해서 가능성이 높은 λ를 추출할 수 있다. 이 λ를 기반으로 α를 구하고, 이 값에 따라 각 단어들의 주제 배정을 업데이트한다. 이를 수렴할 때까지 반복하게 되면 모든 단어들이 적절한 주제에 배치되고 이에 따라 문헌의 주제 분포와 주제의 단어 분포, 그리고 메타데이터별 하이퍼 파라미터 λ를 얻는다. 최종적으로 산출된 λ값을 통해서 해당 메타데이터에 속하는 문헌들의 주제 분포를 예상할 수 있다.

앞서 기술한대로 DMR의 수행 목적인 불확실성 단어를 포함한 문장의 주제 특성을 파악하기 위해 불확실성 단어를 포함한 743,679건의 문장을 대상으로 추출한 SemMedDB의 의미적 술어 결과를 적용했다. DMR의 입력 파일의 형태는 식별 가능한 고유 ID와 토픽 분포의 기준으로 사용할 메타데이터, 그리고 토픽을 구성하는 단어들의 집합으로 구성된다. 본 연구에서 고유 ID는 PMID, 메타데이터는 출판연도, 단어들은 의미적 술어의 주어와 목적어 개체들을 대상으로 설정했다. 개체들은 구(phrase)형태로 구성되어 있으므로 하나의 개체로 인식하기 위해 공백(white space)을 언더스코어('_')로 변환하였다. 한 문장에 포함된 주어 개체와 목적어 개체를 공백으로 구분하여 함께 나열한 후 PMID를 기준으로 동일 문헌에 속한 개체들을 모두 통합했다. DMR 토픽 모델링의 입력 정보의 예시는 〈표 2〉와 같다. PMID '19542758'의 출판연도는 2009년이며 한 문장에서 'Klinefelter's_Syndrome'와 'Boys' 개체가 출현했다. PMID '15982456'의 2005년 문헌에서는 두 문장에서 총 4개의 개체가 출현했다.

〈표 2〉 DMR 토픽 모델링의 입력 정보 예시

PMID	Year	Entities
19542758	2009	Klinefelter's_Syndrome Boys
2031881	1991	Excision Therapeutic_procedure
15982456	2005	Embryo House_mice Encounter_due_to_In_vitro_fertilization Human

DMR 실험은 자연어 처리(Natural Language Processing, NLP)와 문헌 분류(document classification), 클러스터링 등 자바 기반의 다양한 기계 학습 도구를 제공하는 MALLET (MAchine Learning for LanguagE Toolkit)의 토픽 모델링을 이용했다(McCallum, 2002). 토픽의 개수는 20으로 설정하였고, 샘플링의 횟수를 의미하는 이터레이션(iteration)은 1000, 특정 토픽들이 다른 토픽들에 비해 더 우세하도록 표현하기 위해 모델이 데이터에 적합(fit)하게 하이퍼 파라미터를 최적화하는 인터벌(interval)은 100으로 설정했다.

4. 실험 결과 분석

4.1 생의학 지식 결과 통계 분석

불확실성 단어 문장 수는 743,679건으로 전체 문장의 10.59%를 차지하며, 불확실성 단어를 포함하지 않는 문장은 6,279,701건으로 89.41%이다. 또한 각 문장에서 SemMedDB를 통해 추출한 의미적 술어의 수는 불확실성 단어의 데이터 집합이 1,233,683건으로 10.71%, 불확실성 단어를 포함하지 않는 데이터 집합이 10,287,240건으로 89.29%를 차지한다. SemMed

DB로부터 추출한 전체 11,520,923건의 생의학 명제인 의미적 술어의 특성을 확인하기 위해 각 생의학 정보의 빈도수와 비율을 확인했다.

〈표 3〉은 빈도수 기반 상위 10개의 개체 정보를 나타낸다. 총 156,222건의 개체가 출현했으며 개체는 주어와 목적어를 합산하여 총 23,041,846건의 빈도수이다. 1순위 단어인 'Patients'는 200만건 이상의 빈도수로 전체 단어의 8.77%를 차지하여 나머지 단어들에 비해 월등히 높은 값을 가졌다.

〈표 3〉 상위 10개 개체 정보

순위	개체	빈도	비율(%)
1	Patients	2,020,564	8.77
2	Therapeutic procedure	425,601	1.85
3	Child	216,834	0.94
4	Disease	195,592	0.85
5	Woman	177,709	0.77
6	Human	177,186	0.77
7	Cells	160,376	0.7
8	Malignant Neoplasms	106,337	0.46
9	Rattus norvegicus	104,120	0.45
10	Neoplasm	103,994	0.45

동사는 11,520,923건의 데이터에서 62개의 고유한 관계 유형이 출현했다. 〈표 4〉는 상위 10개의 관계 유형을 나타낸다. 관계 유형 'PROCESS_OF'가 전체 데이터의 22.82%를 차지했다.

각 개체의 통합된 의미 유형은 〈표 5〉와 같다. 고유한 의미 유형은 총 133건이며 총 빈도수는 23,041,846건이다. 'Human'이 1순위로 나타났으며 전체 개체 의미 유형의 13%를 차지했다.

〈표 4〉 상위 10개 관계 유형 정보

순위	관계 유형	빈도	비율(%)
1	PROCESS_OF	2,628,529	22.82
2	LOCATION_OF	1,643,974	14.27
3	PART_OF	1,237,617	10.74
4	TREATS	1,176,844	10.21
5	ISA	841,408	7.3
6	AFFECTS	516,480	4.48
7	USES	495,823	4.3
8	COEXISTS_WITH	483,557	4.2
9	INTERACTS_WITH	312,521	2.71
10	CAUSES	308,078	2.67

〈표 5〉 상위 10개 개체 의미 유형 정보

순위	개체 의미 유형	빈도	비율(%)
1	Human	2,994,615	13
2	Disease or Syndrome	2,815,994	12.22
3	Nucleic Acid, Nucleoside, or Nucleotide	2,070,369	8.99
4	Therapeutic or Preventive Procedure	1,499,137	6.51
5	Body Part, Organ, or Organ Component	1,207,647	5.24
6	Neoplastic Process	1,154,228	5.01
7	Gene or Genome	898,629	3.9
8	Pharmacologic Substance	680,547	2.95
9	Cell	666,495	2.89
10	Finding	615,097	2.67

의미적 술어의 정보는 총 11,520,923건의 데이터에서 3,871,653건의 고유한 의미적 술어가 출현했다. 〈표 6〉은 상위 10개의 의미적 술어 정보를 나타낸다. 모든 의미적 술어가 목적어로 'patients'를 포함하고 있다. 또한 관계 유형인 술어는 1건을 제외한 모든 상위 관계 유형이

'PROCESS_OF'로 구성되어 있다. 1회 출현한 의미적 술어는 2,821,937건으로 전체 데이터의 24%를 차지했다.

〈표 6〉 상위 10개 의미적 술어 정보

순위	의미적 술어(S-P-O)	빈도	비율(%)
1	Therapeutic procedure TREATS Patients	24,081	0.21
2	Disease PROCESS_OF Patients	20,930	0.18
3	Operative Surgical Procedures TREATS Patients	18,259	0.16
4	Symptoms PROCESS_OF Patients	16,347	0.14
5	Malignant Neoplasms PROCESS_OF Patients	13,051	0.11
6	Lesion PROCESS_OF Patients	12,223	0.11
7	Excision TREATS Patients	11,296	0.1
8	Diabetic PROCESS_OF Patients	10,959	0.1
9	Neoplasm PROCESS_OF Patients	10,280	0.09
10	Schizophrenia PROCESS_OF Patients	10,268	0.09

의미적 술어의 두 개체인 주어와 목적어 쌍의 빈도수를 산출한 결과는 〈표 7〉과 같다. 두 개체는 알파벳 순서로 정렬하여 중복 값을 합산한 고유 쌍으로 구성했다. 'Patients'와 'Therapeutic procedure' 쌍이 33,378회 출현하여 0.29%로 가장 높은 빈도수와 비율을 차지했다. 반면 1회 출현한 개체 쌍은 2,034,896건으로 전체 데이터의 17.66%를 차지했다.

〈표 7〉 상위 10개 개체 쌍 정보

순위	개체1	개체2	빈도	비율(%)
1	Patients	Therapeutic procedure	33,378	0.29
2	Disease	Patients	21,863	0.19
3	Operative Surgical Procedures	Patients	20,350	0.18
4	Patients	Symptoms	16,877	0.15
5	Malignant Neoplasms	Patients	13,677	0.12
6	Excision	Patients	12,505	0.11
7	Lesion	Patients	12,367	0.11
8	Diabetic	Patients	10,960	0.1
9	Neoplasm	Patients	10,919	0.09
10	Complication	Patients	10,372	0.09

4.2 불확실성 데이터 집합의 연도별 시계열 분석

불확실성 단어 기반 생의학 지식의 시계열적 변화 추이를 파악하기 위해 연도 기반 불확실성 단어 데이터 집합의 특성을 살펴보았다. 〈표 1〉의 최종 데이터 집합에서 연도 데이터가 영문자를 포함하거나 영문자와 숫자의 조합으로 구성되어 연도 구분이 불가한 데이터는 제외한 후 데이터 집합의 통계를 산출하였다.

SemMedDB의 의미적 술어를 기반으로 전체 데이터 집합에서 불확실성 단어 데이터 집합의 연도별 출현 빈도와 비율을 확인하였으며 불확실성 단어 데이터 집합의 연도별 출현 빈도와 비율은 〈표 8〉과 〈그림 4〉와 같다. 연도는 문헌 단위로 산출되지만 본 결과는 SemMedDB로 추출된 의미적 술어의 문헌 수를 기반으로 가중치가 부여된 결과이다. 즉, 한 문헌에서 불확실성 단어를 포함한 문장의 의미적 술어가 3건일 경우 해당 문헌의 연도는 3회 카운트 된다.

〈표 8〉 연도별 데이터 집합

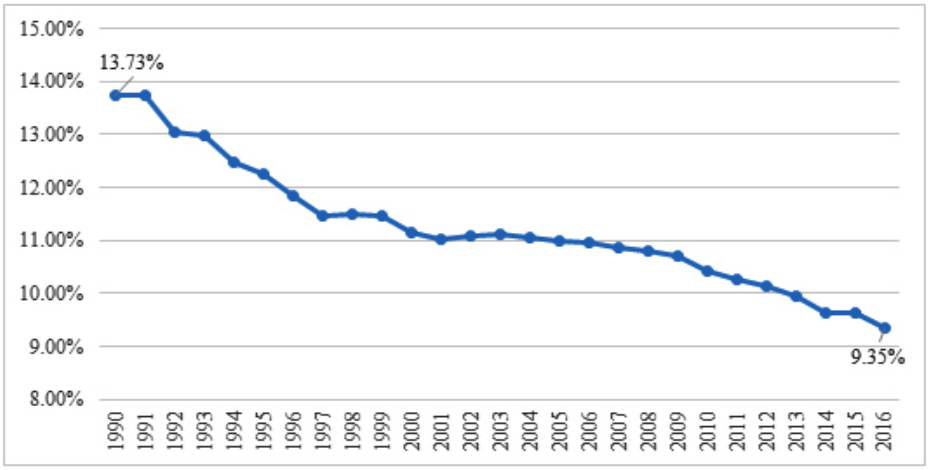
연도	전체 데이터 집합	불확실성 단어 데이터 집합	비율(%)
1990	146,981	20,184	13.73
1991	156,500	21,512	13.75
1992	173,659	22,641	13.04
1993	188,008	24,407	12.98
1994	203,614	25,429	12.49
1995	226,089	27,743	12.27
1996	246,997	29,283	11.86
1997	265,898	30,516	11.48
1998	284,662	32,698	11.49
1999	299,565	34,395	11.48
2000	329,418	36,702	11.14
2001	338,473	37,357	11.04
2002	354,233	39,284	11.09
2003	379,794	42,197	11.11
2004	407,665	45,014	11.04
2005	431,915	47,481	10.99
2006	455,079	49,898	10.96
2007	475,413	51,610	10.86
2008	502,736	54,294	10.80
2009	527,326	56,487	10.71
2010	563,628	58,725	10.42
2011	606,280	62,191	10.26
2012	673,008	68,139	10.12
2013	730,240	72,648	9.95
2014	776,431	74,771	9.63
2015	816,731	78,623	9.63
2016	850,843	79,540	9.35
총계	11,411,186	1,223,769	10.72

〈표 8〉의 연도별 데이터 집합의 분포에서 확인할 수 있듯이 총 11,411,186건 중 10,187,417건(89.28%)의 불확실성 단어가 포함되지 않은 데이터 집합과 1,223,769건(10.72%)의 불확실성 단어 데이터 집합으로 구성된다. 두 데이터 집합 모두 시간의 흐름에 따라 데이터 집합이 증가하는 추세를 보였다. 〈그림 4〉는 연도별 불확실성 단어 데이터 집합의 비율을 나타낸다. 1990년에 13.73%에서 2016년 9.35%까지 지속적으로 감소하는 패턴을 보였다.

4.3 DMR 토픽 모델링 결과 분석

4.3.1 토픽 모델링 결과 분석

불확실성 단어 데이터에 속한 개체들의 특성을 면밀히 살펴보기 위해 SemMedDB를 통해



〈그림 4〉 연도별 불확실성 단어 데이터 집합의 비율

추출된 개체들의 토픽 구성을 살펴보았으며 연도의 흐름에 따른 개체 기반 토픽의 분포 변화를 확인했다.

20개 토픽의 상위 5개 대표 개체는 〈표 9〉와 같다. 각 토픽의 레이블은 상위 개체와 고유 개체를 중심으로 공통된 주제를 나타낼 수 있도록 연구자에 의한 1차 레이블링 후 전문가에 의한 2차 검증을 통해 완성되었다. 또한 상위 30개까지의 개체들을 확인하여 중복 개체와 고유 개체를 확인했으며, 총 600개의 개체 중 해당 토픽에만 출현한 고유한 개체는 '*'로 표시하였다. 〈표 10〉은 토픽 모델링 결과 20개 토픽을 구성하는 개체들의 빈도 순위를 나타낸다. 총 435개의 고유 개체 중 3회 이상 출현한 34개의 개체 정보이다. 'patients'와 'disease' 개체가 총 13회 출현하여 가장 높은 순위를 차지했다.

4.3.2 연도별 토픽 분포 패턴 분석

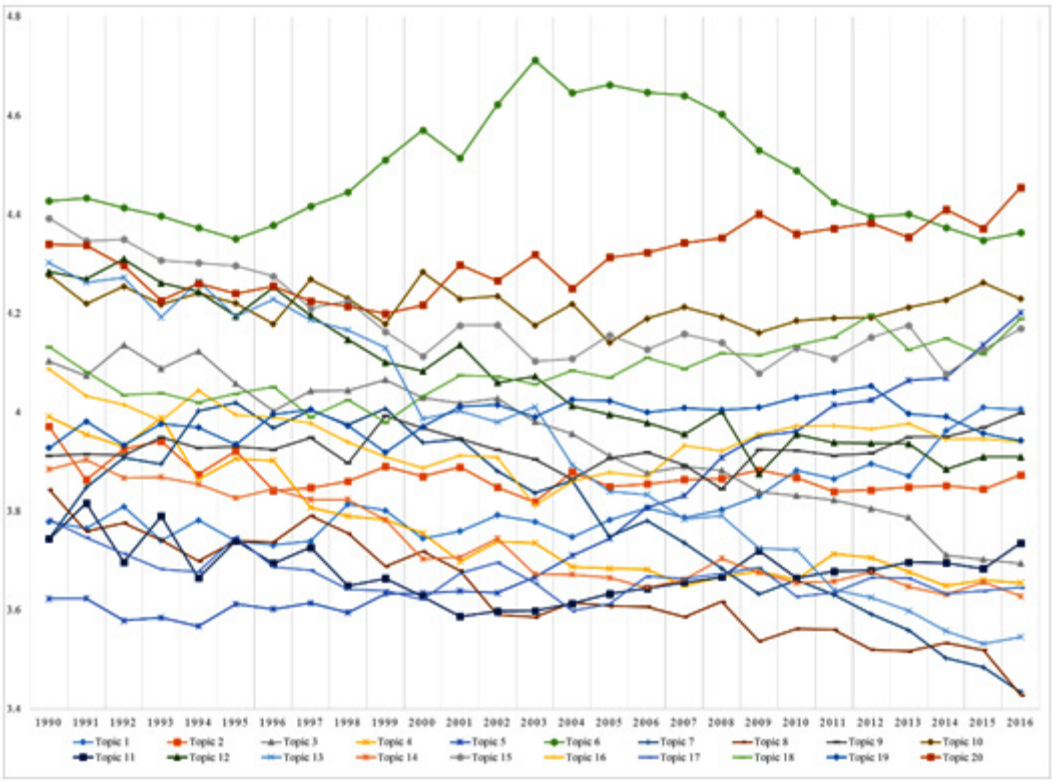
불확실성 단어를 포함한 개체기반 토픽의 특성을 시계열적으로 분석하기 위해 연도 메타데이터를 포함한 DMR 토픽 모델링 결과로 각 토픽의 연도별 분포를 확인했다. 〈그림 5〉는 총 20개 토픽의 연도별 분포를 나타낸다. 1990년과 2016년의 각 토픽 분포의 시계열적 차이를 살펴본 결과 20개 토픽 중 14개 토픽이 2016년에 더 작은 값을 가졌다. 두 연도의 토픽 분포 값의 차이에 대한 총 20개 토픽의 평균은 0.125로 나타났다. 또한 데이터의 연도별 분포를 객관적으로 확인하기 위해 〈표 11〉과 같이 각 토픽별 표준편차와 순위를 계산하였다. 토픽의 표준편차가 크다는 것은 토픽의 연도별 분포의 차이가 크다는 것으로 해당 토픽 연구의 불확실성의 변화가 크다고 해석할 수 있으며, 표준편차가 작다는 것은 토픽의 연도별 분포의 차이가 작은 것으로 해당 토픽 연구의 불확실성의 변화가 작은 것으로 해석할 수 있다. 표준편차 범위는 토픽 9(parkinson_disease)의 0.033부터 토픽 13(neurons)의 0.255까지이다.

〈표 9〉 20개 토픽의 상위 5개 개체 토픽 모델링 결과

Topic 1 malignant_neoplasms	Topic 2 congenital_abnormality	Topic 3 cell_membrane	Topic 4 infection	Topic 5 diabetes
patients	child	proteins	patients	persons*
operative_surgical_procedures*	infant*	cells	infection*	individual
malignant_neoplasms	adult	ervk*	disease	obesity*
therapeutic_procedure	infant_newborn	amino_acids*	tuberculosis*	diabetes*
excision*	disease	enzymes*	pneumonia*	elderly*
Topic 6 antipsychotic	Topic 7 gene	Topic 8 neoplasm	Topic 9 parkinson__disease	Topic 10 pregnancy
patients	cells	neoplasm	patients	woman
atypical_antipsychotic*	interleukin*	lesion	disease	male_population_group
antipsychotic_agents*	human	malignant_neoplasms	syndrome	patients
schizophrenia*	cd*	breast*	parkinson_disease*	unexplained_infertility*
therapeutic_procedure	tp*	lung*	family	lipids
Topic 11 myocardial_infarction	Topic 12 carcinogens	Topic 13 neurons	Topic 14 biopsy	Topic 15 autoimmune_diseases
patients	human	rattus_norvegicus	patients	patients
coronary_arteriosclerosis*	rattus_norvegicus	neurons*	diagnosis	disease
therapeutic_procedure	mus	cells	magnetic_resonance_imaging	serum
heart_failure*	brain	brain	lesion	eye*
myocardial_infarction*	liver	human	biopsy*	antibodies*
Topic 16 wounds_and_injuries	Topic 17 crohns_disease	Topic 18 therapeutic_procedure	Topic 19 DNA and RNA	Topic 20 epilepsy
patients	patients	therapeutic_procedure	disease	epilepsy*
injury	symptoms	patients	genes*	water*
fracture*	disease	pharmaceutical_preparations*	human	human
wounds_and_injuries*	crohns_disease*	adrenal_cortex_hormones*	dna*	network*
congenital_abnormality*	pain	prophylactic_treatment*	dna_sequence*	seizures*

〈표 10〉 토픽 모델링 결과 개체 빈도

순위	개체	빈도	순위	개체	빈도
1	patients	13	18	woman	3
2	disease	13	19	rattus_norvegicus	3
3	human	10	20	male_population_group	3
4	therapeutic_procedure	9	21	genes	3
5	cells	6	22	mus	3
6	complication	6	23	brain	3
7	operative_surgical_procedures	5	24	lung	3
8	symptoms	5	25	liver	3
9	syndrome	5	26	biopsy	3
10	child	4	27	consensus_sequence	3
11	individual	4	28	fever	3
12	lesion	4	29	asthma	3
13	diagnosis	4	30	pathogenesis	3
14	malignant_neoplasms	4	31	dementia	3
15	adult	4	32	rheumatoid_arthritis	3
16	proteins	3	33	house_mice	3
17	neoplasm	3	34	management_procedure	3

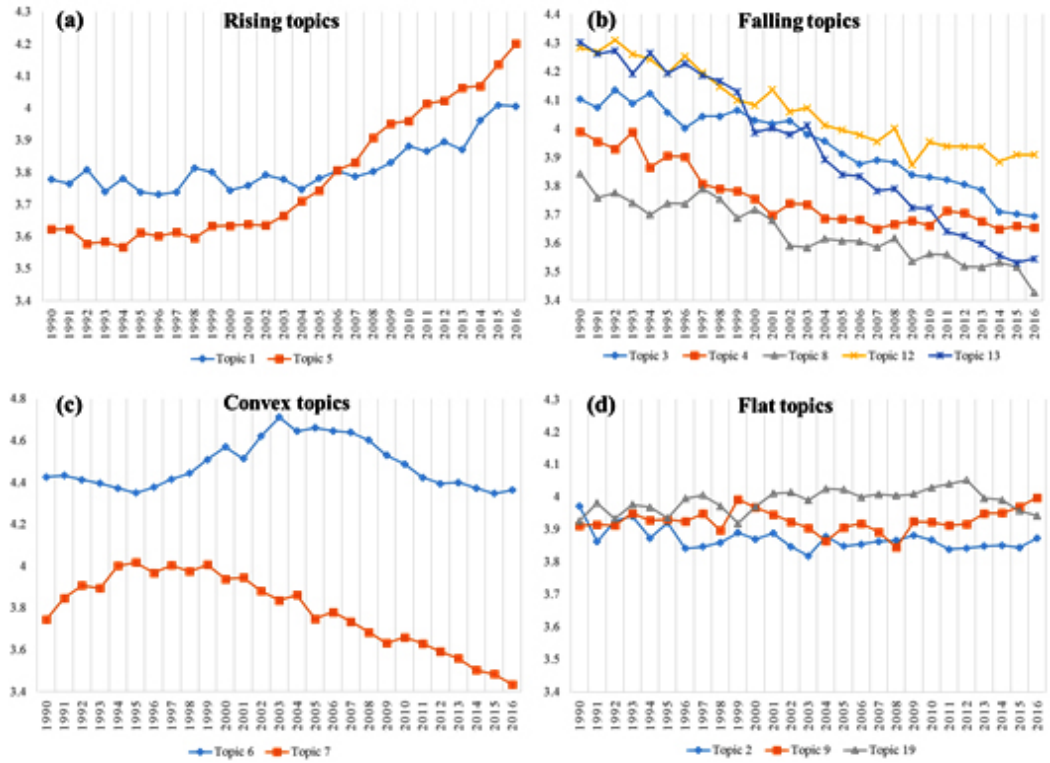


〈그림 5〉 20개 토픽 분포

〈표 11〉 토픽의 표준편차와 순위

토픽	표준편차	순위	토픽	표준편차	순위
Topic 1	0.077	11	Topic 11	0.055	15
Topic 2	0.034	19	Topic 12	0.138	4
Topic 3	0.132	5	Topic 13	0.255	1
Topic 4	0.111	7	Topic 14	0.091	9
Topic 5	0.198	2	Topic 15	0.088	10
Topic 6	0.111	6	Topic 16	0.059	13
Topic 7	0.175	3	Topic 17	0.042	16
Topic 8	0.103	8	Topic 18	0.055	14
Topic 9	0.033	20	Topic 19	0.035	18
Topic 10	0.035	17	Topic 20	0.067	12

보다 면밀한 패턴 분석을 수행하기 위해 패턴의 특성에 따라 상승(rising), 하강(falling), 볼록(convex), 평평(flat)의 4가지의 패턴으로 구분했다. 〈그림 6〉은 각 패턴별 대표적인 토픽의 연도별 분포를 나타낸다.



〈그림 6〉 토픽 분포의 4가지 패턴: (a) rising, (b) falling, (c)convex, (d) flat

(a)는 상승하는 패턴으로 토픽 1과 5(malignant_neoplasms, diabetes)가 시간의 흐름에 따라 토픽의 확률 분포가 점차 증가하고 있다. (b)는 하강하는 패턴으로 토픽 3, 4, 8, 12, 13(cell_membrane, infection, neoplasm, carcinogens, neurons)이 시간의 흐름에 따라 토픽 분포가 점차 감소하고 있다. (c)는 볼록한 패턴을 보이는 토픽 6과 7(antipsychotic, gene)이다. 토픽 6(antipsychotic)은 2003년도에 전체 토픽 분포 중에서 최댓값인 4.71을 가진 토픽이며, 토픽 7(gene)은 1995년도에 해당 토픽 내에서 최댓값 4.02를 가진 토픽이다. (d)는 (a)-(c)의 패턴을 제외하고 일반적인 패턴을 보인 대부분의 토픽 중에서 표준편차가 가장 작은 대표적인 3개의 토픽 2, 9, 19(congenital_abnormality, parkinson_disease, DNA and RNA)이다.

전체적인 토픽의 분포는 상승하는 패턴보다 하강하는 패턴의 토픽이 더 많다는 점을 확인할 수 있었다. 이는 앞서 설명한 연도별 불확실성 단어 데이터 집합의 비율이 점차 낮아지는 것과 동일한 패턴으로 해석할 수 있다. 하강패턴은 해당 토픽에 대한 연구들을 통해 불확실성이 감소하며 시간의 흐름에 따라 발전하는 안정화된 토픽으로 이해할 수 있다. 반면 상승패턴은 불확실성이 증가하며 해당 분야에 대한 연구자들의 학술 논문에 표현된 주장이 상충되거나 논의가 많이 이루어지고 있는 단계로 해석할 수 있다. 볼록한 토픽은 특정 시기에 일시

적으로 불확실성이 높아진 토픽으로 논쟁과 불일치 이후 특정 패러다임이 설립되거나 합의가 형성되어 학술 커뮤니티의 연구 영역에서 불확실성이 감소하는 방향으로 해석할 수 있다. 또한 평평한 토픽은 시간의 흐름에 따라 큰 변화를 보이지 않는 토픽들로 해당 연구 분야에 대한 연구자들의 결과물이 관심있게 이루어지지 않거나, 이미 오래전부터 안정화된 토픽들로 해석이 가능하다. 즉, 생의학 학술 문헌 내의 생의학 개체 기반으로 구성된 토픽을 분석함으로써 학술 연구의 불확실성의 측면에서 각 패턴에 따른 연구 발전 동향을 파악할 수 있었다.

5. 결 론

본 연구에서는 생의학 학술 문헌의 불확실성의 패턴을 파악하기 위한 연구로 문장 내에 존재하는 불확실성 단어 기반 생의학적 지식의 특성을 시간의 흐름에 따라 살펴보고자 했다. 이를 위해 과학적 불확실성에서 지식의 상태가 불완전한 상황을 표현하는 불확실성 단어 196개를 대상으로 PubMed에서 1990년도부터 2016년까지 총 27년간의 생의학 학술 문헌 데이터를 수집하여 SemMedDB의 의미적 술어 기반 불확실성 단어가 포함된 문장의 생의학 개체와 개체 간 관계성의 특성을 파악하였다. 전체 데이터 집합 내의 불확실성 단어 집합의 연도별 비율을 확인한 결과 1990년 13.73%에서 2016년 9.35%로 지속적으로 감소하는 패턴을 보였다. 특히 생의학 학술 문헌에서 중요한 주제적 의미를 담고 있는 생의학 개체를 대상으로 DMR 토픽 모델링 기법을 적용하여 시계열적 개체 패턴을 분석했다. 20개의 토픽은 전문가의 검증을 통해 레이블 되었고 토픽의 시계열적 패턴은 상승, 하강, 볼록, 평평의 4가지 패턴으로 구분되었다. 각 패턴은 해당 연구 분야의 불확실성의 정도와 연결지어 해석할 수 있으며, 불확실성 단어 집합의 비율이 감소하는 패턴과 동일하게 6개 토픽을 제외한 14개 토픽은 1990년대비 2016년 토픽의 분포가 더 낮은 값을 가져 전반적으로 최신년도로 갈수록 하강 패턴을 보이는 점을 확인했다. 결론적으로 생의학 학술 문헌내의 과학적 지식의 표현은 시간의 흐름에 따라 불확실성이 감소하는 경향으로 발전하고 있었다.

본 연구는 기존의 연구에서 시도되지 않았던 불확실성 단어 기반 시간의 흐름에 따른 과학적 지식의 변화 패턴에 대해 종합적으로 분석한 연구로 시사점을 지닌다. 생의학 학술 문헌의 명제로 활용도가 높은 SemMedDB의 의미적 술어를 이용하여 생의학 개체와 개체간의 관계성을 파악하였고, 불확실성 단어가 포함된 문장의 생의학 개체를 DMR 토픽 모델링 기반으로 시계열적 분석을 수행함으로써 기존의 연구에서 제안할 수 없었던 불확실성의 패턴 변화를 종합적으로 파악하였다는 점에서 의의가 있다.

향후 연구로는 현재 연구에서 수행하였던 불확실성 단어 기반의 데이터 수집에서 더 나아가 전체 학술 문헌 내에서 불확실성의 패턴을 확인하기 위해 PubMed의 전체 데이터를 이용하여 SemMedDB가 포괄하는 모든 데이터 집합을 기반으로 문헌 내 생의학적 지식을 발견하고자 한다. 기 확립된 명제를 대상으로 연도별 분석을 수행하고, 추출된 의미적 술어에 'bursts'

감지(detection)기법(Kleinberg, 2003)을 적용한다. 시간의 흐름에 따라 생의학 명제가 언제, 어떻게 학술적으로 주목을 받게 되며 문장의 불확실성이 변화하는지에 대해 불확실성 단어가 포함된 문장과 함께 분석하고자 한다.

참 고 문 헌

Acedo, F. J., & Casillas, J. C. (2005). Current paradigms in the international management field: An author co-citation analysis. International Business Review, 14(5), 619-639. http://dx.doi.org/10.1016/j.ibusrev.2005.05.003

An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. Scientometrics, 88(1), 133-144. http://dx.doi.org/10.1007/s11192-011-0374-1

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990-2004. Journal of the American Society for Information Science and Technology, 58(7), 947-957. https://doi.org/10.1002/asi.20567

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(suppl_1), D267-D270. https://doi.org/10.1093/nar/gkh061

Callon, M., Rip, A., & Law, J. (Eds.). (1986). Mapping the dynamics of science and technology: Sociology of science in the real world. Springer.

Cambrosio, A., Limoges, C., Courtial, J. P., & Laville, F. (1993). Historical scientometrics?: Mapping over 70 years of biological safety research with co-word analysis. Scientometrics, 27(2), 119-143. https://doi.org/10.1007/BF02016546

Chang, Y. W., & Huang, M. H. (2012). A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. Journal of the American Society for Information Science and Technology, 63(1), 22-33. https://doi.org/10.1002/asi.21649

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal

of biomedical informatics, 34(5), 301-310. https://doi.org/10.1006/jbin.2001.1029

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for information Science and Technology, 57(3), 359-377. https://doi.org/10.1002/asi.20317

Chen, C., Song, M., & Heo, G. E. (2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. Journal of Informetrics, 12(1), 158-180. https://doi.org/10.1016/j.joi.2017.12.004

Chen, K., & Guan, J. (2011). A bibliometric investigation of research performance in emerging nanobiopharmaceuticals. Journal of Informetrics, 5(2), 233-247. https://doi.org/10.1016/j.joi.2010.10.007

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. Journal of Informetrics, 5(1), 146-166. https://doi.org/10.1016/j.joi.2010.10.002

Culnan, M. J. (1986). The intellectual development of management information systems, 1972-1982: A co-citation analysis. Management Science, 32(2), 156-172. https://doi.org/10.1287/mnsc.32.2.156

Culnan, M. J. (1987). Mapping the intellectual structure of MIS, 1980-1985: A co-citation analysis. Mis Quarterly, 341-353. https://www.jstor.org/stable/248680

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing & Management, 37(6), 817-842. https://doi.org/10.1016/S0306-4573(00)00051-0

Falahati, R. (2006, February). The use of hedging across different disciplines and rhetorical sections of research articles. In Proceedings of the 22nd NorthWest Linguistics Conference (NWLC22), 99-112.

Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010, July). The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning - Shared Task (pp. 1-12). Association for Computational Linguistics.

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association, 1(2), 161-174. https://doi.org/10.1136/jamia.1994.95236146

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National

Academy of Sciences, 101(suppl 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Heo, G. E., Kang, K. Y., Song, M., & Lee, J. H. (2017). Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. BMC Bioinformatics, 18(7), 251. https://doi.org/10.1186/s12859-017-1640-x

Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. (2006). Exploiting semantic relations for literature-based discovery. In AMIA annual symposium proceedings (Vol. 2006, p. 349). American Medical Informatics Association.

Hyland, K. (1998). Hedging in scientific research articles (Vol. 54). John Benjamins Publishing.

Jensen, J. D. (2008). Scientific uncertainty in news coverage of cancer research: Effects of hedging on scientists' and journalists' credibility. Human Communication Research, 34(3), 347-369. https://doi.org/10.1111/j.1468-2958.2008.00324.x

Jeong, Y. K., Heo, G. E., Kang, K. Y., Yoon, D. S., & Song, M. (2016). Trajectory analysis of drug-research trends in pancreatic cancer on PubMed and ClinicalTrials. gov. Journal of Informetrics, 10(1), 273-285 https://doi.org/10.1016/j.joi.2016.01.003

Jin, Y., Myaeng, S. H., & Jung, Y. (2007). Use of place information for improved event tracking. Information Processing & Management, 43(2), 365-378. https://doi.org/10.1016/j.ipm.2006.07.007

Kilicoglu, H., Rosemblat, G., & Rindflesch, T. C. (2017). Assigning factuality values to semantic relations extracted from biomedical research literature. PloS One, 12(7), e0179926. https://doi.org/10.1371/journal.pone.0179926

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics, 28(23), 3158-3160. https://doi.org/10.1093/bioinformatics/bts591

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery, 7(4), 373-397. https://doi.org/10.1023/A:1024940629314

Lakoff, G. (1972). Hedges: A study in meaning criteria and the logic of fuzzy concepts. Papers from the eighth regional meeting, Chicago Linguistic Society, Chicago: University of Chicago Linguistics Department, 8, 183-228. https://doi.org/10.1007/978-94-010-1756-5_9

Light, M., Qiu, X. Y., & Srinivasan, P. (2004). The language of bioscience: Facts, speculations, and statements in between. In HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases.

Liu, D. R., Omar, H., Liou, C. H., Chi, H. C., & Hsu, C. H. (2015). Recommending blog articles

based on popular event trend analysis. Information Sciences, 305, 302-319. https://doi.org/10.1016/j.ins.2015.02.003

Liu, G. Y., Hu, J. M., & Wang, H. L. (2012). A co-word analysis of digital library field in China. Scientometrics, 91(1), 203-217. http://dx.doi.org/10.1007/s11192-011-0586-4

Malhotra, A., Younesi, E., Gurulingappa, H., & Hofmann-Apitius, M. (2013). 'HypothesisFinder:' A strategy for the detection of speculative statements in scientific text. PLoS Computational Biology, 9(7), e1003117. https://doi.org/10.1371/journal.pcbi.1003117

Malin, B., & Carley, K. (2007). A longitudinal social network analysis of the editorial boards of medical informatics and bioinformatics journals. Journal of the American Medical Informatics Association, 14(3), 340-348. http://dx.doi.org/10.1197/jamia.M2228

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 992-999).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (pp. 3111-3119).

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. Journal of the American Society for Information Science and Technology, 62(10), 1933-1953. http://dx.doi.org/10.1002/asi.21602

Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv Preprint arXiv:1206.3278.

Nerur, S. P., Rasheed, A. A., & Natarajan, V. (2008). The intellectual structure of the strategic management field: An author co-citation analysis. Strategic Management Journal, 29(3), 319-336. https://doi.org/10.1002/smj.659

Newman, D. J., & Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. Journal of the American Society for Information Science and Technology, 57(6), 753-767. https://doi.org/10.1002/asi.20342

Peters, H., & Van Raan, A. (1991). Structuring scientific activities by co-author analysis: An expercise on a university faculty level. Scientometrics, 20(1), 235-255. https://doi.org/10.1007/BF02018157

Pilkington, A., & Meredith, J. (2009). The evolution of the intellectual structure of operations

management－1980-2006: A citation/co-citation analysis. Journal of Operations Management, 27(3), 185-202. https://doi.org/10.1016/j.jom.2008.08.001

Ravetz, J. R. (1973). Scientific knowledge and its social problems. Transaction publishers.

Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics, 36(6), 462-477. https://doi.org/10.1016/j.jbi.2003.11.003

Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. Scientometrics, 6(6), 381-400. https://doi.org/10.1007/BF02025827

Rizomilioti, V. (2006). Exploring epistemic modality in academic discourse using corpora. In Information Technology in Languages for Specific Purposes, 53-71. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-28624-2_4

Sebastian, Y., Siew, E. G., & Orimaye, S. O. (2017). Emerging approaches in literature-based discovery: Techniques and performance review. The Knowledge Engineering Review, 32. https://doi.org/10.1017/S0269888917000042

Solti, I., Cooke, C. R., Xia, F., & Wurfel, M. M. (2009, November). Automated classification of radiology reports for acute lung injury: Comparison of keyword and machine learning based natural language processing approaches. In 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, 314-319. IEEE. https://doi.org/10.1109/BIBMW.2009.5332081

Song, M., Heo, G. E., & Lee, D. (2015). Identifying the landscape of Alzheimer's disease research with network and content analysis. Scientometrics, 102(1), 905-927. https://doi.org/10.1007/s11192-014-1372-x

Song, M., Kim, S., Zhang, G., Ding, Y., & Chambers, T. (2014). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. Journal of the Association for Information Science and Technology, 65(2), 352-371. https://doi.org/10.1002/asi.22970

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of Latent Semantic Analysis, 427(7), 424-440.

Szarvas, G., Vincze, V., Farkas, R., & Csirik, J. (2008, June). The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 38-45. Association for Computational Linguistics.

Uzun, A. (2002). Library and information science research in developing countries and eastern

european countries: A brief bibliometric perspective. International Information & Library Review, 34(1), 21-33. https://doi.org/10.1080/10572317.2002.10762561

Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics, 9(11), S9. https://doi.org/10.1186/1471-2105-9-S11-S9

Vold, E. T. (2006). Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study. International Journal of Applied Linguistics, 16(1), 61-87. https://doi.org/10.1111/j.1473-4192.2006.00106.x

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. Journal of the American SOCIEty for Information Science, 49(4), 327-355. https://doi.org/10.1002/(SICI)1097-4571(19980401)49:4<327:AID-ASI4>3.0.CO;2-4

Wilbur, W. J., Rzhetsky, A., & Shatkay, H. (2006). New directions in biomedical text annotation: Definitions, guidelines and corpus construction. BMC Bioinformatics, 7(1), 356. https://doi.org/10.1186/1471-2105-7-356

Zehr, S. C. (1999). Scientists' representations of uncertainty. Communicating Uncertainty: Media Coverage of New and Controversial Science, 3-21.

Zerva, C., Batista-Navarro, R., Day, P., & Ananiadou, S. (2017). Using uncertainty to link and rank evidence from biomedical literature for model curation. Bioinformatics, 33(23), 3784-3792. https://doi.org/10.1093/bioinformatics/btx466

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic-coupling analysis. Journal of the American Society for Information Science and Technology, 59(13), 2070-2086. https://doi.org/10.1002/asi.20910

Zhao, L. M., & Zhang, Q. P. (2011). Mapping knowledge domains of Chinese digital library research output, 1994-2010. Scientometrics, 89(1), 51-87. http://dx.doi.org/10.1007/s11192-011-0428-4