생의학 학술 문헌의 불확실성 기반 지식 동향 분석에 관한 연구*

Knowledge Trend Analysis of Uncertainty in Biomedical Scientific Literature

허고은 (Go Eun Heo)**
송 민 (Min Song)***

초 록

불확실성이란 정보의 합의나 현존하는 지식 부족으로 인해 명제의 지식이 불완전한 상태를 의미한다. 과학적 지식의 불확실성을 연구하는 학술문헌의 양은 시간이 흐름에 따라 기하급수적으로 증가하고 있으며, 이에 따라 새로운 지식이 발견되고 연구가 발전하고 있다. 이처럼 시간의 흐름은 지식의 불확실성의 패턴을 발견하는데 중요한 요인이 될 수 있음에도 불구하고 기존의 연구들은 불확실성 단어의 단순 출현 빈도를 기반으로 특정 학문 영역에서 불확실성의 특성을 파악해왔다. 따라서, 본 연구에서는 구축한 불확실성 단어를 생의학 영역의 불확실성 연구에 적용하여 시간의 흐름에 따른 불확실성의 변화와 패턴을 파악하고자 한다. 시간의 흐름에 따른 생의학 지식의 패턴을 분석하기 위해 대표 개체 페어, 동사 유형, 대표 개체의 패턴을 살펴보았으며 선형 회귀 분석을 통해 유의성 검증을 수행했다. 개체 페어 분석에서는 17건 중 7건의 개체 페어가 유의하게 감소하는 패턴을 보였다. 10개의 대표적인 동사 유형은 모두 시간이 흐름에 따라 유의하게 감소했다. 대표 개체의 연도별 상대적 중요도 분석에서는 유의하게 상승과 하강 패턴을 보이는 개체들의 불확실성 증감을 분석했다.

ABSTRACT

Uncertainty means incomplete stages of knowledge of propositions due to the lack of consensus of information and existing knowledge. As the amount of academic literature increases exponentially over time, new knowledge is discovered as research develops. Although the flow of time may be an important factor to identify patterns of uncertainty in scientific knowledge, existing studies have only identified the nature of uncertainty based on the frequency in a particular discipline, and they did not take into consideration of the flow of time. Therefore, in this study, we identify and analyze the uncertainty words that indicate uncertainty in the scientific literature and investigate the stream of knowledge. We examine the pattern of biomedical knowledge such as representative entity pairs, predicate types, and entities over time. We also perform the significance testing using linear regression analysis. Seven pairs out of 17 entity pairs show the significant decrease pattern statistically and all 10 representative predicates decrease significantly over time. We analyze the relative importance of representative entities by year and identify entities that display a significant rising and falling pattern.

키워드: 텍스트 마이닝, 불확실성, 의미적 술어, 버스티니스, 동향 분석
text mining, uncertainty, semantic predication, burstiness, trend analysis

---

* 본 연구는 박사학위논문을 수정・요약한 것임.
** 연세대학교 문헌정보학과 연구교수(goeun.heo@yonsei.ac.kr) (제1저자)
*** 연세대학교 문헌정보학과 교수(min.song@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2019년 5월 20일 ■ 최초심사일자 : 2019년 6월 19일 ■ 게재확정일자 : 2019년 6월 24일
■ 정보관리학회지, 36(2), 175-199, 2019. [http://dx.doi.org/10.3743/KOSIM.2019.36.2.175]

1. 서 론

불확실성(uncertainty)이란 정보의 합의나 현존하는 지식 부족으로 인해 명제의 지식이 불완전한 상태를 의미한다. 컴퓨터 과학 영역에서의 불확실성은 불완전한 관찰(observability)이나 비결정성(non-determinism), 또는 이들 모두로 인해 발생한다고 보았다(Russell, Norvig, & Intelligence, 1995). 언어학적 이론에서 불확실성은 불확실성의 양태(modality)의 개념과 관련되어 있다. 인식 양태(epistemic modality)는 화자가 전하고자 하는 내용인 명제(proposition)에 대해 얼마나 많은 확실성과 증거를 가지고 있는지, 지식의 믿음, 신뢰성의 정도, 평가 또는 판단과 관련되어 있다(Palmer, 2014). 과학적 불확실성은 명제가 현재 상태로는 참도 거짓도 아닌 불확실한 것을 의미하며 이러한 지식의 불확실성을 “Epistemic uncertainty”로 표현할 수 있다. 만약 현 세계의 지식에 기초하여 현재 참과 거짓을 결정할 수 없다면 지식적으로 불확실한 상태인 것이다(Szarvas, Vincze, Farkas, Móra, & Gurevychet, 2012).

연구자는 연구 질문을 형성하고 연구방법을 선정하는 것부터 발견점을 해석하고 다른 연구자들과 커뮤니케이션하는 모든 연구의 단계에서 불확실성을 다룬다(Cordner & Brown, 2013). 특히, 학술 커뮤니티에서 관련 연구자들 간의 합의를 이루어가는 과정에서 이전의 발견점이나 학술적 명제가 사실이 되기 위해 논쟁이 필요하다(Bourdieu, 1975; Shwed & Bearman, 2010). 연구자들은 동일한 과학적 질문에 대해 다양하게 데이터를 생성하고 이러한 데이터들은 연구자들 사이에 동의를 얻을 수도 있고, 얻지 못할 수도 있다. 연구자들 간의 의견 불일치 정도가 높을수록 혼란과 논란이 발생한다(Ioannidis & Trikalinos, 2005). 즉, 연구자는 불일치, 모순, 상충되는 발견점이 나타나거나 긴급한 위기를 해결하기 위해 대립되는 패러다임이 제안될 때 심화된 불확실성에 직면한다(Kuhn, 1970). 이러한 논쟁과 불일치 속에서 특정 패러다임이 설립되거나 합의가 형성되면 학술 커뮤니티의 연구 영역에서 불확실성이 감소하게 된다. 이처럼 과학적 지식을 얻기 위해 불확실성은 반드시 거쳐가야 하는 필수적인 단계로 인식되며 모순과 상충의 혁명적인 변화로 발생하는 과학적 지식을 이해하기 위해 지식의 상태(epistemic status)의 흐름을 파악할 필요가 있다.

기존의 불확실성 연구들(Friedman, Alderson, Austin, Cimino, & Johnson, 1994; Hyland, 1998; Falahati, 2006; Rizomilioti, 2006; Vold, 2006)은 특정 학문 영역에서 불확실성 단어의 단순 출현 빈도 기반으로 불확실성의 특성을 파악해왔다. 학술적 지식은 연도의 흐름에 따라 새로운 지식이 발견되고 논의되므로 시간의 흐름에 따른 불확실성의 패턴 변화를 살펴볼 필요가 있다. 따라서 본 연구에서는 텍스트 마이닝 기반으로 생의학 문헌을 처리하여 생의학 문장 내에서 시간의 흐름에 따른 불확실성 단어 기반 생의학 지식의 특성을 분석하고자 한다. 이를 위해 불확실성 단어가 포함된 문장으로 데이터 집합을 구성하였고 생의학 지식으로 대변되는 개체와 이들 간의 관계성을 연도별로 동향을 파악하였다. 본 연구 목적에 따라 아래와 같이 세부적인 연구 수행 방법을 구성하였다.

- 전체 데이터 집합 중 불확실성 단어 데이터 집합의 출현 빈도와 비율 기반 특성을 발견하여 대표적인 생의학 지식을 선정한다.
- 대표적인 생의학 개체 페어와 동사 유형(predicate)의 출현 빈도와 비율 기반 패턴을 발견하고 통계분석을 수행한다.
- 대표적인 생의학 개체의 중요성을 파악하는 방법인 버스티니스(Burstiness)를 적용하여 패턴을 발견하고 통계분석을 수행한다.

불확실성 단어가 포함된 문장의 생의학적 지식의 특성을 종합적으로 분석하기 위해 PubMed에서 196개 불확실성 단어(Chen, Song, & Heo, 2018)들을 개별적으로 수집하였고 수집연도는 연도별 특성을 분석하기에 적합한 1990년부터 2016년까지 한정했다. XML형식으로 다운받은 데이터는 대량의 데이터 처리에 적합한 SAX (Simple API for XML) XML 파서를 통해 PMID, Year, Title, Abstract 정보를 추출했다. 다음으로 PubMed의 학술 문헌 데이터에 대한 UMLS(Unified Medical Language System)의 의미적 술어(semantic predication)를 보유하고 있는 SemMedDB(Semantic Medline Database)를 기반으로 관련 정보를 추출했다. 의미적 술어를 가진 문장들 중 부정(negation)의 의미를 가진 문장을 확인하고 제외하기 위해 규칙 기반의 NegEx(Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001) 알고리즘의 응용 버전인 GenNegEx(Solti, Cooke, Xia, & Wurfel, 2009)를 적용하여 최종 데이터 집합을 구성했다. SemMedDB 결과에서 생의학 지식인 개체, 개체 유형, 동사 유형, 의미적 술어, 개체 페어 간의 출현 빈도와 비율을 확인했다. 더불어 불확실성 단어를 포함한 데이터 집합 기반 대표적인 개체 페어와 동사 유형이 시간의 흐름에 따라 어떠한 패턴을 보이는지 분석했다. 또한 대표적인 개체들 간의 상대적 중요도를 의미하는 버스티니스(Madsen, Kauchak, & Elkan, 2005) 값을 기반으로 연도에 따른 불확실성 변화를 분석했다. 대표적인 개체 페어, 동사 유형, 개체의 연도별 불확실성 증감의 차이는 선형 회귀 분석을 통해 회귀식을 도출하였고 통계적 유의성을 검증하였다.

2. 이론적 배경

Hedging은 인식 양태의 한 부분으로 일반적으로 문장 내에서 조동사, 형용사, 부사, 어휘적 동사로 표현된다. 예를 들면, perhaps, I guess, may be, quite, sort of 등이 될 수 있다. Hedging은 학술 연구의 글쓰기에서 연구자가 새로운 지식을 글로 표현할 때 주장을 올바르게 전달하는 수단으로 사용되며 독자들이 글을 이해하고, 평가하고, 명제적 정보에 반응할 수 있도록 돕는 수사적 장치(rhetorical device)로서 중대한 의미를 지닌다(Hyland, 1998).

Hedging과 불확실성 관련 연구는 크게 언어학 관점과 NLP 관점에서의 연구가 주로 행해졌으며, NLP 연구에서는 대부분 생의학 텍스트 또는 의학 영역의 데이터를 사용하였다. 본 연구의 범위에 따라 NLP 관점에서 컴퓨터 언어학 영역에 초점을 맞춘 연구들을 살펴보고자 한다.

2.1 Hedging과 불확실성 주석 지침 개발 연구

학술문헌 내의 Hedging과 불확실성 단어와 관련하여 정보의 범위와 유형에 대해 지침을 개발하고 주석을 수행하여 코퍼스를 구축한 연구들이 있다. Wilbur, Rzhetsky, Shatkay(2006)는 생의학 영역의 텍스트 마이닝 기법에 도움을 줄 수 있는 주석 지침을 개발했다. 학술 문헌의 문장을 5가지의 질적 차원인 초점(focus), 극성(polarity), 확실성(certainty), 증거(evidence), 방향성(directionality)으로 구분했다. Hedging과 관련된 확실성과 증거는 0부터 3까지의 총 4개의 단계로 강도를 구분했다. 12명의 평가자가 독립적으로 101개의 문장에 대해 주석을 수행했고, 70~80%의 일치율을 보였다. Thompson, Nawaz, McNaught, Ananiadou(2011)는 생의학 이벤트 코퍼스의 메타 지식(meta-knowledge)을 강화하기 위해 주석 스키마를 구축했다. 각 이벤트는 5개의 다양한 메타 지식인 지식 유형(knowledge type), 확실성 수준(certainty level), 극성(polarity), 매너(manner), 정보원(source)으로 주석을 처리했다. 이 중 확실성 수준은 may, might, perhaps를 포함하거나 sometimes, rarely, scarcely 등의 불확실한 단어를 포함하는 1단계와, likely, probably, suggest, indicate 또는 normally, often, frequently와 같은 단어가 출현하는 약한 추측의 2단계, 이벤트가 확실한 3단계로 구분했다. 두 명의 평가자에 의해 스키마가 구축되었고, 0.84~0.93의 kappa 값으로 평가자 간의 일치율을 보였다.

Hedging의 한 측면인 부정 표현 발견 연구의 일환으로 Chapman, Bridewell, Hanbury, Cooper, Buchanan(2001)은 문장 내에서 부정을 의미하는 구를 발견하고 부정 범위에 포함된 의학 용어들을 확인할 수 있는 NegEx 정규 표현 알고리즘을 개발했다. Szarvas, Vincze, Farkas, Csirik(2008)과 Vincze, Szarvas, Farkas, Móra, Csirik(2008)은 생의학 텍스트의 부정과 불확실성을 다루는 연구에 활용될 수 있는 자원을 제공하기 위해 BioScope 코퍼스를 구축했다. 부정(negative)과 추측성(speculative) 키워드의 토큰 레벨과 언어학적 범위의 문장 레벨의 주석도 포함했다. 대표 평가자가 지침을 설정하였고 두 명의 언어학 전문가에 의해 주석이 수행되었다.

기존의 특정 유형이나 현상에 국한되어 개발된 방법론의 한계점을 극복하여 Szarvas et al.(2012)은 다양한 영역과 장르에 적용이 가능한 불확실성 단어 발견 모델을 제안했다. 의미적 불확실성(semantic uncertainty)을 지식(epistemic)과 가설(hypothetical)의 두 카테고리로 분류했다. 가설 불확실성은 역설적인(paradoxical)것과 지식이 아닌 양상(non-epistemic modality)으로 구분된다. 역설적인 불확실성은 하위 카테고리로 조사(investigation)와 조건(condition)을 포함한다. 지식이 아닌 양상에서는 의견(doxastic)과 다이나믹 불확실성이 있다. 장르와 도메인 간의 불확실성 단어를 인식하기 위해 세 가지 코퍼스의 주석을 정규화 했으며, CoNLL(Computational Natural Language Learning) 2010 학회에서 수행되었던 자질집합을 이용하여 토큰 기반의 분류와 시퀀스 레이블(Conditional Random Fileds, CRF) 모델로 불확실성 단어 발견의 결과를 비교했다. 주석의 비용을 최소화하면서 성능을 높이기 위한 방법으로는 이러한

두 가지의 접근법을 통합할 필요가 있다고 제안했다.

Vincze(2013)는 담화(discourse-)와 화용론(pragmatics-) 관련 불확실성을 살펴보았다. 언어학적 현상을 분류하였고, 위키피디아 데이터를 기반으로 코퍼스를 구축했다. Szarvas et al.(2012)과 유사하게 불확실성 단어를 사용했으며, 두 명의 언어학 평가자가 4,530건의 위키피디아 데이터를 담화 단계의 불확실성 유형인 weasel, peacock, hedge로 구분하여 태깅했다.

Chen, Song, Heo(2018)는 초기 불확실성 단어 리스트를 정의하고 두 층의 인공 신경망으로 이루어진 텍스트 처리기법인 Word2Vec(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)을 적용하여 연관 단어를 자동적으로 추출했다. 추출된 단어들은 두 명의 평가자에 의해 실제 불확실성 단어를 분류하여 최종적으로 196개의 단어를 구축하였다. 기존 연구들과 달리 학문 영역에 제한되지 않는 일반성을 갖춘 단어로 구성하였다.

이처럼 대부분의 연구들은 수작업 기반으로 주석을 수행함으로써 불확실성 단어와 hedging 단어를 발견하고 불확실성을 의미하는 범위를 지정하여 코퍼스를 구축했다. 또한 구축한 코퍼스를 대상으로 규칙 기반 또는 기계 학습 기반의 접근법을 제안하여 성능평가를 수행했다.

2.2 학문의 Hedging과 불확실성 패턴 연구

특정 학문 영역에서의 불확실성 단어 분포에 대한 연구나 학문 영역 간 단어의 출현 빈도 기반의 특성을 비교한 연구들이 수행되었다. Rizomilioti(2006)는 다른 학문 영역의 학술 문헌에서 불확실한 표현의 빈도를 비교하기 위해 고고학, 문학 비평, 생물학에서 인식 양태를 표현하는 언어학적 패턴을 분석하였다. 가장 많은 불확실성 단어를 포함한 학문은 고고학인 반면, 문학 비평 논문은 가장 낮은 빈도를 보였다. Falahati(2006)는 의학, 화학, 심리학 학술 문헌의 hedging 분포를 확인하였다. 가장 많은 hedge를 포함한 학문은 심리학 학술 문헌이며 나머지 두 학문은 hedge 사용 빈도가 낮았다. Vold(2006)는 영어, 프랑스어, 노르웨이어 세 가지 언어와 언어학과 의학의 두 다른 학문 영역에서 hedge의 사용 빈도를 비교했다. hedge를 많이 사용한 언어는 영어와 노르웨이어였다. 이들은 유사한 패턴을 보였는데 모두 게르만어파이며, 유사한 언어 전략을 공통적으로 가지고 있기 때문인 것으로 확인했다. 또한 프랑스어로 작성한 저자들에 비해 텍스트가 더 비평적이고 뚜렷한 표현을 나타낸다는 점을 발견했다. 이러한 연구 결과 다양한 기술 또는 학문 영역에서 불확실성을 표현하거나 불확실성 단어를 사용하는 방식이 상이하다는 점을 확인했다. 최근 연구로 Heo(2019)는 생의학 학술 문헌을 대상으로 UMLS에서 제공하는 의미적 술어를 기반으로 생의학 명제를 분석하고, DMR 토픽 모델링을 수행하여 생의학 개체의 시계열적 패턴을 파악하였다. 시간이 흐름에 따라 과학적 지식의 표현이 불확실성이 감소하는 패턴으로 발전하고 있다는 점을 확인하였다.

학술 연구의 출판물은 시간적 정보를 제공하는 연도 정보가 중요한 역할을 함에도 불구하고 기존 불확실성 연구에서는 특정 학문 영역이나, 학문 영역 간에 존재하는 불확실성 단어

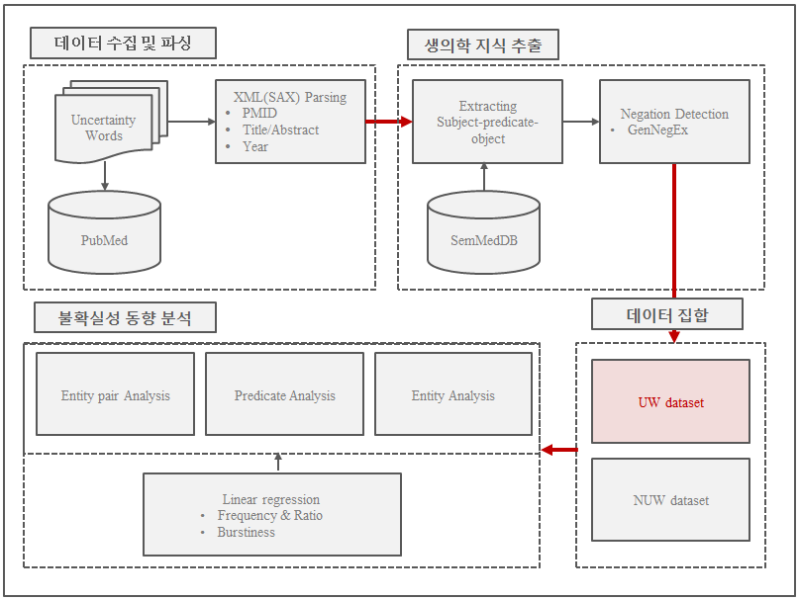
의 단순 출현 빈도로 특성을 분석하였다(Heo, 2019). 학술적 지식은 연도의 흐름에 따라 새로운 지식이 발견되고, 분석되며 논의되기에 이러한 불확실성의 패턴을 시계열적으로 파악할 필요가 있다. 따라서 본 연구에서는 불확실성 단어를 포함한 데이터 집합에서 생의학 지식의 시계열 분석을 통해 기존 연구에서 제안되지 않았던 방식으로 과학적 지식의 발전 행태를 종합적으로 파악하고자 한다.

3. 연구 설계

본 연구의 목적인 불확실성 단어를 생의학 영역의 학술문헌에 적용하여 시간의 흐름에 따른 불확실성 변화와 패턴을 파악하기 위해 〈그림 1〉과 같이 연구를 설계하였다. 각 처리 과정은 다음 절부터 상세히 기술한다.

3.1 데이터 수집

Chen, Song, Heo(2018)는 초기 불확실성 단어 집합 61개를 기반으로 Word2Vec(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)을 적용하여 불확실성 단어를 자동적으로 확장했다. 또한 393개의 자동으로 추출한 후보 단어 중 두 명의 평가자에 의해 불확실성 단어 여부를 파악하는 후처리 과정을 거쳐 최종적으로 196개의 불확실성 단어를 구축하였다. 확장된 불확실성 단어는 다양한 기계학습 알고리즘을 통해 분류 평가를 수행하였으며 F1척도 기반 0.62~0.76의 결과를 보였다. 따라서 본 연구에서는 Chen, Song, Heo(2018)의 196개 단어를 대상으로 생의학 학술 문헌을 포함하고 있는



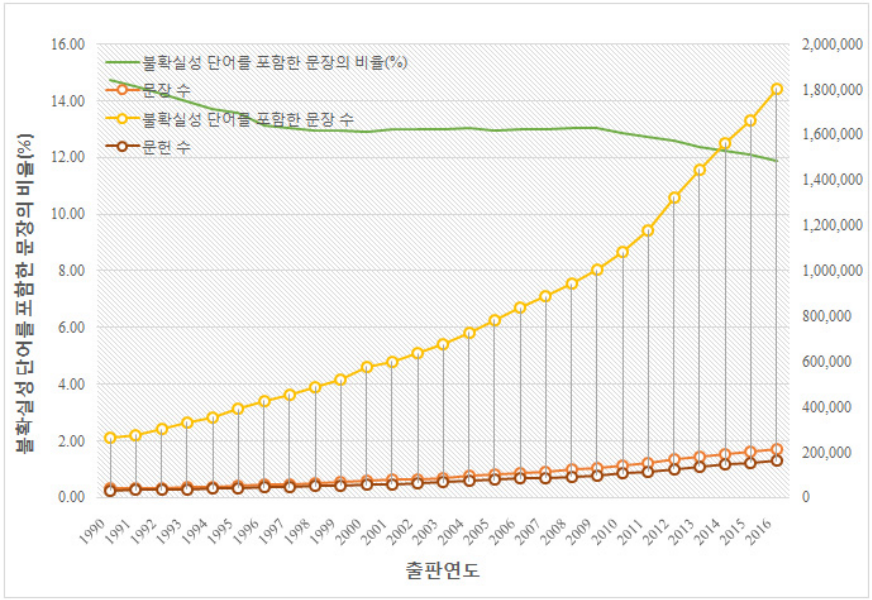
〈그림 1〉 연구 개요

검색엔진인 PubMed에서 각 단어를 질의어로 제목과 초록에 해당 단어가 포함되어 있는 문헌 데이터를 수집했다. 연도별 분석을 수행하기 위해 1990년도부터 2016년까지 총 27년으로 출판 연도를 설정하여 수집하였다. 총 문헌 데이터는 2,489,466건으로 구성되었다. 앞선 선행연구에서 기술한 Heo(2019)의 연구에서도 이와 동일한 데이터 집합을 사용하였다.

3.2 데이터 파싱

PubMed에서 수집한 문헌 데이터는 XML 형식으로 구성되어 있다. XML이란 구조화된 문서를 웹상에서 구현할 수 있도록 텍스트를 정의하는 어휘와 기호로 구성되는 마크업 언어(markup language)로 1996년 W3C(World Wide Web Consortium)에서 제안한 웹문서 표준형식이다. XML 데이터 파일은 선언, 루트 요소(root elements), 속성(attributes), 태그 및 데이터 값(values)으로 구분된다. 총 45개의 요소(elements) 중 본 연구에서 필요한 네 가지 요소인 〈PMID〉, 〈PubDate〉, 〈ArticleTitle〉, 〈Abstract〉와 〈AbstractText〉를 SAX 파서를 적용하여 태그명, 속성명, 속성값 및 요소 내용을 추출했다.

데이터 통계는 〈그림 2〉와 〈표 1〉과 같다. 문헌 수는 총 2,110,181건이며 총 문장 수는 21,541,600건이다. 이 중 불확실성 단어를 포함하는 문장은 총 2,755,482건으로 평균적으로 12.79%의 비율을 차지했다. 또한, 불확실성 단어를 포함하지 않는 문장은 18,786,118건으로 87.21%이다. 불확실성 단어를 포함하고 있는 문장의 6.82배이다.



〈그림 2〉 불확실성 단어 기반 데이터 집합의 통계 그래프

〈표 1〉 불확실성 단어 기반 데이터 집합의 통계

출판연도	문헌 수	문장 수	불확실성 단어 포함 문장 수	문헌 당 문장 수	불확실성 단어 포함 문장 비율
1990	31,208	262,163	38,642	8.4	14.74
1991	32,486	277,387	40,241	8.54	14.51
1992	34,385	301,277	42,944	8.76	14.25
1993	36,877	332,169	46,462	9.01	13.99
1994	38,672	355,027	48,648	9.18	13.7
1995	41,690	389,466	52,922	9.34	13.59
1996	43,977	425,996	55,970	9.69	13.14
1997	46,224	455,775	59,498	9.86	13.05
1998	49,322	488,546	63,366	9.91	12.97
1999	52,064	518,353	67,050	9.96	12.94
2000	57,507	577,884	74,456	10.05	12.88
2001	59,910	601,014	78,106	10.03	13
2002	63,345	634,636	82,375	10.02	12.98
2003	67,150	675,068	87,802	10.05	13.01
2004	72,337	726,659	94,752	10.05	13.04
2005	77,735	783,505	101,611	10.08	12.97
2006	83,047	837,386	108,823	10.08	13
2007	87,569	886,404	115,052	10.12	12.98
2008	93,522	942,076	122,766	10.07	13.03
2009	99,243	1,003,748	130,684	10.11	13.02
2010	105,808	1,082,026	139,305	10.23	12.87
2011	114,075	1,178,789	150,282	10.33	12.75
2012	126,376	1,323,379	166,894	10.47	12.61
2013	136,001	1,448,048	178,931	10.65	12.36
2014	145,004	1,565,389	191,542	10.8	12.24
2015	152,589	1,665,577	201,910	10.92	12.12
2016	162,058	1,803,853	214,448	11.13	11.89
총계(평균)	2,110,181	21,541,600	2,755,482	(9.92)	(13.10)

3.3 생의학 지식 추출

생의학 지식인 개체와 개체 간의 관계성을 추출하기 위해 SemMedDB를 적용하였다. SemMedDB는 PubMed의 제목과 초록으로부터 의미적 술어를 추출하여 제공하는 의미해석 프로그램인 SemRep(Rindflesch & Fiszman, 2003)의 데이터베이스이다. 이는 미국 국립의학도서관(National Library of Medicine, NLM)에서 제공하는 생의학 분야의 온톨로지인 UMLS의 지식 정보원(Bodenreider, 2004)을 기반으로 한다. 의미적 술어는 트리플 구조인 주어-서술어-목적어 형태로 구성되어 있으며 생의학 영역에서 유의미한 지식을 발견하는데 도움을 주는 지식 자원으로 알려져 있다.

최신 버전인 SemMedVER30에서 포함하고

있는 총 5개의 테이블(CITATIONS, GENERIC_CONCEPT, PREDICATION, PREDICATION_AUX, SENTENCE) 중 PREDICATION 테이블을 참조하였으며 이는 총 12개의 필드로 구성되어 있다. PMID를 기반으로 데이터 집합에 포함된 총 11,520,923건의 생의학 지식들을 추출했다. 트리플 구조를 가지는 고유 문헌 수는 총 1,768,757건이며, 고유 문장 수는 7,023,380이다. 본 연구에서 필요한 7가지 정보인 주어명(SUBJECT_NAME), 주어의 의미 유형(SUBJECT_SEMTYPE), 동사 유형(PREDICATE), 목적어명(OBJECT_NAME), 목적어의 의미 유형(OBJECT_SEMTYPE), 문헌 ID(PMID), 문장 ID(SENTENCE_ID)의 정보를 추출했다.

〈표 2〉는 불확실성 단어 'consensus'에 대한 생의학 지식 추출 결과로 PMID 8425050 문헌에 총 9건의 트리플 구조가 존재한다. 문장 ID를 확인해보면 총 5건의 문장에 대한 결과인 점을 확인할 수 있다.

또한 SENTENCE 테이블의 문장을 문장 ID를 기반으로 추출하였고 이들을 대상으로 문장 내에서 196개의 불확실성 단어가 포함되어 있는지 확인했다. 결과적으로 불확실성 단어를 포함하는 문장은 749,120건이며 고유 문헌 수는 643,103건이다. 이에 대한 의미적 술어는 총 1,242,704건으로 전체 의미적 술어 중 10.79%를 차지했다.

3.4 부정 표현 발견

불확실성 단어 데이터 집합을 구분하기에 앞서 불확실성 단어가 포함된 문장에 불확실성 단어를 기준으로 부정 표현이 존재하는지 여부를 확인했다. 문장의 문맥을 파악하여 실제 문장의 의미가 지식의 불확실한 상태를 의미하지 않는 경우 분석 대상에서 제외하여 정교한 데이터 집합을 구성하기 위한 필수적인 단계로 판단하였다. 임상 텍스트로부터 부정 표현을 발견하기 위한 규칙 기반의 NegEx(Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001) 알고리즘에서 확장된 알고리즘인 부정 표현의 상태와 문장 내 문맥 정보를 함께 표현하는 Solti, Cooke, Xia, Wurfel(2009)의 GenNegEx를 기반으로 부정 표현 발견을 수행했다. GenNegEx v1.2의 성능 평가 결과 정확률 93%, 재현율 95%, F척도 94%의 높은 성능을 보인 알고리즘이다.

〈표 2〉 SemMed DB를 이용한 생의학 지식 추출 결과

주어	주어 의미 유형	동사 유형	목적어	목적어 의미 유형	문헌 ID	문장 ID
Polysiphonia boldii	alga	ISA	Algae, Red	alga	8425050	35598845
Polysiphonia boldii	alga	LOCATION_OF	Base Sequence	nusq	8425050	35598845
Polysiphonia boldii	alga	ISA	Algae, Red	alga	8425050	35599000
Polysiphonia boldii	alga	LOCATION_OF	Phycoerythrin	aapp	8425050	35599000
Intergenic Region	bacs	PART_OF	Phycoerythrin	aapp	8425050	35599252
Porphyridium cruentum	alga	ISA	Algae, Red	alga	8425050	35599754
Porphyridium cruentum	alga	LOCATION_OF	Phycoerythrin	aapp	8425050	35599754
Pes	bpoc	LOCATION_OF	Consensus Sequence	nusq	8425050	35599882
Pes	bpoc	LOCATION_OF	Aspartate	aapp	8425050	35599882

〈표 3〉 불확실성 단어 데이터 집합의 부정 표현 문장

문장 ID	불확실성 단어	문장	부정 표현 발견 구
136504191	uncommon	CONCLUSIONS: In children with pacemakers implanted forAVB, NSVT is not uncommon and may be associated with increased mortality.	not uncommon and may be associated with increased
33917522	atypical	We report that haloperidol andfluphenazine, classical neuroleptics, cause a generalized reduction in theactivity of NADH: ubiquinone oxidoreductase (complex I) in the rat brain in vivo, an effect that was not observed with the atypical neuroleptic, clozapine.	not observed with the atypical neuroleptic

NegEx 알고리즘에서 정의한 트리거 단어를 이용하여 총 749,120건의 불확실성 단어가 포함된 문장의 부정 표현을 확인한 결과 5,441건의 문장에서 부정 표현이 발견되었다. 이는 문장 내에 속한 불확실성 단어를 기준으로 불확실성 단어를 부정하는 의미를 지니므로 불확실성 단어를 포함한 문장 집합에서 제외하였다. 〈표 3〉은 부정 표현 문장으로 인식된 두 문장의 예시를 나타낸다.

결과적으로 최종 분석 데이터 집합의 문헌 수는 총 1,768,757건이다. 이 중 불확실성 단어 문장에 해당하는 문헌 수는 640,232건이며, 불확실성 단어를 포함하지 않는 문장에 해당하는 문헌 수는 1,701,959건이다. 전체 문장 수는 총 7,023,380건이며 불확실성 단어 문장 수는 743,679건으로 전체 문장의 10.59%를 차지했다. 반면 불확실성 단어를 포함하지 않는 문장은 6,279,701건으로 비율은 89.41%이다. 또한 각 문장에서 추출된 의미적 술어의 수는 불확실성 단어 데이터 집합이 1,233,683건으로 10.71%, 불확실성 단어를 포함하지 않는 데이터 집합이 10,287,240건으로 89.29%를 차지했다.

4. 실험 결과 분석

불확실성 단어 기반 생의학 지식의 패턴을 연도별 흐름에 따라 확인하기 위해 분석 대상이 되는 대표적인 개체 페어를 객관적으로 선정하였고, 출현 비율을 기반으로 대표 개체 페어와 대표 동사 유형을 통계적으로 분석하였다. 또한 개체의 상대적 중요도를 의미하는 버스티니스 값을 기반으로 통계 분석을 수행하여 연도별 개체의 상대적 중요도가 어떻게 변화하는지 살펴보고자 했다.

4.1 연도별 대표 개체 페어 선정

시간의 흐름에 따른 개체 페어의 동향을 비교분석하기 위해 각 데이터에서 개체 페어의 출현 빈도 비율이 0.05%이상인 불확실성 단어 비포함 데이터 집합의 46개 페어와 불확실성 단어 데이터 집합의 59개 페어 중 두 데이터 집합에 중복으로 출현한 34개 페어를 대표 페어로 선정했다. 그 후 생의학적으로 유의미한 페어를 추출하기 위해 전문가의 검증을 거쳐 페어를 필터링 했다. 필터링 과정은 첫째로, 해당

개체의 의미 유형이 의학 용어를 의미하지 않거나 일반적인 단어들을 포함할 경우 제외한다. (e.g. Genetic Function, Health Care Activity, Injury or Poisoning, Mammal, Quantitative Concept, Research Activity, and Research Device, etc.) 둘째로, 개체의 의미 유형이 의학 용어를 의미하지만 해당 개체가 광범위한 단어일 경우 제외한다(e.g. Pharmacotherapy, Pharmaceutical Preparations, Disease, Operative Surgical Procedures, Critical Illness, Excision, Lesion, Complication, Human, Patients, Therapeutic procedure, etc.). 이 때, 두 개체 페어를 함께 통합하여 의미가 있을 경우 전문가의 판단에 의해 제외하지 않았다. 최종적으로 17개의 대표 페어를 선정하였으며 각 데이터 집합 내에서의 각 개체 페어의 순위와 빈도수는 〈표 4〉와 같다. 이는 불확실성 단어 데이터 집합을 기준으로 정렬한 결과이다.

Antibiotics-Therapeutic procedure 페어를 제외한 16개의 페어가 모두 Patients와 페어를 형성했다. 두 데이터 집합은 전반적으로 유사한 순위를 보이고 있으나 몇 가지 페어는 상이한 순위를 보였다. 대표적으로 Coronary Arteriosclerosis-Patients 페어는 불확실성 단어 데이터 집합에서 5순위인데 반해 불확실성 단어 비포함 데이터 집합에서는 15순위였다. 불확실성 단어 비포함 데이터 집합 대비 불확실성 단어 데이터 집합에서 순위가 높게 나타났으므로 두 개체 관계에 대한 과학적 논의가 상대적으로 불확실한 단계로 남아있다는 점을 유추할 수 있다.

〈표 4〉 17개의 대표적인 개체 페어

번호	개체 페어	UW datasets		NUW datasets		All datasets	
		순위	빈도수	순위	빈도수	순위	빈도수
1	Coronary Arteriosclerosis-Patients	5	2,284	15	7,810	13	10,094
2	Malignant Neoplasms-Patients	6	1,991	5	11,686	5	13,677
3	Patients-Schizophrenia	9	1,382	12	8,982	11	10,364
4	Neoplasm-Patients	10	1,278	9	9,641	9	10,919
5	Cerebrovascular accident-Patients	11	1,268	13	8,118	14	9,386
6	Diabetic-Patients	14	1,115	8	9,845	8	10,960
7	Malignant neoplasm of breast-Patients	16	911	10	9,415	12	10,326
8	Heart failure-Patients	18	804	14	7,955	15	8,759
9	Parkinson Disease-Patients	21	784	16	7,658	16	8,442
10	Diabetes-Patients	22	783	28	5,208	29	5,991
11	Antibiotics-Therapeutic procedure	25	722	30	5,028	30	5,750
12	Myocardial Infarction-Patients	26	693	33	4,780	33	5,473
13	Alzheimer's Disease-Patients	27	647	23	5,811	25	6,458
14	Patients-Rheumatoid Arthritis	28	616	19	7,200	19	7,816
15	Kidney Failure, Chronic-Patients	30	599	22	5,942	24	6,541
16	Obesity-Patients	32	573	24	5,742	27	6,315
17	Multiple Sclerosis-Patients	33	572	31	5,012	31	5,584

4.2 연도별 대표 개체 페어 통계 분석

연도별 분석을 수행하기 위해 17개 페어의 1990~2016년까지의(총 27년) 연도별 출현 횟수를 산출했다. 전체 데이터 집합에서 각 페어의 연도별 분포 변화는 최신 연도로 갈수록 지속적으로 증가하는 패턴을 보였으며, Malignant Neoplasms-Patients 페어가 2016년에 총 1,116회 출현하여 가장 높은 출현 빈도를 나타냈다.

빈도 데이터는 전체 데이터 집합 중 불확실성 단어 데이터 집합의 비율로 환산하여 불확실성 단어의 증감을 확인했다. 불확실성 단어 데이터 집합에서의 17개 개체 페어의 연도별 출현 비율에 대한 평균, 최댓값, 최솟값, 범위(최댓값-최솟값), 표준편차는 〈표 5〉와 같다.

출판연도에 따라 불확실성 단어 분포의 비율이 변화하는지를 파악하기 위해 연구가설을 다음과 같이 설정했다.

H_A = 시간이 흐름에 따라 불확실성 단어 데이터 집합의 개체 페어 출현 비율에 영향을 미칠 것이다.

선형 회귀 분석을 수행하였고 회귀식을 도출한 결과는 〈표 6〉과 같다. 회귀식은 y = ax + b의 선형 함수(linear function)로 표현된다. 두 회귀계수(coefficient of regression)중 a는 기울기(slope)이며, b는 절편(intercept)이다. 결정계수(coefficient of determination)는 선형 모델의 적합성을 판단하기 위한 척도로 추정된 회귀선이 관찰 값에 얼마나 적합한지를 특정한다. 결정계수는 R2(R-square)로 표시하며 1에 가까울수록 회귀직선이 자료에 적합하며 0에 가까울수록 회귀직선의 적합도가 낮다.

〈표 5〉 대표 개체 페어의 기술 통계

번호	개체 페어	평균	최댓값	최솟값	범위	표준편차
1	Coronary Arteriosclerosis-Patients	22.42	28.89	13.61	15.28	3.23
2	Malignant Neoplasms-Patients	15.3	25.27	11.59	13.69	2.81
3	Patients-Schizophrenia	14.56	32	8.77	23.23	5.17
4	Neoplasm-Patients	11.95	17.35	8.09	9.25	2.02
5	Cerebrovascular accident-Patients	13.68	18.18	8.82	9.36	2.46
6	Diabetic-Patients	10.33	14.93	7.17	7.75	1.8
7	Malignant neoplasm of breast-Patients	9.52	15.79	5.05	10.74	2.17
8	Heart failure-Patients	9.75	16.36	6.56	9.81	2.72
9	Parkinson Disease-Patients	11.11	20.83	6.99	13.84	3.76
10	Diabetes-Patients	13.51	23.08	6.12	16.95	3.64
11	Antibiotics-Therapeutic procedure	12.65	17.84	7.69	10.15	2.25
12	Myocardial Infarction-Patients	14.11	25.64	6.88	18.76	5.08
13	Alzheimer's Disease-Patients	11.27	20	6.56	13.44	3.76
14	Patients-Rheumatoid Arthritis	8.59	12.66	4.84	7.82	2.01
15	Kidney Failure, Chronic-Patients	8.85	15.09	4.44	10.65	2.54
16	Obesity-Patients	9.22	17.78	3.45	14.33	3.51
17	Multiple Sclerosis-Patients	12.15	29.79	5.49	24.3	5.32

〈표 6〉 17개 개체 페어의 선형 회귀 분석 결과

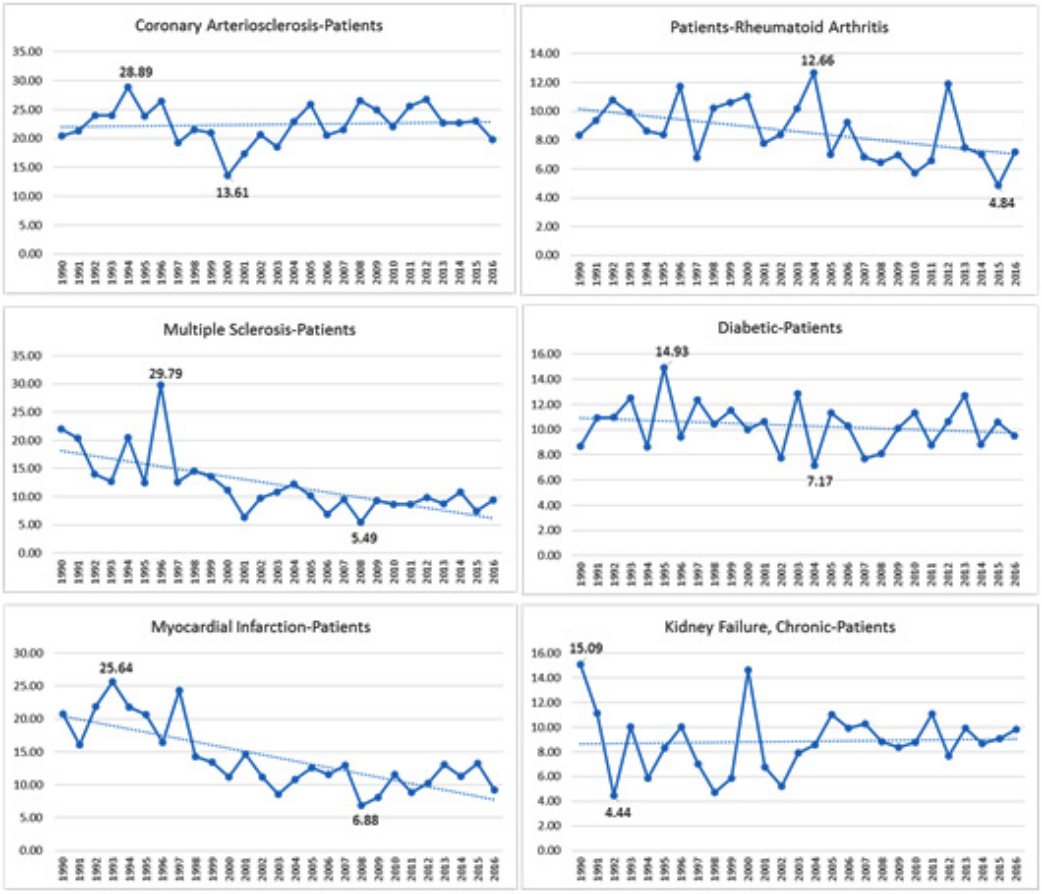
번호	개체 페어	회귀식	결정계수	유의확률
1	Coronary Arteriosclerosis-Patients	y = 0.0299x + 22.003	0.005	0.72
2	Malignant Neoplasms-Patients	y = -0.2149x + 18.309	0.354	0.001*
3	Patients-Schizophrenia	y = -0.2253x + 17.713	0.115	0.083
4	Neoplasm-Patients	y = -0.0912x + 13.227	0.124	0.072
5	Cerebrovascular accident-Patients	y = -0.0628x + 14.555	0.04	0.32
6	Diabetic-Patients	y = -0.0436x + 10.936	0.036	0.345
7	Malignant neoplasm of breast-Patients	y = -0.1225x + 11.234	0.193	0.022*
8	Heart failure-Patients	y = -0.1079x + 11.261	0.096	0.117
9	Parkinson Disease-Patients	y = -0.3375x + 15.834	0.489	4.99E-05**
10	Diabetes-Patients	y = -0.0743x + 14.547	0.025	0.427
11	Antibiotics-Therapeutic procedure	y = -0.041x + 13.225	0.02	0.48
12	Myocardial Infarction-Patients	y = -0.4881x + 20.946	0.561	6.95E-06**
13	Alzheimer's Disease-Patients	y = -0.3452x + 16.106	0.512	2.75E-05**
14	Patients-Rheumatoid Arthritis	y = -0.1208x + 10.283	0.218	0.014*
15	Kidney Failure, Chronic-Patients	y = 0.0162x + 8.6226	0.002	0.805
16	Obesity-Patients	y = -0.0353x + 9.7139	0.006	0.698
17	Multiple Sclerosis-Patients	y = -0.4592x + 18.575	0.452	1.22E-04**

$p^*<.05$, $p^{**}<.01$

17개의 대표적인 개체 페어 중 Coronary Arteriosclerosis-Patients와 Kidney Failure, Chronic-Patients 페어를 제외하고 총 15개 페어는 기울기가 음수(-)로 연도가 증가할 때 불확실성 단어 분포의 비율은 감소하는 추세를 나타냈다. 특히 기울기가 0.4881로 가장 높은 값을 가진 Myocardial Infarction-Patients 페어는 연도의 변화에 불확실성 단어 분포의 비율이 영향을 많이 받는다는 것을 의미한다.

가설 검정을 위해 산출한 유의확률(P-Value)은 95%의 신뢰도에서 세 가지 개체 페어인 Malignant Neoplasms-Patients, Malignant neoplasm of breast-Patients, Patients-Rheumatoid Arthritis의 연도별 변화가 유의하게 나타났으며, 99%의 신뢰도에서 네 가지 개체 페어인 Parkinson Disease-Patients, Myocardial Infarction-Patients, Alzheimer's Disease-Patients, Multiple Sclerosis-Patients가 유의하게 나타났다. 총 7개의 개체 페어가 연도의 변화에 따라 불확실성 단어 분포의 비율이 유의하게 감소하고 있음을 통계적으로 확인했다.

보다 상세한 분석을 수행하기 위해 17개 대표적인 개체 페어 중 서로 다른 특성을 지니는 대표적인 개체 페어를 선정했다. 개체 페어 선정의 기준은 1) 평균 기반, 2) 표준편차 기반, 3) 기울기 기반으로 최댓값, 최솟값을 가지는 총 6개 개체 페어를 선정했다. 연도별 비율 그래프는 〈그림 3〉과 같으며 각 그래프마다 최댓값과 최솟값을 데이터 레이블로 표현했다. 그래프의 변동(fluctuation) 중 중간 연도에 정점(peak)을 보이는 경우는 해당 개체와 관련하여 새로운 연구 주제나 과학적 이슈가 출현하



〈그림 3〉 시계열적 상이한 특성을 보이는 6개 페어의 연도별 출현 비율

거나 현존하는 연구 문제가 논란이 많은 상태로 해석할 수 있다. 반면 중간 연도에 비율이 급격히 감소하는 패턴을 보이는 경우 과학적 이슈가 합의를 통해 안정화된 시기로 이해할 수 있다. 각 그래프의 전체적인 추세를 확인하기 용이하도록 선형 추세선을 표현했다.

4.2.1 평균 기반 개체 페어 분석

우선, 각 페어의 연도별 출현 비율의 평균이 전체 비율의 평균보다 높다는 것은 해당 개체 페어의 관계성이 상대적으로 불확실하다는 것을 의미한다. 이는 생의학 학술 문헌에서 두 개체에 대한 언급이 불확실성 단어와 함께 출현한 비율이 높으므로 불확실한 주장이 많다는 것을 의미하고 과학적 탐구가 더 이루어져야 할 개체 관계로 추정된다. 평균이 22.42%로 가장 높은 Coronary Arteriosclerosis-Patients 페어는 그래프의 기울기가 양수(+)값을 가져 미미하지만 연도별 불확실성의 비율이 상승하는 패턴을 보였다. 이는 타 개체 페어와는 달리 해당 개체 페어의 불확실성 단어의 비율이 감소하지 않은 경우로 불확실성이 높은 개체 관계로

해석할 수 있다.

반면, 각 페어의 연도별 출현 비율의 평균이 전체 비율의 평균보다 낮다는 것은 해당 개체 페어의 관계성이 상대적으로 확실하다는 것을 의미한다. 이는 두 개체에 대한 연구 내용이 보다 확실성을 가지는 경우가 많으므로 안정화된 개체 페어로 추정할 수 있다. 평균이 8.59%로 가장 낮은 Patients-Rheumatoid Arthritis 페어는 표준편차도 2.01로 낮게 나타났으므로 불확실성이 낮은 상태가 지속적으로 안정화된 페어로 해석할 수 있다. 특히, 최댓값은 2004년의 12.66%로 모든 개체 페어의 최댓값 중 가장 낮은 최댓값을 가졌고, 범위 또한 7.82%로 모든 개체 페어 중 2번째로 가장 낮은 범위를 가졌다. 표준편차도 Diabetic-Patients 페어 다음으로 낮은 값을 보였다. 이를 통해 상대적으로도 상승・하강 변동의 폭이 작고 불확실성이 낮은 대표적인 페어인 점을 확인할 수 있다. 이 페어는 95% 신뢰도에서 연도별 불확실성 비율 변화가 유의하게 나타났다.

4.2.2 표준편차 기반 개체 페어 분석

각 페어의 연도별 출현 비율의 표준편차가 높다는 것은 연도별 비율 변동의 폭이 크다는 것으로 해당 개체 페어의 관계성이 상대적으로 불안정 하다는 것을 의미한다. 이는 각 개체 페어 출현 비율의 기울기에 따라 두 가지 해석이 가능하다. 기울기가 음수로 하강하는 추세에서 표준편차가 높다면 이전에 비해 연구의 변화가 급진적으로 확실성으로 나아가고 있으며, 반면 기울기가 양수로 상승하는 추세에서 표준편차가 높으면 해당 연구의 변화가 급진적으로 불확실성으로 나아가는 것을 시사한다. Multiple Sclerosis-Patients 페어는 표준편차가 5.32로 가장 높은 값을 가졌다. 개체 페어의 최댓값은 1996년의 29.78%이며 2008년에 5.49%의 최솟값을 가져 범위가 24.3%로 전체 페어 중 가장 높은 범위를 가졌다. 가장 높은 기울기 값을 보인 Myocardial Infarction-Patients 개체 다음으로 2번째로 높은 기울기인 -0.4592 를 가진 개체 페어로 시간이 흐름에 따라 불확실성이 두드러지게 감소한 대표적인 개체 페어이다. 이 개체 페어는 99%의 신뢰도에서 연도별 비율 변화가 유의하게 나타났다.

각 페어의 연도별 출현 비율의 표준편차가 낮다는 것은 연도별 비율 변동의 폭이 작으므로 해당 개체 페어의 관계성이 상대적으로 안정화되어 있다. 만약 불확실성 단어의 출현 비율이 높은 상태에서 비율 변동의 값이 작은 경우는 해당 개체와 관련한 연구가 발전되지 않고 정체되어 있는 상태로 해석할 수 있다. Diabetic-Patients 페어는 표준편차가 1.8로 가장 낮은 값을 가졌다. 이는 평균이 10.33%이며 최댓값은 1995년에 14.93%로 앞서 언급한 평균이 가장 낮은 Patients-Rheumatoid Arthritis 페어 다음으로 낮은 값을 가졌다. 또한 범위는 7.75%로 가장 낮은 값을 가져 시간의 흐름에 따른 비율 변동의 폭이 작은 안정적인 페어로 확인할 수 있다.

4.2.3 기울기 기반 개체 페어 분석

개체 페어의 기울기가 가장 큰 것은 연도에 따른 비율의 증감이 높은 것으로 연도가 비율에 미치는 영향을 나타내므로 결정계수와 연결지어 생각해볼 수 있다. Myocardial Infarction-Patients 개체 페어는 기울기가 음수인 0.4881로

전체 페어에서 가장 높은 기울기를 가졌다. 출현 비율 평균은 14.11%이며 범위도 상대적으로 높은 값인 18.76%를 가졌다. 결정계수도 0.561로 가장 높은 값을 가진 페어로 시간이 흐름에 따라 불확실성의 비율이 낮아진 대표적인 개체 페어이다. 이 페어는 99% 신뢰도에서 연도별 불확실성 비율 변화가 유의하게 나타났다.

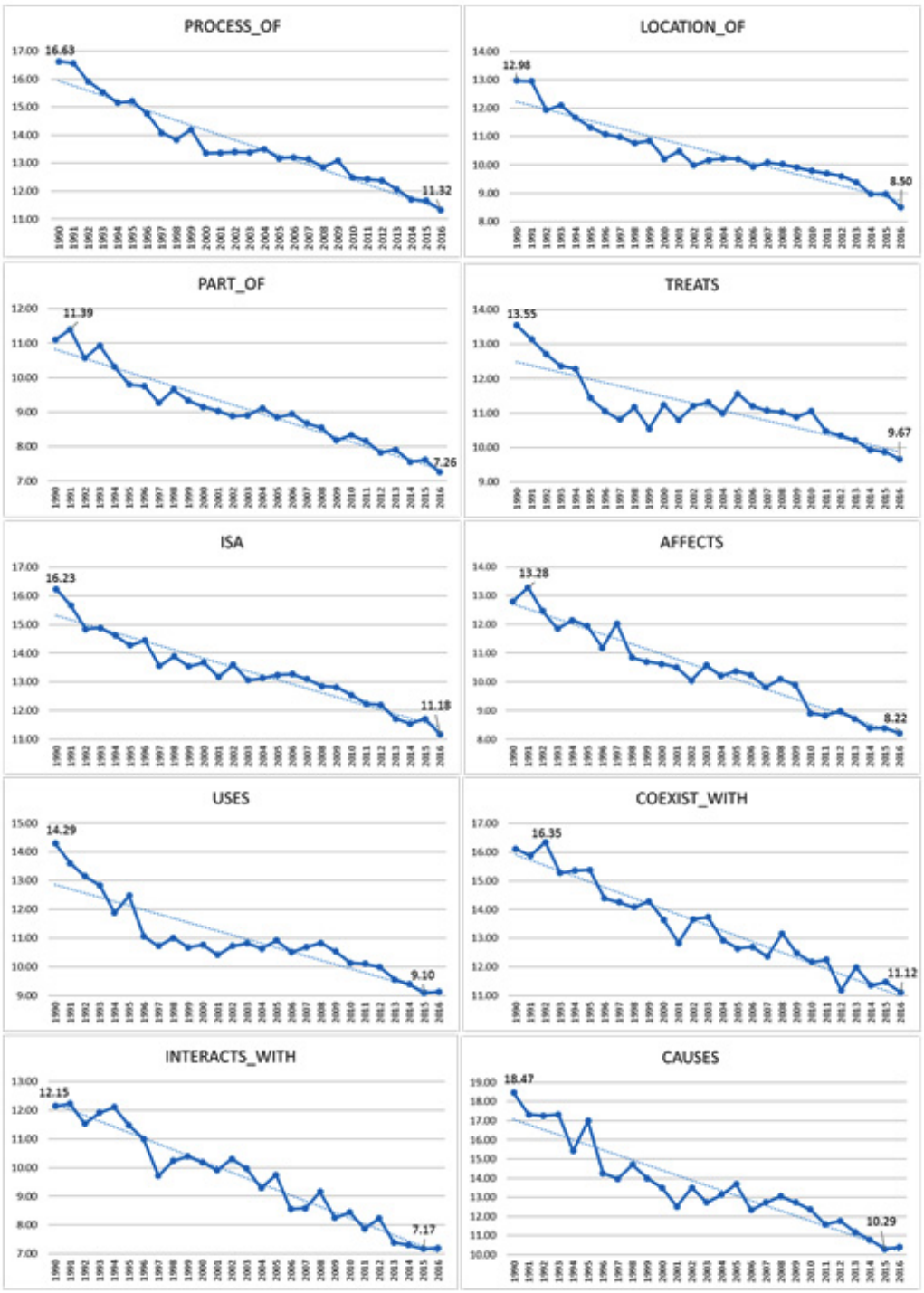
반면, 개체 페어의 기울기가 0.0162로 가장 작은 페어는 Kidney Failure, Chronic-Patients이다. 그래프의 기울기가 양수로 시간의 흐름에 따른 개체 관계의 변동은 미세하지만 상승하는 패턴을 보였다. 결정계수도 0.002로 가장 낮은 값을 가진 페어로 시간의 흐름에 따른 비율의 변화가 영향을 미치지 않는 페어이다. 이는 해당 개체 페어에 대한 주장에서 불확실성 단어를 포함한 비율이 8.85%로 평균이 가장 낮은 Patients-Rheumatoid Arthritis 다음으로 낮은 평균을 보였다. 1990년에 15.09%로 최댓값을 가진 이후 1991년 11.11%로, 1992년에는 최솟값인 4.44%까지 비율이 낮아졌다가 2000년에 14.63%로 비율이 높아졌다. 2000년대 중반 이후부터 변동의 폭이 작아졌으나 20014년부터 증가하는 패턴을 보이고 있다. 평균 기반으로 확인했을 때 연구가 이미 발전되어 있는 상태에서 지속되는 개체 페어로 판단할 수 있다.

4.3 연도별 대표 동사 유형 통계 분석

17개의 대표적인 개체 페어 중 위 〈그림 3〉의 6개 개체 페어를 비롯한 총 16개의 개체 페어는 동사 유형이 PROCESS_OF 관계이다. 예외적으로 Antibiotics-Therapeutic procedure 페어는 USES 관계를 가졌다. 따라서 이 두 동사 유형을 포함하여 상위 10순위의 대표적인 동사 유형들이 연도별로 어떠한 분포를 보이는지 살펴보기 위해 각 동사 유형의 전체 데이터 집합 대비 불확실성 단어가 포함된 데이터 집합의 연도별 비율 변화를 확인했다. 그래프로 표현한 결과는 〈그림 4〉와 같으며, 10개의 모든 동사 유형이 최신연도로 갈수록 불확실성 단어 문장 내 동사 유형의 출현 분포의 비율이 감소하는 추세를 보였다.

동사 유형의 연도별 출현 비율의 기술 통계는 〈표 7〉과 같다. 동사 유형 중에서 가장 많은 분포를 차지하면서 본 연구에서 대표적인 개체들 간의 관계를 설명하는 PROCESS_OF는 10개의 동사 유형 중 가장 높은 평균값인 13.64%를 가졌다. 불확실성 단어 출현 비율의 평균이 높다는 것은 해당 동사 유형의 불확실성의 정도가 높은 것으로 판단할 수 있다. TREATS는 범위와 표준편차가 가장 낮은 동사 유형으로 연도에 따른 비율의 변화가 가장 낮아 불확실성의 변동이 작은 동사 유형으로 확인되었다. CAUSES의 경우 1990년에 18.47%로 10개 동사 유형 중 가장 높은 비율을 차지하였고, 2015년에 10.29%로 비율의 최솟값을 보였다. 범위가 8.18%로 가장 높은 값을 가졌으며 표준편차 기준으로도 높은 편차를 보였다. CAUSES는 연도의 흐름에 따라 불확실성 감소의 비율이 타 동사 유형에 비해 높게 나타나 불확실성의 변화가 급진적임을 확인했다.

또한, 연도와 불확실성 단어 분포의 선형관계를 통계적으로 설명하기 위해 다음과 같이 연구 가설을 설정하였고 선형 회귀 분석을 수행하였다.



〈그림 4〉 10개 동사 유형의 연도별 출현 비율

〈표 7〉 상위 10개 동사 유형의 기술 통계

순위	동사 유형	평균	최댓값	최솟값	범위	표준편차
1	PROCESS_OF	13.64	16.63	11.32	5.3	1.42
2	LOCATION_OF	10.47	12.98	8.5	4.48	1.11
3	PART_OF	9.08	11.39	7.26	4.14	1.08
4	TREATS	11.18	13.55	9.67	3.88	0.93
5	ISA	13.37	16.23	11.18	5.05	1.21
6	AFFECTS	10.45	13.28	8.22	5.07	1.4
7	USES	10.97	14.29	9.1	5.19	1.28
8	COEXISTS_WITH	13.45	16.35	11.12	5.22	1.53
9	INTERACTS_WITH	9.64	12.22	7.17	5.05	1.59
10	CAUSES	13.63	18.47	10.29	8.18	2.2

〈표 8〉 상위 10개 동사 유형의 선형 회귀 분석 결과

순위	동사 유형	회귀식	결정계수	유의확률
1	PROCESS_OF	y = -0.176x + 16.108	0.926	1.15E-15**
2	LOCATION_OF	y = -0.1348x + 12.361	0.898	6.50E-14**
3	PART_OF	y =-0.134x + 10.952	0.933	3.56E-16**
4	TREATS	y = -0.1x + 12.583	0.707	4.05E-08**
5	ISA	y = -0.1497x + 15.466	0.925	1.43E-15**
6	AFFECTS	y = -0.1742x + 12.885	0.942	6.18E-17**
7	USES	y = -0.146x + 13.01	0.791	5.68E-10**
8	COEXISTS_WITH	y = -0.1894x + 16.098	0.934	2.66E-16**
9	INTERACTS_WITH	y = -0.1988x + 12.424	0.946	2.24E-17**
10	CAUSES	y = -0.2646x + 17.335	0.876	8.19E-13**

p**〈.01

H_A = 시간이 흐름에 따라 불확실성 단어 데이터 집합의 동사 유형 출현 비율에 영향을 미칠 것이다.

각 동사 유형의 선형회귀식과 결정계수, 유의확률은 〈표 8〉과 같다. 앞서 설명한 바와 같이 모든 동사 유형이 연도의 변화에 따라 불확실성 단어 포함 데이터 집합문장의 출현 비율이 감소하였고 특히, CAUSES의 기울기가 0.2646으로 가장 크게 감소하는 패턴을 보였다. 회귀식의 결정계수는 0.707~0.946까지 높은 설명력을 가졌다. 유의확률은 99%의 신뢰도에서 모두 귀무가설을 기각하여 시간의 흐름에 따라 동사 유형을 포함한 불확실성 단어의 비율이 유의하게 감소하였다.

4.4 연도별 대표 개체의 버스티니스 기반 통계 분석

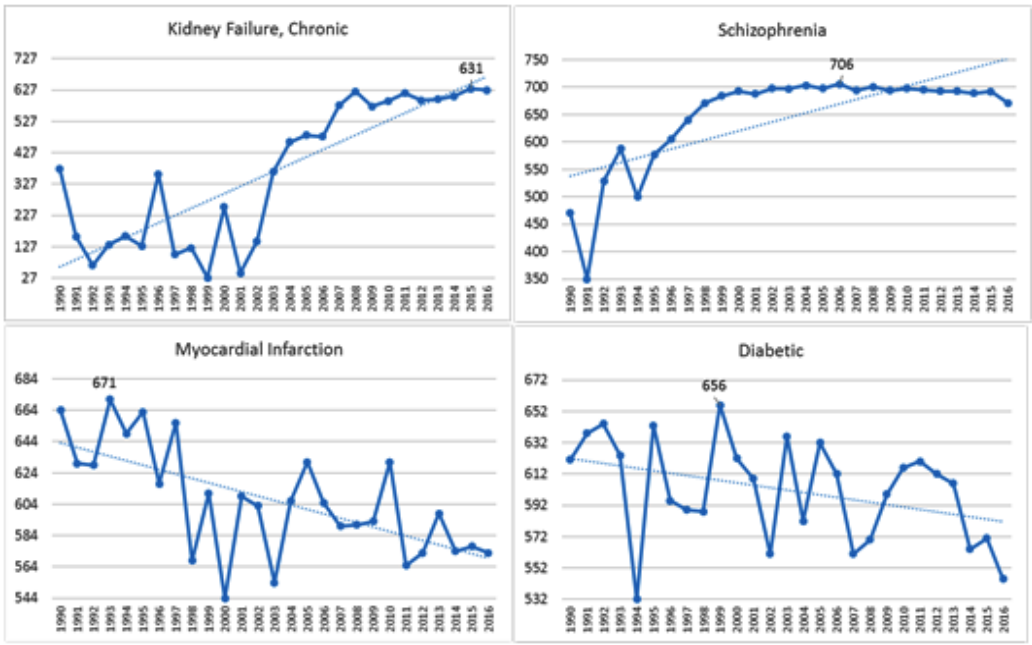
출현 비율 기반 분석과 더불어 불확실성 단어 데이터 집합에서 개체의 상대적 중요도를 연도별 흐름으로 파악하기 위해 개체의 버스티니스 값을 산출했다(Madsen, Kauchak, & Elkan,

2005). 'Burstiness'란 한 문헌에 여러 번 등장한 단어의 행태를 의미(Church & Gale, 1995; Katz, 1996)하는 것으로 단순 출현 빈도의 개념과는 달리 특정 문헌에 집중적으로 나타나 식별능력이 높은 중요성을 가지는 단어들을 추출하기 위한 유용한 방법으로 알려져 있다.

각 연도별 최대 3,000개의 개체에 대해 버스티니스 값을 추출하였고, 1년 이상 값이 출현한 9,706건의 개체 중 모든 연도에 버스티니스 값이 추출된 726건을 분석 대상으로 삼았다. 버스티니스 값은 연도 내에서 상대적으로 산출되므로 연도별 분석을 수행하기 위해 해당 값의 순위를 산출했고 그래프 산출 시 불확실성의 증감과 일치하도록 표현하기 위해 순위에 역순을 취했다. 또한 실제 값의 연도별 비율을 산출하였다.

앞서 분석한 대표적인 17개 개체 페어의 고유한 19개 개체의 순위와 비율을 살펴보았다. 일반적인 단어로 분류된 Patients와 Therapeutic procedure는 순위의 차이가 발생하지 않고 독보적으로 1, 2위를 차지한 개체였다. 나머지 17개 개체들은 변동의 폭의 차이는 있지만 모두 연도별 흐름에 따라 상승 또는 하강하는 패턴을 보였다. 이 중 순위 변동의 폭이 큰 각 패턴별 대표적인 2건의 개체는 〈그림 5〉와 같다. 상단의 Kidney, Failure, Chronic 개체와 Schizophrenia 개체는 상승하는 패턴을 보인 대표적인 개체들이다. 또한 하단의 Myocardial Infarction, Diabetic 개체는 연도별 흐름에 따라 하강하는 패턴을 보인 대표적인 개체들이다.

시간이 흐름에 따라 상승하는 패턴은 불확실성 단어 포함 데이터 집합 내 개체의 버스티니



〈그림 5〉 대표적인 4개 개체의 연도별 출현 비율

스 값 기반의 상대적 중요도가 높아진 것으로 해당 개체의 불확실성에 대한 주장이 증가한 것이다. 이는 개체에 대한 학술적 이슈나 문제가 발전하는 형태이며, 해당 개체를 포함하는 주제의 프레임이 만들어지고 있는 단계로 해석할 수 있다. 반면, 시간이 흐름에 따라 하강하는 패턴은 해당 개체의 버스티니스 값이 감소하여 불확실성 단어 문장 내에서 개체의 상대적 중요도가 낮아진 것이다. 개체에 대한 학술적 주장의 불확실성이 감소한 것으로 학술적 문제가 해결되어 안정화된 단계로 해석할 수 있다. 즉, 해당 개체의 순위가 연도의 흐름에 따라 높아진다면 불확실성이 증가하며, 반면 낮아진다면 불확실성이 감소한 것으로 해석할 수 있다.

〈표 9〉에서 확인할 수 있듯이, Kidney, Failure, Chronic 개체는 4개의 개체 중 버스티니스 값 기반의 비율과 순위가 가장 낮게 나타난 개체로 1990년대에는 순위의 변동이 보이다가 1999년에 최소 순위인 27위를 차지하였다. 이 후 2001년부터 급속도로 증가하는 패턴을 보였고 2015년에는 631순위로 상대적 중요도가 가장 높아졌다. 범위가 604이며, 표준편차가 219.42로 연도의 흐름에 따라 가장 큰 변화를 보인 대표 개체이다. 특히 앞서 분석한 기준에 따라 2000년대 들어 학술적 문제가 논란이 많은 상태의 단계로 해석할 수 있다. Schizophrenia는 4개의 개체 중 버스티니스 값 기반의 비율과 순위가 가장 높게 나타난 개체로 1991년에 최소 순위인 350위를 차지하였고 1994년 이후 급격히 증가하는 패턴을 보였다. 2006년에 706순위로 Schizophrenia의 불확실성에 대한 학술적 진술이 가장 높게 나타났다. 2000년대 들어 순위 변동의 폭이 줄어들어 지속적으로 불확실성이 높은 단계가 유지되는 단계로 해석이 가능하다.

〈표 9〉 버스티니스 기반 대표적인 4개 개체의 기술 통계

통계	기준	개체명			
		Schizophrenia	Myocardial Infarction	Diabetic	Kidney Failure, Chronic
전체 순위 (726)	순위	75	110	116	343
	비율	43	111	121	257
대표 개체 순위 (19)	순위	10	13	15	19
	비율	5	14	15	19
평균	순위	645	606	602	367
	비율	0.0013	0.0007	0.0007	0.0004
최댓값	순위	706	671	656	631
	비율	0.0021	0.0013	0.001	0.0008
최솟값	순위	350	544	532	27
	비율	0.0003	0.0005	0.0004	0.0001
범위	순위	356	127	124	604
	비율	0.0018	0.0008	0.0005	0.0007
표준편차	순위	87.36	34.5	31.92	219.42
	비율	0.0005	0.0002	0.0001	0.0002

반면, 시간의 흐름에 따라 하강하는 패턴을 보인 Myocardial Infarction은 변동이 존재하지만 1993년 671순위로 최대 순위를 보였으며 2000년에 544순위로 최소 순위를 보인 개체이다. 2010년 이후 변동 폭이 줄어들면서 안정화를 보이고 있다. Diabetic 개체는 1990년대 초반 600대의 순위에서 1994년에 532순위로 순위가 급격히 하강한 후 1995년에 다시 증가하였으며, 이후 1999년에 656순위로 최대순위를 가진 후 변동을 보인 개체이다. 2010년 이후부터 지속적으로 감소하는 패턴을 보였다. 앞서 설명한 바와 같이 불확실성 단어 문장 내 해당 개체의 상대적 중요도가 감소한 것이므로 개체와 관련한 주제의 학술적 문제가 해결되어가는 과정으로 해석할 수 있다.

대표되는 4개 개체의 연도별 흐름에 따른 버스티니스 값의 차이의 유의성 검증을 수행하기 위해 다음과 같이 연구 가설을 설정하였다.

H_A = 시간이 흐름에 따라 불확실성 단어 데이터 집합의 개체의 상대적 중요도에 영향을 미칠 것이다.

선형 회귀 분석을 통해 회귀식과 결정계수, 유의확률, 그리고 평균적인 잔차(추정값과 관측값의 차이)의 크기를 나타내는 평균 제곱 오차(Mean Squared Error, MSE)를 산출한 결과는 〈표 10〉과 같다.

연도별 726개 개체의 상대적 중요도를 비율로 환산하였기에 회귀식의 기울기와 Y절편은 매우 작은 값을 가지지만, 추정된 회귀식은 모두 99%의 신뢰도에서 귀무가설을 기각하여 연도에 따라 불확실성이 유의하게 증가 또는 감소하는 것을 통계적으로 확인했다. 평균 제곱오차도 모두 작은 값을 가져 실제 값들이 회귀선에 근접해 있으므로 높은 선형 관계를 가진다는 점을 확인했다.

불확실성 단어가 포함된 데이터 집합에서 생의학 지식으로 대변되는 생의학 개체와 관계성을 분석하여 연도별 흐름을 살펴볼 수 있었으며, 두 개체와 관계의 발전 단계를 불확실성 단어 출현 비율의 패턴과 더불어 버스티니스 값의 상대적 중요도 기반 순위와 비율을 기반으로 불확실성 기반 생의학 지식의 동향에 대한 종합적인 분석이 가능했다.

5. 결 론

본 연구에서는 생의학 학술 문헌의 불확실성 기반 생의학 지식의 동향을 파악하기 위한 연구로 문장 내에 존재하는 불확실성 단어 문장 기반 생의학 개체와 관계의 특성을 확인하여 지식의 흐름을 살펴보고자 했다.

〈표 10〉 대표적인 4개 개체의 선형 회귀 분석 결과

순위	개체	회귀식	결정계수	유의확률	평균오차제곱
1	Schizophrenia	y = 5E-05x + 0.0006	0.451	1.26E-04**	1.73E-07
2	Myocardial Infarction	y = -2E-05x + 0.001	0.544	1.12E-05**	2.21E-08
3	Diabetic	y = -1E-05x + 0.0008	0.384	0.001**	1.43E-08
4	Kidney Failure, Chronic	y = 2E-05x + 2E-05	0.743	7.68E-09**	1.41E-08

p**〈.01

PubMed에서 불확실성 단어 196건을 질의어로 하여 문헌 집합을 수집하였고 전처리를 거쳐 SemMedDB를 적용하여 생의학 지식을 추출하였다. 전문가의 검증을 통해 선별된 17개 개체 페어의 연도별 불확실성 단어 출현 비율을 분석했고 회귀 모형을 통해 연도의 흐름에 따른 불확실성의 변화의 유의성 검증을 수행했다. 개체 페어 7건을 비롯하여 총 10건의 동사 유형은 모두 연도별 흐름에 따라 유의하게 불확실성 단어가 감소하는 패턴을 보였다. 또한 개체의 식별력을 의미하는 버스티니스 값을 연도별로 산출하여 대표 개체의 유의한 불확실성 증감 패턴을 확인하였다. 이를 통해 생의학 학술적 지식의 흐름을 불확실성 단어의 증감을 기반으로 분석할 수 있는 가능성을 검토하였다.

본 연구는 기존에 수행된 단순 출현 빈도 기반 불확실성 연구들에서 더 나아가 시간의 흐름에 따른 불확실성 단어의 특성 변화를 분석하여 대표 개체와 관계성의 통계 기반 유의한 증감 패턴을 확인했다는데 의의가 있다. 이는 문헌 내에서 유의미한 지식을 발견하고 분석함으로써 지식의 발전과 동향을 살펴보았다는 점에서 정보학적인 시사점을 지닌다.

향후 연구로는 불확실성 현상을 기반으로 예측이 가능한 모형을 제안하고자 한다. 이 때 시간적인 변인 외에 불확실성의 증감에 영향을 미치는 외부 변인을 함께 고려하여 예측력을 높이고 학술 지식의 발전에 영향을 미치는 일반적인 관계를 제안할 것이다. 또한 정확한 결과를 도출하기 위해 문장 내 출현한 불확실성 단어와 SemMedDB기반으로 추출된 의미적 술어간 거리의 근접성에 따라 데이터 처리를 수행할 필요가 있다. 생의학 개체와 불확실성 단어 간의 실질적인 관계성을 지정하는 방법을 추가하여 유의미한 결과 분석이 가능하도록 할 예정이다. 더불어 본 연구 결과가 임상의학 전문가들에게 유의미한 정보로 활용되기 위해 의학 전문가의 의학적 검증을 통해 연구결과의 효용성을 입증하고 생의학 가설 발견에 대해 논의하고자 한다.

참 고 문 헌

허고은 (2019). 토픽 모델링 기반 과학적 지식의 불확실성의 흐름에 관한 연구. 정보관리학회지, 36(1), 191-213. http://dx.doi.org/10.3743/KOSIM.2019.36.1.191

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

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. Information (International Social Science Council), 14(6), 19-47. https://doi.org/10.1177/053901847501400602

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., 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

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. Natural Language Engineering, 1(2), 163-190. https://doi.org/10.1017/S1351324900000139

Cordner, A., & Brown, P. (2013). Moments of uncertainty: Ethical considerations and emerging contaminants. In Sociological Forum, 28(3), 469-494. https://doi.org/10.1111/socf.12034

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.

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

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

Ioannidis, J. P., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. Journal of Clinical Epidemiology, 58(6), 543-549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. Natural Language Engineering, 2(1), 15-59. https://doi.org/10.1017/S1351324996001246

Kuhn, T. S. (1970). The structure of scientific revolutions. University of Chicago Press.

Madsen, R. E., Kauchak, D., & Elkan, C. (2005). Modeling word burstiness using the dirichlet distribution. In Proceedings of the 22nd International Conference on Machine Learning, (August): 545-552. https://doi.org/10.1145/1102351.1102420

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, 3111-3119.

Palmer, F. R. (2014). Modality and the English modals. Routledge.

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

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

Russell, S., Norvig, P., & Intelligence, A. (1995). Artificial intelligence: A modern approach prentice-hall. Englewood cliffs, NJ.

Shwed, U., & Bearman, P. S. (2010). The temporal structure of scientific consensus formation. American Sociological Review, 75(6): 817-840. https://doi.org/10.1177/0003122410388488

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

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.

Szarvas, G., Vincze, V., Farkas, R., Móra, G., & Gurevych, I. (2012). Cross-genre and cross-domain detection of semantic uncertainty. Computational Linguistics, 38(2), 335-367. https://doi.org/10.1162/COLI_a_00098

Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. BMC bioinformatics, 12(1), 393. https://doi.org/10.1186/1471-2105-12-393

Vincze, V. (2013). Weasels, hedges and peacocks: Discourse-level uncertainty in Wikipedia articles. International Joint Conference on Natural Language Processing, (October): 383-391. Nagoya, Japan.

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

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

• **국문 참고문헌에 대한 영문 표기**
(English translation of references written in Korean)

Heo, G. E. (2019). The stream of uncertainty in scientific knowledge using topic modeling. Journal of the Korean Society for Information Management, 36(1), 191-213. http://dx.doi.org/10.3743/KOSIM.2019.36.1.191