Word2Vec과 WordNet 기반 불확실성 단어 간의 네트워크 분석에 관한 연구

Network Analysis between Uncertainty Words based on Word2Vec and WordNet

허 고 은 (Go Eun Heo)*

목 차

초 록

과학에서 지식의 불확실성은 명제가 현재 상태로는 참도 거짓도 아닌 불확실한 상태를 의미한다. 기존의 연구들은 학술 문헌에 표현된 명제를 분석하여 불확실성을 의미하는 단어를 수동적으로 구축하고 구축한 코퍼스를 대상으로 규칙 기반, 기계 학습 기반의 성능평가를 수행해왔다. 불확실성 단어 구축의 중요성은 인지하고 있지만 단어의 의미를 분석하여 자동적으로 확장하고자 하는 시도들은 부족했다. 한편, 계량정보학이나 텍스트 마이닝 기법을 이용하여 네트워크의 구조를 파악하는 연구들은 다양한 학문분야에서 지적 구조와 관계성을 파악하기 위한 방법으로 널리 활용되고 있다. 따라서, 본 연구에서는 기존의 불확실성 단어를 대상으로 Word2Vec을 적용하여 의미적 관계성을 분석하였고, 영어 어휘 데이터베이스이자 시소러스인 WordNet을 적용하여 불확실성 단어와 연결된 상위어, 하위어 관계와 동의어 기반 네트워크 분석을 수행하였다. 이를 통해 불확실성 단어의 의미적, 어휘적 관계성을 구조적으로 파악하였으며, 향후 불확실성 단어의 자동 구축의 확장 가능성을 제시하였다.

ABSTRACT

Uncertainty in scientific knowledge means an uncertain state where propositions are neither true or false at present. The existing studies have analyzed the propositions written in the academic literature, and have conducted the performance evaluation based on the rule based and machine learning based approaches by using the corpus. Although they recognized that the importance of word construction, there are insufficient attempts to expand the word by analyzing the meaning of uncertainty words. On the other hand, studies for analyzing the structure of networks by using bibliometrics and text mining techniques are widely used as methods for understanding intellectual structure and relationship in various disciplines. Therefore, in this study, semantic relations were analyzed by applying Word2Vec to existing uncertainty words. In addition, WordNet, which is an English vocabulary database and thesaurus, was applied to perform a network analysis based on hypernyms, hyponyms, and synonyms relations linked to uncertainty words. The semantic and lexical relationships of uncertainty words were structurally identified. As a result, we identified the possibility of automatically expanding uncertainty words.

키워드: 텍스트 마이닝, 계량정보학, 불확실성, Word2Vec, 워드넷, 네트워크 분석
Text Mining, Bibliometrics, Uncertainty, Word2Vec, WordNet, Network Analysis

* 연세대학교 문헌정보학과 연구교수(goeun.heo@yonsei.ac.kr / ISNI 0000 0004 7707 1202)
논문접수일자: 2019년 7월 16일 최초심사일자: 2019년 8월 5일 게재확정일자: 2019년 8월 16일
한국문헌정보학회지, 53(3): 247-271, 2019. [http://dx.doi.org/10.4275/KSLIS.2019.53.3.247]

1. 서 론

불확실성이란 어떤 특정한 무언가에 대해 현재 상태가 아직 해결이 나지 않은, 확실하지 않은, 변화 가능성이 있는 상태를 의미한다. 불확실성과 관련된 연구로 언어학적인 측면에서 추측(speculation)을 의미하는 hedging은 인식양태(epistemic modality)의 한 부분으로 학술연구의 문장 내에서 조동사, 형용사, 부사, 동사로 표현되며 다양한 관점으로 해석되어 오고 있다. 즉, 학술 연구의 집필에 중요한 역할을 하는 hedging은 아직 입증되지 않은 새로운 주장을 올바르게 전달하는 수단으로 중대한 의미를 지닌다(Hyland 1996, 1998).

기존의 불확실성을 밝히기 위한 연구들은 주로 언어학적인 관점과 자연어 처리 연구에서 수행되어 왔다. 생의학 영역에서 정보의 범위와 유형에 대해 지침을 개발하고 주석을 수행하여 코퍼스를 구축해왔다(Szarvas, Vincze, Farkas and Csirik 2008; Vincze, Szarvas, Farkas, Móra and Csirik 2008; Szarvas, Vincze, Farkas, Móra and Gurevychet 2012; Vincze 2013). 또한 해당 문장이 불확실한 문장인지에 대한 텍스트 자동 분류 문제를 다루기 위해 규칙 기반, 지도 학습 기반, 약한 지도학습 기반으로 불확실성을 인식하는 접근법을 제안했다(Light, Qiu and Srinivasan 2004; Medlock and Briscoe 2007; Kilicoglu and Bergler 2008; Szarvas 2008; Thompson, Nawaz, McNaught and Ananiadou 2011, Malhotra, Younesi, Gurulingappa and Hofmann-Apitius 2013).

이처럼 불확실성이란 화자가 전하고자 하는 명제(proposition)에 대해 얼마나 많은 확실한 증거를 가지고 있는지와 관련되어 있으므로(Palmer 2014) 학술문헌 내에 표현된 지식의 믿음, 신뢰의 정도, 평가 및 판단과 연결된 불확실성을 의미하는 단어를 규정하는 것과 더불어 이러한 불확실성 단어들 간의 구조적인 관계성을 살펴보는 것이 중요하다. 하지만 기존 연구들은 불확실성 단어들의 특성을 불확실성의 정도에 따라 단계를 구분하여 분석하는데 그쳤으며 단어들 간의 실질적인 관계성을 살펴보지는 않았다. 또한 기존의 연구들은 불확실성 단어와 hedging을 수작업으로 구축하는데 초점을 두었으며, 규칙 기반 또는 기계 학습 기반의 접근법을 제안하여 구축한 코퍼스를 기반으로 성능평가를 수행했다. 수작업의 비용을 최소화하기 위해 자동화된 시스템을 제한하였지만 학술 영역 간 불확실성 단어의 이질성(heterogeneity)으로 인해 새로운 영역과 장르에 불확실성 단어를 발견하기 위한 학습 데이터를 구축하기 위해서는 약 3,000건에서 5,000건의 문장을 수작업으로 구축(Szarvas, Vincze, Farkas, Móra and Gurevychet 2012)해야 하는 한계점이 존재한다.

따라서 본 연구에서는 구축된 불확실성 단어들의 관계성을 딥러닝 기법으로 잘 알려져 있는 두 층의 인공신경망 기법인 Word2Vec과 영어 시소러스인 WordNet을 기반으로 단어 간의 의미적 관계와 어휘적 관계를 중심으로 구조적인 특성을 확인하고자 한다. Chen, Song and Heo(2018)에서 제안한 불확실성 단어 리스트를 대상으로 Word2Vec 방법론을 적용한 불확실성 단어 간의 의미적 관계성을 살펴보았다. 또한 WordNet 기반 불확실성 단어 간의 상하관계와 더불어 각 불확실성 단어의 동의어를 확인함으로써 단어들 간의 구조적인 특성을 살

펴보고 어휘적 단어 확장 가능성을 확인했다. 분석 결과는 단어와 단어 간의 관계성을 파악하는데 용이한 소셜 네트워크의 분석 기법인 네트워크 분석을 수행하여 시각적으로 확인했다.

2. 이론적 배경

2.1 자연어 처리 기반 불확실성 연구

기존의 불확실성과 관련하여 자연어 처리(Natural Language Processing, NLP) 접근을 통해 불확실성 단어를 추출하고 범위를 자동으로 인식하는 방법론을 개발한 연구들이 시도되었다. 컴퓨터 언어학(computational linguistics) 커뮤니티의 CoNLL(Computational Natural Language Learning)은 1999년부터 시작된 경쟁적으로 공유된 과업을 제공하는 컨퍼스이다(Farkas, Vincze, Móra, Csirik and Szarvas 2010). CoNLL-2010에서는 불확실성 발견을 두 단계로 나누어 실험을 수행했다. 첫 단계는 불확실한 정보를 포함하는 문장을 확인하는 과업이며 두 번째 단계로는 문장 내의 추측성을 지니는 텍스트의 범위(span)를 인식하는 과업이다. 또한 두 가지의 학문 영역인 생의학 영역의 문헌 데이터와 위키피디아 백과사전의 데이터 영역으로 구분된다. 생의학 학술 문헌에서 전형적인 hedge 단어는 4가지 카테고리로 구성된다: 1) 조동사, 2) hedging 동사 또는 추측 내용이 포함된 동사, 3) 형용사 또는 부사, 4) 접속사.

생의학 문헌에서는 시퀀스 레이블로 알려진 조건 임의 필드(Conditional Random Fields, CRF) 기법이 좋은 성능을 보였고, 위키피디아 데이터에서는 bag-of-word 기반의 문장 분류가 가장 좋은 성능을 보였다. 또한 텍스트 범위를 인식하는 시스템은 클래스 레이블의 수와 기계 학습 기법에 따라 성능의 차이를 보였다. CoNLL에서는 수작업 코퍼스를 활용하여 기계학습으로 불확실성 단어를 발견하는 방법들이 개발되었다. 방법론으로 토큰 분류(Fernandes, Crestana and Milidiú 2010; Sánchez, Li and Vogel 2010) 또는 시퀀스 레이블 접근법(Li, Shen, Gao and Wang 2010; Rei and Briscoe 2010; Tang, Wang, Wang, Yuan and Fan 2010; Zhang, Zhao, Zhou and Lu 2010)을 적용했다. Rei와 Briscoe(2010), Zhang, Zhao, Zhou and Lu(2010)는 해당 단어의 의존(dependency) 관계 유형도 함께 이용했다.

Hedging 관련 연구들은 추측성 문장과 추측성이 아닌 문장에 대한 텍스트 자동 분류 문제를 다루었다. 이들은 규칙 기반, 지도 학습 기반, 약한 지도 학습 기반으로 불확실성을 인식하는 접근법을 제안했다(Light, Qiu and Srinivasan 2004; Medlock and Briscoe 2007; Kilicoglu and Bergler 2008; Szarvas 2008; Thompson, Nawaz, McNaught and Ananiadou 2011; Malhotra, Younesi, Gurulingappa and Hofmann-Apitius 2013). Kilicoglu와 Bergler(2008)는 Hyland(1998)의 어휘적 hedging을 WordNet과 UMLS SPECIALIST lexicon을 이용하여 반자동적인 방법으로 확장했다. 첫째로, WordNet의 동의어 집합인 신셋(synonym sets, synsets)을 확인하여 동의어를 포함했고, 코퍼스에 출현한 단어만 포함시켰다. 또한 동사와 형용사의 명사화(nominalization)를 수행했다. 반자동화된

가중치 스키마를 통해 추측성의 강도를 더 정확하게 확인할 수 있었다.

대부분의 연구들은 수작업 기반으로 주석을 수행함으로써 불확실성 단어와 hedging 단어를 발견하고 불확실성을 의미하는 범위를 지정하여 코퍼스를 구축했다. 또한 규칙 기반 또는 기계 학습 기반의 접근법을 제안하여 구축한 코퍼스를 기반으로 성능평가를 수행했다. 수작업 구축의 비용을 최소화하기 위한 자동화된 시스템들을 제안했음에도 불구하고 학술 영역 간 불확실성 단어의 이질성(heterogeneity)으로 인해 새로운 영역과 장르에 불확실성 단어를 발견하기 위한 학습 데이터를 구축하기 위해서는 약 3,000건에서 5,000건의 문장을 수작업으로 구축(Szarvas, Vincze, Farkas, Móra and Gurevychet 2012)해야 하는 한계점이 존재한다.

2.2 소셜 네트워크 분석과 불확실성

소셜 네트워크 분석이란 1980년대 이후 수학의 그래프 이론에서 비롯된 분석 방법론으로 네트워크 구조를 행위자 또는 사물로 표현되는 노드와 이들 간의 상호작용을 의미하는 링크를 기반으로 표현한다. 현대 사회학의 핵심적인 기술로 정보학적인 관점에서 계량정보학이나 텍스트 마이닝 기법을 적용하여 소셜 구조를 파악할 수 있으며 현재는 사회학, 정보학, 커뮤니케이션학, 경영학 등을 비롯하여 학문을 망라하여 사회구조와 관계성, 속성을 분석한다. 특히 최근에는 소셜 네트워크 서비스(SNS)의 등장으로 다양한 분야에서 연구가 증가하였으며 영향력이 증대되었다.

문헌정보학 영역에서의 계량정보학 분석 기법은 저널 논문이나 인용 분석과 같은 출판물의 영향력을 통계적으로 측정하기 위한 분석 방법이다. 전통적으로 동시인용 분석(Co-citation analysis) 또는 동시출현단어 분석(Co-word analysis)을 이용하여 학문분야의 지적 구조를 파악하는 연구들이 수행되어 왔다. 허고은과 송민(2013)은 저자동시인용 분석과 동시출현단어 분석을 기반으로 의학과 정보학을 융합한 학제적 학문인 의료정보학 분야의 지적구조를 파악하였다. 네트워크 분석을 통해 융합 학문들 간의 관계를 파악하고 의료정보학 분야의 학문적 성향을 살펴보았다. Daim, Rueda, Martine and Gerdsri(2006)는 신흥 기술을 예측하기 위해 유용한 데이터로 활용 할 수 있는 계량정보학과 특허 분석을 통해 새로운 기술 영역에 대한 예측을 수행하였다. 또한 Madani와 Weber(2016)는 Web of Science 데이터베이스에서 143개 논문을 대상으로 계량정보학 분석과 키워드 기반의 네트워크 분석을 적용하였다. 네트워크의 클러스터 분석을 통해 특허 분석이 어떠한 단계로 발전하는지 동향을 살펴보았다. Geaney, Scutaru, Kelly, Glynn and Perry(2015)는 1951년부터 2012년까지의 제2형 당뇨병 연구 결과를 계량정보학 분석 방법을 이용하여 분석했다. 제목, 저자의 소속, 출판연도 등의 메타데이터를 추출하여 국가와 기관, 저널 등을 순위화하여 분석하였다.

Kostoff, Del, Humenik, Garcia and Ramirez(2001)는 텍스트 마이닝과 계량정보학의 인용 마이닝을 기반으로 특정 연구가 다른 연구, 기술 개발 또는 응용에 영향을 미치는 경로를 식별하고 이용자 집단의 기술과 인프라의 특성을

식별하는 방법을 설명하였다. 인용 계량정보학과 텍스트 마이닝의 결합을 통해 시너지 효과를 낼 수 있으며 연구의 영향력을 식별할 수 있다는 점을 강조했다.

특히, 문헌의 연도별 데이터를 기반으로 시기를 구분하여 계량정보학 분석과 텍스트 마이닝 기법을 결합한 동향 분석 연구들이 수행되어 왔다. Song, Heo and Kim(2014)은 2000년부터 2011년까지의 생물정보학 분야의 컨퍼런스 데이터를 DBLP로부터 수집한 후 네 시기로 구분하여 학문의 발전 동향을 살펴보았다. Heo, Kang, Song and Lee(2017)는 생물정보학 분야의 연구 발전 동향을 종합적으로 살펴보기 위해 1996년부터 2015년까지의 데이터를 수집하여 네시기로 구분하였고 ACT(Author-Conference-Topic) 모델을 적용하여 3개의 대표적인 메타데이터인 저자, 저널, 핵심 구(key-phrase) 분석을 통해 학문 분야의 발전 패턴을 분석하였다.

또한 특정 질병 연구에 대한 분석을 시도한 연구들이 있다. Song, Heo and Lee(2015)는 알츠하이머 병 관련 연구 동향을 네트워크 분석과 내용 분석을 기반으로 파악하였다. 네트워크 분석에는 5개의 대표적인 중심성 지표를 기반으로 알츠하이머 병 관련 개념과 개념 간의 관계를 살펴보았고 내용 분석으로는 문헌 내 단어들 간의 분포를 기반으로 주제를 발견하는 DMR 토픽 모델링을 적용하여 연도별 시계열적 분포 변화를 분석하였다. Jeong, Heo, Kang, Yoon and Song(2017)은 췌장암 관련 약물 연구의 동향을 두 상이한 데이터 자원인 학술 문헌과 임상실험 데이터를 기반으로 비교분석한 연구를 수행하였다. DMR 토픽 모델링 기법을 적용하여 약물 클러스터 감지 분석과 시계열적 약물 연구 동향 분석을 수행하였고 약물 기반의 네트워크를 분석하였다.

최근의 연구로 Chen, Song and Heo(2018)는 불확실성을 의미하는 불확실성 단어를 정의하고 초기 단어와 Word2Vec 모델을 기반으로 확장된 단어들의 관계성을 시각화하여 네트워크 분석을 수행했다. 또한 허고은(2019)은 불확실성 단어를 기반으로 문헌 내 생의학 지식의 시계열적 흐름을 살펴보기 위해 DMR 토픽 모델링을 적용하여 생의학 개체의 불확실성 기반 토픽의 동향을 살펴보았다. 연구 결과 과학적 지식의 표현이 불확실성이 감소하는 패턴으로 연구가 발전하고 있음을 확인하였다. 더불어 허고은, 송민(2019)은 불확실성 단어를 기반으로 생의학 학술문헌에서 시간의 흐름에 따른 생의학 지식의 불확실성의 변화를 살펴보았다. 선형 회귀 분석을 수행하여 대표적인 생의학 개체와 동사 유형의 증감 패턴의 유의성을 확인하였으며 버스티니스 값을 기반으로 생의학 개체의 상대적 중요도를 연도별 흐름으로 파악하였다.

기존 연구들은 학술 영역 간의 특수성을 기반으로 특정 영역에 국한된 불확실성 단어들을 구축해왔다. 불확실성 단어는 영역의 범위를 망라하여 활용되어 오기에 범용적으로 활용할 수 있는 일반화된 불확실성 단어들을 확장할 필요가 있다. 또한 기존의 연구들은 수작업 기반으로 코퍼스를 구축해왔고, 구축한 코퍼스를 대상으로 성능평가를 수행하는 연구들을 수행해왔다. 시간과 비용 소모가 큰 작업들을 최소화하기 위해서는 텍스트 마이닝 기법을 적용한 자동화된 불확실성 단어 구축과 확장이 필요하

다. 더불어 기존의 연구들은 불확실성 단어를 구축하는 과정의 중요성을 인식하고 있지만 단어들 간의 사전적, 어휘적 의미를 기반으로 불확실성 단어의 특성과 단어 간의 관계성을 파악한 연구들은 시도되지 않았다. 불확실성 단어들 간의 관계성을 구조적으로 파악한다면 불확실성 단어의 구축에 사전 정보를 제공할 수 있을 뿐만 아니라 자동화된 방법론에 정확성과 효율성을 높일 수 있다.

따라서 본 연구에서는 기존의 Chen, Song and Heo(2018)의 연구에서 구축한 196개의 일반적인 불확실성 단어를 기반으로 이들 간의 의미적 관계성을 Word2Vec을 적용하여 살펴보며, WordNet을 기반으로 불확실성 단어들 간의 사전적인 관계성을 분석하고자 한다.

3. 연구 설계

3.1 Word2Vec 모델

Word2Vec 모델은 최근에 기계 학습 커뮤니티에서 중점적으로 관심을 가지게 된 기법으로 두 층의 인공 신경망으로 이루어진 텍스트 처리기법이다(Mikolov, Sutskever, Chen, Corrado and Dean 2013). 데이터 집합의 모든 단어들이 고유한 벡터로 변경되는 단어 임베딩을 통해 단어 간의 유사성을 계산한다. Word2Vec에서 단어 임베딩을 구하는 과정을 신경 단어 임베딩이라고 하며 CBOW(Continuous Bag of Words)와 skip-gram의 두 종류의 학습 방법이 있다. CBOW는 주변 단어의 맥락을 통해 목적 단어를 예측하는 것이고, skip-gram은 한 단어를 입력하여 주변에 올 수 있는 단어를 예측하는 것이다. Word2Vec 모델은 대표적으로 PubMed Word2Vec 모델과 구글 뉴스 Word2Vec 모델이 있다. PubMed Word2Vec 모델은 PubMed에서 제공하는 총 22,120,269건의 문헌으로 window size를 5로 설정하여 skip-gram 모델인 계층적 소프트맥스 학습(hierarchical softmax training)을 사용하였고, 자주 출현한 단어의 서브샘플링 임계값(frequent word subsampling threshold)을 0.001로 하여 총 2,351,706단어 간의 200차원의 벡터를 구성한 모델이다(Pyysalo, Ginter, Moen, Salakoski and Ananiadou 2013).

구글 뉴스 Word2Vec 모델은 약 천억개 단어의 구글 뉴스 데이터셋에서 0.00001(과학적 표기법: 1e-5)의 임계값을 이용한 서브샘플링과 각 긍정적인 예시 당 3개의 부정적인 예시를 이용한 부정 샘플링(negative sampling)으로 CBOW 알고리즘을 통해 사전학습(pre-trained)하여 300만 단어와 구(phrase)로 구성된 300차원의 벡터를 포함하고 있는 모델이다(Mikolov, Sutskever, Chen, Corrado and Dean 2013). 본 연구에서는 기존에 구축된 196개 불확실성 단어들 간의 각 모델 기반 유사성을 확인했다.

3.2 WordNet

WordNet은 1895년 프린스턴 대학(Princeton university)의 심리학과 교수 George A. Miller의 지도하에 인지과학 연구소에서 개발한 영어 어휘 데이터베이스로 영어 단어를 예시나 정의와 함께 신셋으로 불리는 동의어(synonyms) 집합으로 구성한 사전과 시소러스의 조합형태이다(Miller, Beckwith, Fellbaum, Gross and

Miller 1990). 가장 최신버전은 2012년 11월에 배포된 3.1버전으로 총 206,941개 단어 의미 쌍(word-sense pairs)에 대해 117,659개의 신셋으로 구성된 155,287개의 단어를 포함하고 있으며 압축된 크기는 12메가바이트이다. WordNet은 명사, 동사, 형용사, 부사의 어휘 카테고리(syntactic category)가 있으며, 동일한 어휘 카테고리에 속한 단어들은 신셋으로 그룹화되어 총 45개의 lexicographer 파일을 가진다. 모든 신셋은 의미적 관계(semantic relations)에 따라 다른 신셋들과 연결된다. 단어들 간의 관계를 표현하는 포인터(pointers)는 단어의 의미(meaning) 관계를 표현하는 의미적(semantic) 포인터와 단어의 형태(form)로 관계를 표현하는 어휘적(lexical) 포인터가 있다. 예를 들어 명사는 상위어(hypernyms), 하위어(hyponyms), 등위어(coordinate terms) 등을 포함하며, 동사는 상위어, 등위어 등을 포함한다. 각 단어는 단어의 불규칙한 형태를 줄이기 위해 원형복원(lemmatization) 형태로 배포된다. 예를 들어 과거형 동사 wrote는 동사원형인 write로 원형복원된다.

본 연구에서는 불확실성 단어의 동의어를 비롯하여 가장 중요한 관계성으로 알려진(Banerjee and Pedersen 2002) 상위어와 하위어, 유사어에 주목하여 관계성을 살펴보았다. 포인터 유형으로 hypernym, hyponym, similar to를 사용하여 불확실성 단어에 대한 온톨로지 상의 의미적 관련 단어들을 추출했다. 196개 불확실성 단어 리스트에서 모든 단어들끼리 쌍을 구성하여 관계성을 가지는 단어들의 해당 정보를 추출했다. 또한 동의어 관계는 각 불확실성 단어에 대해 신셋을 추출하였다.

3.3 유사도 산출

불확실성 단어 쌍의 유사도는 WordNet 상의 온톨로지에서 두 단어 간의 유사도를 계산하는 대표적인 Lesk(Banerjee and Pedersen 2002)와 WUP(Wu and Palmer 1994) 알고리즘을 적용하여 산출했다. Lesk(1986)는 기존의 단어 의미 중의성 해소(Word Sense Disambiguation, WSD)를 위한 기본적인 알고리즘의 원리를 WordNet 상에 적용하였다. 이는 두 개념을 직접적으로 연결하는 개념과 함께 단어 정의에서 겹치는 단어의 수를 계산하여 가장 높게 계산된 단어를 추출하는 알고리즘이다(Banerjee and Pedersen 2002).

WUP는 WordNet 상의 두 신셋의 깊이(depth)를 두 단어의 가장 가까운(specific) 공통된 조상(ancestor)을 의미하는 LCS(Least Common Subsumer)의 깊이와 함께 고려하여 유사도를 산출한다. 값은 0부터 1사이이며 두 단어의 개념이 같으면 1이 된다.

$$Score(x,y) = \frac{2 * depth(LCS)}{depth(x) + depth(y)}$$

우선 상하관계를 가지는 527개 단어 쌍의 유사도를 각각 산출했다. 상하관계는 불확실성 단어의 상위어와 하위어로 나타난 단어들을 모두 통합하여 두 단어 중 상위 개념을 가지는 단어를 첫 번째 컬럼에 위치시켜 방향성 있는 네트워크를 구성했다. 또한 동의어 관계의 유사도를 산출하기 위해 불확실성 단어와 동의어로 구성된 196개 단어 쌍의 유사도를 각각 산출했다. 네트워크에서 에지의 가중치가 되는 Lesk

알고리즘과 WUP 알고리즘의 유사도 값은 0과 1사이의 값을 가지므로 시각적으로 구별하기 어려운 점을 고려하여 100을 곱하여 가중치의 차이를 주었고 소수점은 모두 반올림하여 정수형으로 만들었다.

Word2Vec은 196개 불확실성 단어들 간의 모든 쌍을 구성하여 유사도를 산출했다. 이전 절에서 기술했듯이 두 모델인 구글 Word2Vec 모델과 PubMed Word2Vec 모델을 개별적으로 적용하여 단어 쌍 간의 유사도를 산출하였다. 총 19,110건의 단어 쌍으로 구성되어 있으며 WordNet과 동일하게 에지의 가중치에 100을 곱하고 소수점을 반올림하여 정수형으로 만들었다. Word2Vec은 코사인 유사도를 기반으로 유사도를 산출하므로 -1부터 1사이의 값을 가진다. -0.04부터 0.04까지의 값은 0으로 대체되며, -0.05부터의 값은 음수 값이므로 시각화에서 제외되었다.

3.4 네트워크 분석

단어 쌍을 네트워크화하여 분석하는 기법은 단어들 간의 관계성을 시각적으로 파악하기 용이하다. 이는 사회 연결망 분석(Social Network Analysis, SNA)에서 개인과 집단 간의 관계를 노드와 노드 간의 연결로 관계성을 표현하는 방법론에서 유래되었다. 사회 연결망 분석 방법 중 중심성(centrality)이란 네트워크 구조상 중요한 역할을 수행하는 노드를 파악하기 위한 지표로 하나의 노드가 네트워크 중심에 위치한 정도를 파악한다. 중심성은 대표적으로 네 가지의 중심성이 존재한다.

우선 연결정도(degree) 중심성은 한 노드에 얼마나 많은 다른 노드들이 관계를 맺고 있는지를 기준으로 그 노드가 중심에 위치하는 정도를 파악한다(Freeman 1978). 연결정도 중심성은 방향성이 있는 그래프에서 해당 노드로 오는 방향의 연결을 의미하는 내향 중심성(in-degree centrality)과 해당 노드에서 나가는 방향의 연결을 의미하는 외향 중심성(out-degree centrality)으로 구분된다.

근접 중심성(closeness centrality)은 인접 중심성으로도 불리우며 각 노드 간의 거리(distance) 또는 근접성(closeness)을 근거로 중심성을 측정하는 방법이다. 연결정도 중심성과 달리 직접적으로 연결된 노드뿐만 아니라 간접적으로 연결된 모든 노드 간의 거리를 합산하여 중심성을 측정하며 방향성이 있는 그래프에서 방향에 따라 내향 근접성(in-closeness)과 외향 근접성(out-closeness)으로 구분된다(Freeman 1978).

매개 중심성(betweenness centrality)은 네트워크 내에서 한 노드가 담당하는 매개자, 중재자 역할의 정도로 중심성을 측정하는 방법이다(Freeman 1978). 한 노드가 다른 노드들 사이의 최단 경로에 위치할수록 그 노드의 매개 중심성이 높아진다. 예를 들어 A, B, C 노드가 있을 때 A와 C는 B를 통해서만 관계를 맺을 수 있을 경우 B는 매개 중심성이 높은 노드이다.

위세 중심성(prestige Centrality)은 연결된 노드의 중요성에 가중치를 준 뒤 노드의 중심성을 측정하는 방법으로 위세가 높은 강자들과 관계가 많을수록 자신의 영향력 또한 높아진다. 이는 보나시치(Bonacich) 권력 지수 또는 보나시치 중심성 지수라고도 불리운다.

또한 네트워크 분석에서 그룹별 특징을 살펴보

기 위한 커뮤니티 감지(community detection) 분석이 있다. 이는 군집을 이루는 커뮤니티의 구성을 확인하는 군집화 방법으로 Blondel, Guillaume, Lambiotte and Lefebvre(2008)의 모듈성(modularity) 알고리즘이 있다. 네트워크 내에서 비슷한 특성을 지니는 노드들을 그룹화하여 시각화하므로 활용도가 높은 분석 방법이다.

196개 불확실성 단어의 특성을 네트워크화하여 시각적으로 파악하기 위해 WordNet 기반의 시각화와 Word2Vec 기반의 시각화를 수행하였다. 시각화를 위한 입력 파일로 .net 파일을 만들었으며 오픈소스 시각화 툴인 Gephi (Bastian, Heymann and Jacomy 2009)를 이용하였다. WordNet 기반의 시각화는 비방향성 동의어 관계 네트워크와 방향성이 있는 상하관계 네트워크를 개별적으로 시각화하여 각 관계에서 나타난 불확실성 단어의 특성을 파악하고자 했다. Word2Vec은 두 상이한 모델을 기반으로 불확실성 단어 간의 관계성 네트워크를 구성했다.

4. 실험 결과 분석

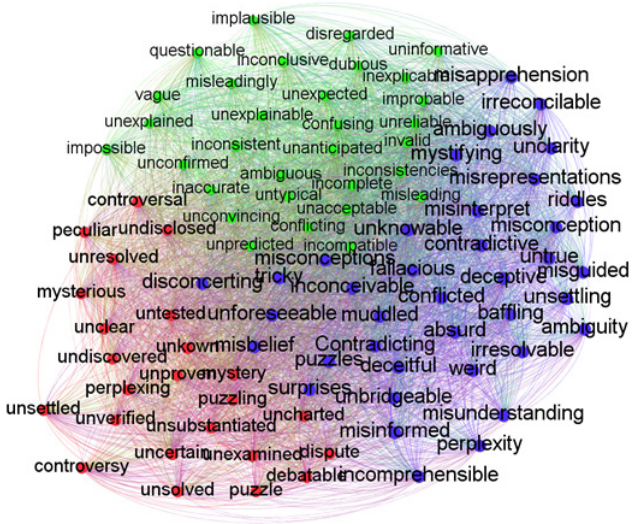
본 장에서는 Word2Vec 기반 네트워크와 WordNet 기반 네트워크를 시각화하여 결과를 분석하였다. 우선 Word2Vec 네트워크 분석에서는 두 가지의 상이한 모델인 구글 Word2Vec 모델과 PubMed Word2Vec 모델을 적용한 불확실성 단어 간의 네트워크를 분석하였으며, WordNet 네트워크 분석에서는 상하관계 네트워크와 동의어 관계 네트워크를 분석하였다. 불확실성 단어 쌍의 유사도는 WordNet상의 온톨로지에서 두 단어 간의 유사도를 계산하는 Lesk와 WUP 알고리즘으로 구분하여 네트워크 결과를 비교 분석하였다.

4.1 Word2Vec 기반 네트워크 분석

〈그림 1〉은 196개 불확실성 단어들의 PubMed Word2Vec 모델을 이용한 네트워크이다. 초기 196개 노드와 19,060개의 에지 중에서 에지 가중치가 음수인 값들을 제외하고 18,942개의 에지로 구성되었다. 연결정도의 범위는 174부터 195까지이며 네트워크의 가시성을 위해 연결정도가 195인 모든 단어와 연결을 유지하고 있는 노드들로 필터를 적용하여 노드 93개(47.45%)와 에지 4,278개(22.58%)로 최종 네트워크를 구성했다. 네트워크 레이아웃은 Fruchterman Reingold 알고리즘을 적용하였다. 노드의 연결정도는 모두 92로 동일하며, 커뮤니티 감지 알고리즘(Blondel, Guillaume, Lambiotte and Lefebvre 2008)을 적용하여 모듈성은 resolution을 1로 설정했을 때 0.067로 3개의 커뮤니티로 구분되었다.

〈표 1〉은 3개 커뮤니티에 대한 결과를 나타낸다. 파랑 색상은 38개의 불확실성 단어로 구성되어 있으며 가장 높은 비율인 40.86%를 차지한 커뮤니티이다. 대표적인 단어들로 misunderstanding, misapprehension, misrepresentations 등의 단어가 속해있다. 또한 32개의 단어로 구성된 초록색상은 전체의 34.41%를 차지하며 빨강 색상은 23개의 단어로 구성되어 24.73%를 차지했다.

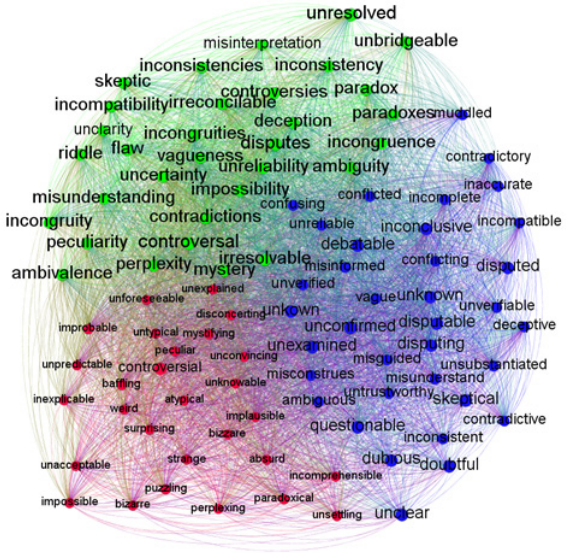
〈그림 2〉는 196개 불확실성 단어들의 구글 Word2Vec 모델을 이용한 네트워크이다. 초기 노드 196개와 에지 18,954개 중 에지 가중치가



〈그림 1〉 PubMed 모델 기반 Word2Vec의 불확실성 단어 네트워크

〈표 1〉 PubMed 모델 기반 Word2Vec 네트워크의 모듈성 알고리즘의 커뮤니티

색상	수	비율	대표적인 단어
파랑	38	40.86%	misunderstanding, irreconcilable, misapprehension, absurd, unsettling, misrepresentations
초록	32	34.41%	confusing, implausible, unconvincing, ambiguous, misleading, questionable
빨강	23	24.73%	puzzling, dispute, perplexing, controversal, mysterious, unresolved



〈그림 2〉 구글 모델 기반 Word2Vec의 불확실성 단어 네트워크

음수인 에지들을 제거하여 18,568개 에지로 구성되었다. 연결정도의 범위는 127~195이며 네트워크의 가시성을 위해 연결정도 192이상인 노드들만을 대상으로 필터링을 적용하여 96개(48.98%) 노드와 4,554개(24.53%)의 에지로 네트워크를 구성했다. PubMed 모델의 네트워크와 동일한 기준인 모든 노드와 연결된 노드만을 대상으로 했을 때 노드 수 9개(4.59%)와 에지 수 36개(0.19%)만 남게 되어 PubMed 네트워크와 유사한 노드 수를 기준으로 맞추었다. 연결정도가 195인 노드는 questionable, unconfirmed, conflicting, unpredictable, inexplicable, inconclusive, baffling, flaw, peculiar 단어이다. 평균 연결정도는 94.875이며, 커뮤니티 감지 알고리즘에서 모듈성은 resolution을 1로 설정하여 0.086으로 3개의 커뮤니티로 구성되었다.

〈표 2〉는 모듈성 기반의 정보와 대표되는 단어를 나타낸다. 〈그림 1〉의 PubMed 모델 기반 Word2Vec의 불확실성 네트워크와 동일하게 비율 기반 색상 정보를 동일하게 구성하였다. 대부분의 불확실성 단어들은 연결정도 중심성이 95이다. 파랑 색상은 36개의 불확실성 단어로 구성되어 있으며 4개 단어인 inaccurate, unsubstantiated, deceptive, unclear는 연결정도 중심성이 94이다. 초록 색상은 32개 불확실성 단어 중 6개 단어인 misunderstanding, riddle, deception, peculiarity, incompatibility, perplexity는 연결정도 중심성이 94이다. 빨강 색상은 28개의 불확실성 단어로 구성되어 있으며 유일하게 improbable 단어가 연결정도 중심성이 93이다.

〈표 3〉은 PubMed 모델과 구글 모델을 적용한 상위 10개 불확실성 단어 쌍 간의 Word2Vec 유사도 값을 나타낸다. PubMed 모델에서는 contradictory와 conflicting 단어가 0.899로 가장 높은 유사도를 가진 단어 쌍이며 상위 10개 단어 쌍 모두 0.8이상의 값이고 실제 상위 26개 단어 쌍이 0.8이상의 값을 가졌다. 상위 10개 단어 쌍 중에서 controversial, uncertain, unclear, unknown, debatable 단어는 2개의 불확실성 단어와 높은 유사도 관계를 보였다.

반면, 구글 모델 기반 불확실성 단어 쌍은 상위 5위의 단어들이 0.8이상의 유사도 값을 가지는 것으로 확인되었으며, 0.7이상의 값을 가지는 단어 쌍은 총 32개 단어 쌍이다. 상위 10개 단어 쌍에서 baffling 단어는 3개 단어인 puzzling, perplexing, mystifying과 높은 유사도 값을 가졌고, 이들 중 puzzling, perplexing은 1순위 유사도 값인 0.832의 값을 가지는 단어 쌍으로 출현하였다. PubMed 모델에 비해 구글 모델은 불확실성 단어의 관계에 대해 상대적으로 낮은 유사도를 보이며 대표되는 단어 출현 빈도수가 적어서 불확실성 단어가 흩어져 있는(sparse) 분포를 띤다는 점을 확인할 수 있었다.

〈표 2〉 구글 모델 기반 Word2Vec 네트워크의 모듈성 알고리즘의 커뮤니티

색상	수	비율	대표적인 단어
파랑	36	37.50%	ambiguous, contradictory, confusing, inconsistent, vague, conflicting
초록	32	33.33%	unresolved, vagueness, inconsistencies, inconsistency, contradictions, skeptic
빨강	28	29.17%	controversial, baffling, puzzling, implausible, absurd, incomprehensible

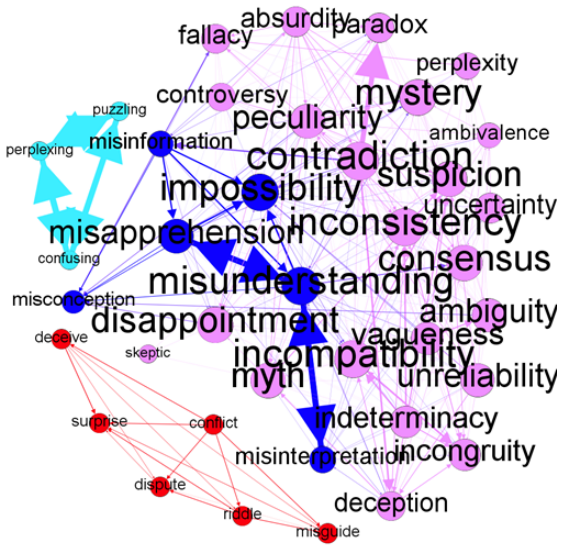
〈표 3〉 Word2Vec 유사도 기반의 상위 10개 불확실성 단어 쌍

순위	PubMed 모델			구글 모델		
	단어	단어	유사도 값	단어	단어	유사도 값
1	contradictory	conflicting	0.899	puzzling	perplexing	0.832
2	controversial	debatable	0.894	weird	strange	0.816
3	uncertain	unclear	0.887	puzzling	baffling	0.816
4	uncertainty	uncertainties	0.863	contradicts	contradicted	0.809
5	controversial	unsettled	0.862	baffling	perplexing	0.807
6	unclear	unknown	0.855	myth	myths	0.797
7	unknown	undefined	0.852	dispute	disputes	0.793
8	unsolved	unresolved	0.849	controversies	controversy	0.792
9	uncertain	debatable	0.849	baffling	mystifying	0.785
10	mysterious	mystery	0.839	contradictory	conflicting	0.785

4.2 WordNet 기반 네트워크 분석

4.2.1 상하관계 네트워크 분석

〈그림 3〉은 상하관계를 가지는 두 단어를 Lesk 알고리즘 기반으로 유사도를 산출한 결과에 대한 시각화 네트워크이다. 각 노드는 불확실성 단어를 의미하고 두 노드를 연결하는 에지는 Lesk 알고리즘의 유사도 값이다. 네트워크의 레이아웃은 Fruchterman Reingold 알고리즘을 적용하였고 가시성을 위해 초기 노드 85개와 초기 에지 339개의 네트워크에서 연결정도를 산출하여 연결정도가 3이상인 노드 38개(47.5%)와 에지 296개(91.93%)로 구성하였다. 에지의 방향성은 부모노드에서 자식노드



〈그림 3〉 Lesk 알고리즘 기반 WordNet의 상하관계 네트워크

로 향한다. 쌍방의 화살표를 가지는 단어들은 해당 단어의 상위어와 하위어로 모두 추출되었으며 이들은 모두 유사어로 분류되어 있었다. 평균 연결정도 중심성 값은 7.789이고, 모듈성은 resolution을 1로 설정하여 0.32이며, 커뮤니티의 개수는 4개로 구분되었다. 〈표 4〉는 각 커뮤니티와 대표적인 단어들의 정보를 나타낸다.

〈표 5〉는 연결정도 중심성이 10 이상인 상위

〈표 4〉 Lesk 알고리즘 기반 네트워크의 모듈성 알고리즘의 커뮤니티

색상	수	비율	대표적인 단어
분홍	23	60.53%	mystery, contradiction, inconsistency, incompatibility, disappointment, peculiarity
파랑	6	15.79%	misunderstanding, impossibility, misapprehension, misinterpretation, misconception
빨강	6	15.79%	deceive, surprise, dispute, riddle, misguide, conflict
하늘	3	7.89%	confusing, perplexing, puzzling

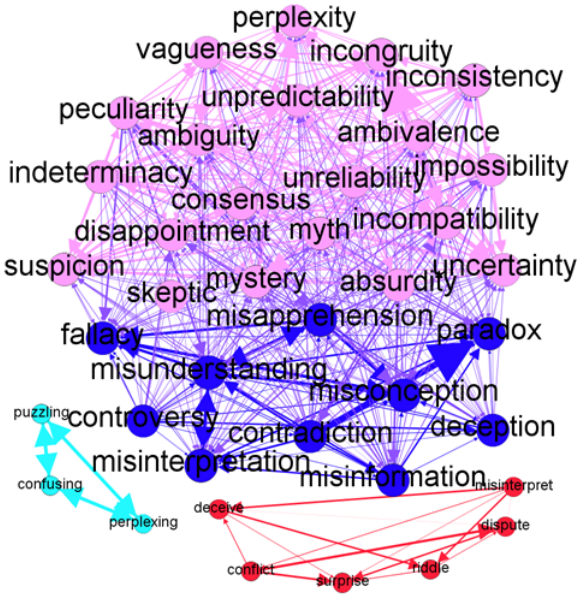
〈표 5〉 연결정도 중심성 값이 10 이상인 불확실성 단어 정보

순위	단어	내향 중심성	외향 중심성	연결정도 중심성
1	contradiction	4	23	27
2	misunderstanding	11	15	26
3	inconsistency	14	12	26
4	incompatibility	17	9	26
5	impossibility	15	10	25
6	consensus	21	3	24
7	disappointment	12	12	24
8	mystery	16	8	24
9	suspicion	18	5	23
10	myth	10	13	23
11	misapprehension	9	13	22
12	peculiarity	4	18	22
13	ambiguity	15	6	21
14	unreliability	0	21	21
15	vagueness	4	16	20
16	indeterminacy	2	17	19
17	incongruity	18	1	19
18	absurdity	1	16	17
19	uncertainty	13	4	17
20	paradox	14	2	16
21	deception	5	11	16
22	fallacy	15	1	16
23	controversy	0	15	15
24	misinterpretation	7	6	13
25	perplexity	12	1	13
26	misinformation	3	9	12
27	ambivalence	4	7	11

27개의 단어와 연결정도 중심성 정보를 나타낸다. 방향성이 있는 네트워크이므로 해당 단어가 화살표 방향을 받을 경우 내향 중심성, 반면 화살표를 방향을 밖으로 향할 경우는 외향 중심성이다. 두 단어의 관계에서 내향 중심성은 상대 노드의 자식 노드, 외향 중심성은 상대 노드의 부모 노드가 된다. 따라서 내향 중심성이 높을 경우 네트워크 내 연결된 관계에서 상대적으로 하위에 속한 단어이며, 외향 중심성이 높을수록 연결된 관계에서 상대적으로 상위에 속한 단어이다. 또한 외향 중심성이 높은 단어는 해당 단어의 의미가 여러 의미로 해석될 수 있는 다의성이 크다는 특징이 존재한다.

내향 중심성의 최댓값은 21로 상위 6번째 단어인 consensus이며 외향 중심성의 최댓값은 23으로 상위 1번째 단어인 contradiction 단어이다. 상위 14번째, 23번째 단어인 unreliability와 controversy는 내향 중심성이 0으로 모든 연결 관계의 방향이 밖으로 향하고 있다. 이를 통해 WordNet 상에서 불확실성 단어 중 상위 계층에 속한 상위어임을 확인할 수 있다. 또한 〈표 5〉에서는 나타나지 않았지만 외향 중심성이 0으로 나타난 단어는 surprise와 skeptic 단어로 내향 중심성이 각각 5, 3으로 출현했다. 이들은 네트워크 내에서 불확실성 단어 중 상대적으로 하위에 속한 단어들이다.

〈그림 4〉는 WUP 알고리즘을 기반으로 두 상하관계 단어의 유사도를 시각화한 네트워크이다. 각 노드는 불확실성 단어를 의미하고 두 노드를 연결하는 에지는 WUP 알고리즘의 유사도 값이다. 가시성을 위해 초기 노드 85개와 초기 에지 486개에서 네트워크 연결정도가 3이상인 노드 39개(46.43%)와 에지 461개(94.86%)로 구성했다. 평균 연결정도 중심성 값은 11.821이



〈그림 4〉 WUP 알고리즘 기반 WordNet의 상하관계 네트워크

고, 모듈성은 resolution을 1로 설정하여 0.129이며 커뮤니티의 개수는 4개로 나타났다. 〈표 6〉은 모듈성 결과를 나타낸다.

빨강으로 구분된 커뮤니티에서 이전 Lesk 알고리즘 기반 네트워크와 비교하여 misguide 단어가 제외되었고 misinterpret 단어가 포함되었다. 하늘색은 이전 네트워크와 동일한 커뮤니티로 그룹화되었고 세 단어는 모두 양방향의 관계를 가지고 있다.

〈표 7〉은 네트워크에 출현한 단어들 중 연결

〈표 6〉 WUP 알고리즘 기반 네트워크의 모듈성 알고리즘의 커뮤니티

색상	수	비율	대표적인 단어
분홍	20	51.28%	mystery, myth, ambiguity, unpredictability, incompatibility, unreliability
파랑	10	25.64%	misunderstanding, misinformation, misapprehension, misinterpretation, misconception
빨강	6	15.38%	deceive, surprise, dispute, riddle, misinterpret, conflict
하늘	3	7.69%	confusing, perplexing, puzzling

〈표 7〉 연결정도 중심성 값이 29 이상인 불확실성 단어 정보

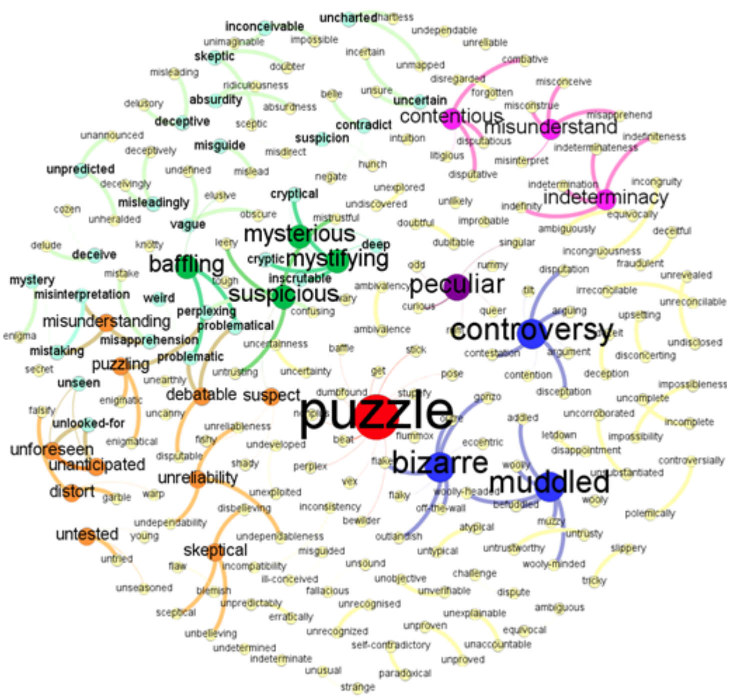
순위	단어	내향 중심성	외향 중심성	연결정도 중심성
1	misunderstanding	12	19	31
2	incompatibility	19	12	31
3	misinterpretation	16	14	30
4	misapprehension	9	21	30
5	misconception	24	6	30
6	fallacy	29	1	30
7	incongruity	27	3	30
8	uncertainty	25	5	30
9	indeterminacy	3	27	30
10	inconsistency	16	14	30
11	mystery	19	10	29
12	perplexity	28	1	29
13	contradiction	4	25	29
14	paradox	24	5	29
15	unpredictability	12	17	29
16	myth	13	16	29
17	vagueness	7	22	29
18	ambiguity	21	8	29
19	peculiarity	5	24	29
20	absurdity	3	26	29
21	misinformation	6	23	29
22	ambivalence	10	19	29
23	suspicion	22	7	29
24	disappointment	14	15	29
25	impossibility	17	12	29
26	unreliability	1	28	29
27	consensus	26	3	29
28	controversy	0	29	29
29	deception	9	20	29
30	skeptic	20	9	29

정도가 29이상인 상위 30개의 단어를 나타낸다. 내향 중심성이 0인 단어는 28번째 단어인 controversy와 <표 7>에 포함되지 않았지만 외향 중심성이 5인 misinterpret가 있다. 반면 외향 중심성이 0인 단어는 surprise로 내향 중심성이 5이다.

Lesk 알고리즘을 기반으로 한 네트워크와 비교하면 WUP를 적용한 네트워크는 상대적으로 에지 수가 많다. 두 동의어 간의 유사도 산출 시 유사성이 0으로 나오는 값들이 더 적게 출현하여 에지 수가 보존되었다. 이는 연결정도 중심성의 값에서도 확인할 수 있다. 즉, Lesk 알고리즘에 비해 WUP 알고리즘이 상하관계에서 유사도를 더 적합하게 산출해내는 특성을 지니는 것으로 판단된다.

4.2.2 동의어 네트워크

<그림 5>는 Lesk 알고리즘을 기반으로 불확실성 단어 기준의 동의어 관계를 시각화한 네트워크이다. 각 노드는 불확실성 단어와 불확실성 단어에 대해 동의어 집합으로 구성된 단어를 의미하고 두 노드를 연결하는 에지는 Lesk 알고리즘의 유사도 값이다. 네트워크 레이아웃은 Fruchterman Reingold 알고리즘을 적용하였다. 초기 254개 노드와 168개의 에지로 구성된 네트워크에서 두 단어의 유사도 값이 0으로 계산된 단독적으로 존재하는 24개의 노드를 제거하여 230개 노드와 168개의 에지로 구성했다. 평균 연결정도 중심성 값은 1.461이다. 네트워크 상의 색상은 연결정도 중심성 값을 기준으로 총 8개로 구분되었다. <표 8>은 해



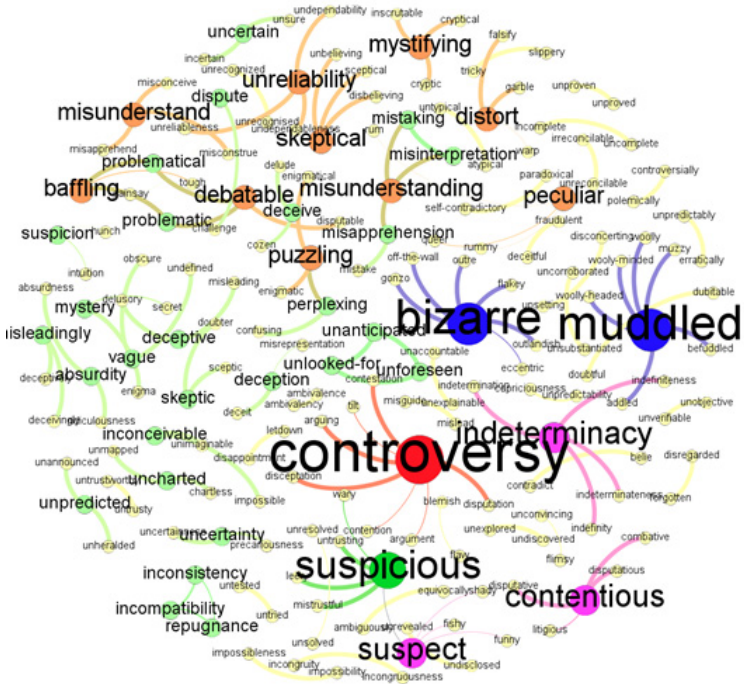
〈그림 5〉 Lesk 알고리즘 기반 동의어 네트워크

〈표 8〉 연결정도 중심성 기반 네트워크 정보

색상	연결정도 중심성	빈도수	비율(%)
노랑	1	181	78.7
연두	2	27	11.74
주황	3	10	4.35
분홍	4	3	1.3
초록	5	4	1.74
보라	6	1	0.43
파랑	7	3	1.3
빨강	12	1	0.43

당 색상에 대한 연결정도 중심성과 네트워크상의 노드 개수와 비율을 나타낸다. puzzle 단어가 연결정도 중심성이 12로 가장 대표적인 노드이다. 하나의 노드 쌍으로만 구성된 노랑색은 총 181개 노드로 전체 노드 중 78.7%를 차지한다.

〈그림 6〉은 WUP 알고리즘 기반의 불확실성 단어 기준 동의어 쌍의 네트워크로 초기 노드 254개와 135개의 에지로 구성된 네트워크에서 60개의 단독으로 존재하는 노드를 제거하여 194개의 노드와 135개의 에지 네트워크를 구성했다. 연결정도 중심성을 기준으로 하는 네트워크 정보는 〈표 9〉와 같다. controversy 단어가 연결



〈그림 6〉 WUP 알고리즘 기반 동의어 네트워크

〈표 9〉 연결정도 중심성 기반 네트워크 정보

색상	연결정도 중심성	빈도수	비율(%)
노랑	1	150	77.32
연두	2	27	13.92
주황	3	10	5.15
분홍	4	3	1.55
초록	5	1	0.52
파랑	6	2	1.03
빨강	7	1	0.52

정도 중심성 7로 가장 대표적인 노드로 나타났다. 연결정도 중심성이 1인 노랑색은 총 150개 노드로 전체 노드 중 77.32%를 차지한다.

두 네트워크의 차이점을 보다 면밀히 분석하기 위해 각 네트워크의 연결정도가 2 이상인 노드들을 대상으로 연결정도를 비교했다. Lesk 알고리즘에서는 49개 단어, WUP 알고리즘에서는 40개 단어가 연결정도 2 이상의 값을 가졌다. 이들을 통합한 결과 총 55개의 고유한 단어로 구성되었다. Lesk 알고리즘 기반 네트워크에 출현하지 않은 단어는 6개이며, WUP 알고리즘 기반 네트워크에 출현하지 않은 단어는 15개이다.

〈표 10〉은 총 55개의 고유한 노드 중 각 네트워크의 연결정도를 통합한 값이 5 이상인 상위 21개 노드의 연결정도 정보를 나타낸다. 이들은 모두 196개의 불확실성 단어에 포함되는 단어들이다.

4번째 단어인 puzzle은 Lesk 알고리즘을 기반으로 한 네트워크에서는 연결정도가 12로 가장 높은 대표 노드였지만 WUP 알고리즘을 기반으로 한 네트워크에서는 출현하지 않았다. 그 이유는 동의어 쌍의 유사도가 모두 0으로 산출되었기 때문이다. 불확실성 단어 puzzle과 동의어 15쌍에 대한 Lesk 알고리즘의 유사도는 〈표 11〉과 같다. WordNet 상에서 puzzle 단어와 동일한 신셋 집합으로 구분되어 있는 단어들이지만 유사도는 전반적으로 낮게 산출되었다. 이 중 beat 단어가 0.25로 가장 높은 유사도를 가지며, mystify, gravel, amaze는 0으로 산출되었다.

또한 〈표 10〉에서 21번째 단어인 mysterious는 Lesk 알고리즘 기반의 네트워크에서 연결정도 중심성이 5이다. WUP 알고리즘에서는 유사도가 0으로 산출되어 출현하지 않았으며, mysterious 단어 쌍의 Lesk 알고리즘 유사도 결과는 〈표 12〉와 같이 deep 단어가 0.13으로 가장 높았고, 나머지 네 단어는 0.06으로 동일한 유사도 값을 가졌다.

종합적으로 두 네트워크를 비교해보면 WUP 네트워크가 Lesk 네트워크에 비해 노드와 에지 수가 적다. 이는 WUP 알고리즘을 적용하였을 때 두 동의어 간의 유사도가 0으로 산출된 결과가 더 많기 때문이다. 이는 앞서 설명한 연결정도 중심성 기반 두 알고리즘의 비교 분석과도 동일한 결과를 보였다. 즉, Lesk 알고리즘이 WordNet 상 동의어 관계의 유사도를 산출할 때 에지의 손실이 적어 관계성을 더 잘 표현하는 것으로 판단된다. 이는 상하관계 네트워크의 결과와는 대비되는 결과이다.

〈표 10〉 두 알고리즘의 연결정도 중심성 기반 상위 단어 정보

순위	단어	연결정도 중심성		총 연결정도 중심성
		Lesk	WUP	
1	controversy	7	7	14
2	bizarre	7	6	13
3	muddled	7	6	13
4	puzzle	12	-	12
5	suspicious	5	5	10
6	peculiar	6	3	9
7	baffling	5	3	8
8	mystifying	5	3	8
9	indeterminacy	4	4	8
10	contentious	4	4	8
11	misunderstand	4	3	7
12	suspect	3	4	7
13	misunderstanding	3	3	6
14	puzzling	3	3	6
15	skeptical	3	3	6
16	distort	3	3	6
17	debatable	3	3	6
18	unreliability	3	3	6
19	unanticipated	3	2	5
20	unforeseen	3	2	5
21	mysterious	5	-	5

〈표 11〉 puzzle 단어 쌍의 Lesk 알고리즘 유사도 값

불확실성 단어	동의어	Lesk 유사도 값
puzzle	beat	0.25
	stick	0.13
	get	0.11
	perplex	0.08
	flummox	0.08
	stupefy	0.08
	nonplus	0.08
	dumbfound	0.08
	bewilder	0.08
	baffle	0.06
	vex	0.04
	pose	0.03
	mystify	0
	gravel	0
	amaze	0

〈표 12〉 mysterious 단어 쌍의 Lesk 알고리즘 유사도 값

불확실성 단어	동의어	Lesk 유사도 값
mysterious	deep	0.13
	cryptic	0.06
	cryptical	0.06
	inscrutable	0.06
	mystifying	0.06

동일한 WordNet의 결과에 다른 두 유사도 알고리즘을 적용하여 네트워크를 시각화했을 때 네트워크의 특성이 상이하게 나타났다. 각 알고리즘은 상하관계 또는 동의어 관계성에 따라 에지를 보존하는 비율이 대비되어 본 연구 데이터에 특정 알고리즘이 관계성 표현에 더 적합하다는 판단을 내리기 어렵지만 두 개 이상의 유사도 알고리즘을 적용하여 네트워크의 특성을 비교 분석함으로써 산출된 결과를 종합적으로 판단할 수 있었다.

지금까지 196개 불확실성 단어들 간의 Word2Vec 기반의 유사도를 산출한 네트워크를 통해 불확실성 단어 간의 연결 관계를 시각적으로 살펴보았다. 또한 196개의 불확실성 단어에 대해 WordNet을 적용하여 불확실성 단어에 대한 상위어, 하위어, 동의어 관계를 시각화했다. WordNet 온톨로지 상의 유사도를 계산하는 Lesk 알고리즘과 WUP 알고리즘을 적용하여 방향성 있는 상하관계 네트워크와 비방향성 동의어 네트워크를 수행하였고 불확실성 단어를 중심으로 단어의 관계성을 시각적으로 확인하였다.

5. 결 론

본 연구에서는 불확실성 단어들을 기반으로 두 층의 인공 신경망으로 이루어진 텍스트 처리 기법인 Word2Vec 기법과 영어 어휘 데이터베이스 사전과 시소러스의 조합 형태인 WordNet을 기반으로 불확실성 단어들 간의 연관성을 종합적으로 분석하고 네트워크를 기반으로 시각화하여 특성을 도출했다. 불확실성 단어들의 WordNet 기반 어휘적 관계성뿐만 아니라 Word2Vec 기반 의미적 관계성을 모두 살펴봄으로써 불확실성 단어들의 특성을 파악하였을 뿐만 아니라 WordNet 상의 동의어를 기반으로 불확실성 단어들을 자동적으로 확장할 수 있는 가능성을 살펴보았다. 특히 Word2Vec 모델 기반의 네트워크 분석에서는 두 가지의 대표적인 모델인 구글 뉴스 모델과 PubMed 학술 문헌 모델을 개별적으로 적용하여 196개의 불확실성 단어들 간의 의미적 유사성을 시각적으로 확인하였다. WordNet 네트워크 분석에서는 WordNet의 온톨로지를 기반으로 단어 쌍의 유사도를 계산하는 대표적인 알고리즘인 Lesk 알고리즘과 WUP 알고리즘을 적용하여 WordNet 기반의 상하관계, 동의어 관계를 살펴봄으로써 단어들의 사전적, 어휘적 특성을 구조적으로 확인하였다.

본 연구는 불확실성 단어들 간의 관계성을 의미적, 구조적으로 파악함으로써 추후 불확실성 단어를 확장할 수 있는 가능성이 있는 연구로

시사점을 지닌다. 특히 동의어 관계에서 기존 불확실성 단어와 빈번히 출현한 단어들을 기반으로 확장 가능성이 있는 후보 단어들을 추출할 수 있다. 이는 기존 선행연구에서 수행된 수작업 기반의 불확실성 단어를 구축하는데 드는 시간과 비용을 감소시키는 방향으로 단어 구축의 자동화에 도움을 제공할 수 있다는데 연구의 의의를 지닌다.

본 연구는 기 구축되어 있는 불확실성 단어 196개를 중심으로 단어들 간의 관계성을 규정지었기 때문에 WordNet 기반 동의어 네트워크에서 확장 가능성이 있는 동의어들 간의 연결성은 포함하지 않았다. 현재 불확실성 단어 리스트에는 포함되지 않았지만 동의어로 정의되어 있는 단어들 간의 관계성을 구조적으로 파악한다면 보다 풍부한 관계성을 파악할 수 있을 것으로 판단된다. 따라서, 향후 연구로는 동의어로 포함된 대표적인 불확실성 단어들을 기반으로 불확실성 단어 리스트를 확장하고자 한다.

참 고 문 헌

[1] 허고은, 송민. 2013. 저자동시인용 분석과 동시출현단어 분석을 이용한 의료정보학 저널의 지적구조 분석. 『정보관리학회지』, 30(2): 207-225

[2] 허고은, 송민. 2019. 생의학 학술 문헌의 불확실성 기반 지식 동향 분석에 관한 연구. 『정보관리학회지』, 36(2): 175-199.

[3] 허고은. 2019. 토픽 모델링 기반 과학적 지식의 불확실성의 흐름에 관한 연구. 『정보관리학회지』, 36(1): 191-213.

[4] Banerjee, S. and Pedersen, T. 2002. "An adapted Lesk algorithm for word sense disambiguation using WordNet." In *International Conference on Intelligent Text Processing and Computational Linguistics*, 136-145. Springer, Berlin, Heidelberg.

[5] Bastian, M., Heymann, S. and Jacomy, M. 2009. "Gephi: an open source software for exploring and manipulating networks." *Icwsm*, 8: 361-362.

[6] Blondel, V. D., Guillaume, J. L., Lambiotte, R. and Lefebvre, E. 2008. "Fast unfolding of communities in large networks." *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

[7] Chen, C., Song, M. and 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

[8] Daim, T. U., Rueda, G., Martin, H. and Gerdsri, P. 2006. "Forecasting emerging technologies:

Use of bibliometrics and patent analysis." *Technological Forecasting and Social Change*, 73(8): 981-1012.

[9] Farkas, R., Vincze, V., Móra, G., Csirik, J. and Szarvas, G. 2010. "The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text." 1-12. Association for Computational Linguistics.

[10] Fernandes, E. R., Crestana, C. E. and Milidiú, R. L. 2010. "Hedge detection using the RelHunter approach." In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, (July): 64-69. Association for Computational Linguistics.

[11] Freeman, L. C. 1978. "Centrality in social networks conceptual clarification." *Social networks*, 1(3): 215-239.

[12] Geaney, F., Scutaru, C., Kelly, C., Glynn, R. W. and Perry, I. J. 2015. "Type 2 diabetes research yield, 1951-2012: bibliometrics analysis and density-equalizing mapping." PloS one, 10(7): e0133009.

[13] Heo, G. E., Kang, K. Y., Song, M. and Lee, J. H. 2017. "Analyzing the field of bioinformatics with the multi-faceted topic modeling technique." *BMC bioinformatics*, 18(7): 251.

[14] Hyland, K. 1996. "Talking to the academy: Forms of hedging in science research articles." *Written communication*, 13(2): 251-281.

[15] Hyland, K. 1998. Hedging in scientific research articles, Vol. 54. John Benjamins Publishing.

[16] Jeong Y. K., Heo, G. E. Kang, K. Y., Yoon, D. S. and Song, M. 2016. "Trajectory analysis of drug-research trends in pancreatic cancer on PubMed and ClinicalTrials." gov. *Journal of Informetrics*, 10(1): 273-285.

[17] Kilicoglu, H. and Bergler, S. 2008. "Recognizing speculative language in biomedical research articles: a linguistically motivated perspective." *BMC bioinformatics*, 9(11): S10.

[18] Kostoff, R. N., del Rio, J. A., Humenik, J. A., Garcia, E. O. and Ramirez, A. M. 2001. Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, 52(13): 1148-1156.

[19] Lesk, M. 1986. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone." In *Proceedings of the 5th annual international conference on Systems documentation*, (pp. 24-26). ACM.

[20] Li, X., Shen, J., Gao, X. and Wang, X. 2010. "Exploiting rich features for detecting hedges and their scope." In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, (July): 78-83. Association for Computational Linguistics.

[21] Light, M., Qiu, X. Y. and Srinivasan, P. 2004. "The language of bioscience: Facts, speculations,

and statements in between." In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, (May): 17-24. Association for Computational Linguistics.

[22] Madani, F. and Weber, C. 2016. "The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis." *World Patent Information*, 46: 32-48.

[23] Malhotra, A., Younesi, E., Gurulingappa, H. and Hofmann-Apitius, M. 2013. "'Hypothesis Finder': a strategy for the detection of speculative statements in scientific text." *PLoS computational biology*, 9(7): e1003117.

[24] Medlock, B. and Briscoe, T. 2007. "Weakly supervised learning for hedge classification in scientific literature." In *ACL*, (June): 992-999.

[25] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. 2013. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, (pp. 3111-3119).

[26] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. 1990. "Introduction to WordNet: An on-line lexical database." *International journal of lexicography*, 3(4): 235-244.

[27] Palmer, F. R. 2014. Modality and the English modals. Routledge.

[28] Pyysalo, S., Ginter, F., Moen, H., Salakoski, T. and Ananiadou, S. 2013. "Distributional semantics resources for biomedical text processing." In: LBM. Tokyo: Database Center for Life Science.

[29] Rei, M. and Briscoe, T. 2010. "Combining manual rules and supervised learning for hedge cue and scope detection." In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, (July): 56-63. Association for Computational Linguistics.

[30] Sánchez, L. M., Li, B. and Vogel, C. 2010. "Exploiting CCG structures with tree kernels for speculation detection." In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, (July): 126-131. Association for Computational Linguistics.

[31] Song, M., Heo, G. E. and Kim, S. Y. 2014. "Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in DBLP." *Scientometrics*, 101(1): 397-428.

[32] Song, M., Heo, G. E. and Lee, D. H. 2014. "Identifying the Landscape of Alzheimer's Disease Research with Network and Content Analysis." *Scientometrics*, 102(1): 905-927.

[33] Szarvas, G. 2008. "Hedge classification in biomedical texts with a weakly supervised selection

of keywords." *Proceedings of ACL-08: HLT*, 281-289.

[34] Szarvas, G., Vincze, V., Farkas, R. and Csirik, J. 2008. "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*, (pp. 38-45). Association for Computational Linguistics.

[35] Szarvas, G., Vincze, V., Farkas, R., Móra, G. and 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

[36] Tang, B., Wang, X., Wang, X., Yuan, B. and Fan, S. 2010. "A cascade method for detecting hedges and their scope in natural language text." In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, (July): 13-17. Association for Computational Linguistics.

[37] Thompson, P., Nawaz, R., McNaught, J. and Ananiadou, S. 2011. "Enriching a biomedical event corpus with meta-knowledge annotation." *BMC bioinformatics*, 12(1): 393.

[38] 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.

[39] Vincze, V., Szarvas, G., Farkas, R., Móra, G. and 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

[40] Wu, Z. and Palmer, M. 1994. "Verbs semantics and lexical selection." In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, (June): 133-138. Association for Computational Linguistics.

[41] Zhang, S., Zhao, H., Zhou, G. and Lu, B. L. 2010. "Hedge detection and scope finding by sequence labeling with normalized feature selection." In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, (July): 92-99. Association for Computational Linguistics.

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

[1] Heo, G. E. and Song, M. 2013. "Examining the Intellectual Structure of a Medical Informatics Journal with Author Co-citation Analysis and Co-word Analysis." *Journal of the Korean*

Society for Information Management, 30(2): 207-225.

[2] Heo, G. E. and Song, M. 2019. "Knowledge Trend Analysis of Uncertainty in Biomedical Scientific Literature." *Journal of the Korean Society for Information Management*, 36(2): 175-199.

[3] 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.