2022년 2월
박사학위논문

개념적 은유 후보 선정을 위한 근원-목표영역 벡터 관계 모델

조선대학교 대학원

컴퓨터공학과

김 형 주

개념적 은유 후보 선정을 위한 근원-목표영역 벡터 관계 모델

Source-Target Domain Vector Relationship Model for Conceptual Metaphor Seed Selection

2022년 2월 25일

조선대학교 대학원

컴퓨터공학과

김 형 주

개념적 은유 후보 선정을 위한 근원-목표영역 벡터 관계 모델

지도교수 김 판 구

이 논문을 컴퓨터공학 박사학위신청 논문으로 제출함

2021년 10월

조선대학교 대학원

컴퓨터공학과

김 형 주

김형주의 박사학위논문을 인준함

심사위원장	조선대학교 교수	양 희 덕	(인)
심사위원	조선대학교 교수	최 준 호	(인)
심사위원	UST 교수	황 명 권	(인)
심사위원	가천대학교 교수	최 창	(인)
심사위원	조선대학교 교수	김 판 구	(인)

2022년 1월

조선대학교 대학원

목 차

표 목 차

그 림 목 차

ABSTRACT

Source-Target Domain Vector Relationship Model for Conceptual Metaphor Seed Selection

Hyoung Ju Kim
Advisor: Pan Koo Kim, Ph.D.
Department of Computer Engineering
Graduate School of Chosun University

This study aims to express mapping through vectorization from one domain (source domain) to another (target domain) on the basis of conceptual metaphoric theory, and to express semantic similarities between words and concepts between words in each area. A model capable of selecting conceptual metapho r candidates was proposed through extracting metaphors from text corpus, semantic vocabulary relationship analysis, sentence structure analysis and patter n extraction, analysis of extracted data, and performance evaluation.

Conceptual candidates were extracted based on the Top2Vec model among the topic modeling methods, and semantic relations based on conceptual metaphor patterns were defined and modeled for extracting concepts and relationships of source and target areas of conceptual metaphor. In addition, in order to select conceptual candidates for the source and target regions, the expanded pattern of Hearst Pattern for recognition of upper words was defined, and upper word candidates were extracted using the conceptual candidate and semantic relationship extraction method.

In the proposed model, the process of extracting the concepts of the source area and the target area is as follows.

First, by applying the Top2Vec model, one of the topic models, related vocabulary similar in meaning to the core concept vocabulary of words in the document was extracted, and second, high-level vocabulary was selected centering on the core conceptual vocabulary. Third, the 'is-A' relationship was extracted based on the upper-level candidate vocabulary, and fourth, the conceptual metaphor in the document was selected based on the extracted source area and target area, and a quantitative comparative evaluation was performed.

The Word2Vec embedding model was used together to evaluate the performance of the source area and target area concept extraction method for the conceptual metaphor proposed in this study. After preprocessing CNN News Corp., a source-target area concept extraction experiment was conducted using the Top2Vec model, and concept extraction performance was compared through three performance indicators: accuracy, precision, reproduction rate, and F-measure.

As a result of this study, it was confirmed that the Top2Vec model used in this paper performed better performance than the model applied Word2Vec when simultaneously extracting the concept of the target area, the concept of the source area, and the conceptual metaphor.

Ⅰ. 서 론

A. 연구 배경

개념적 은유는 언어로 표현되는 특정 은유 표현 집단들 속에 있는 핵심 개념을 추출하여 체계화시킨 것으로서 개념적 은유는 다른 은유법과 차이가 있다. 개념적 은유는 구체적이고 친숙한 경험에 기반을 둔 근원(Source) 영역을 이용하여 추상적이고 모호한 목표 영역을 명확하게 개념화하는 사고방식이며, 은유적 언어 표현이라는 것은 개념적 은유가 구체적인 언어로 표현되는 것이다. 개념적 은유란 우리에게 익숙한 근원(Source) 영역으로써 낯선 목표(Target) 영역을 개념화하는 인지 전략이다. 개념적 은유는 하나의 개념영역을 들어 다른 개념영역을 이해(표현)하는 것이다. 예를 들어 우리가 인생이나 사랑을 여행으로 이해하거나 논쟁을 전쟁으로 이해할 때 인생, 사랑, 논쟁은 표현하려는 영역으로서 '목표영역(Target domain)'이라 하고, 여행, 전쟁은 목표영역을 이해하기 위해 수단이 되는 영역으로서 '근원 영역(Source domain)'이다.

기존 은유 연구는 전통적인 자연어처리 기법으로 품사 패턴, 통계적 분석을 기반으로 연구되었다. 개념적 은유는 어떤 하나의 어휘를 중심으로 찾거나 하나의 품사 혹은 하나의 통사적 패턴으로 명확하게 추출하는데 어려움이 있다. 개념적 은유 인식을 위해 문장의 의미적 문맥 정보 분석이 필요하며 의미적 어휘 관계를 분석하고 문장 구조 분석 및 은유 문장 패턴 추출 필요하다.

문맥 정보 혹은 구조(패턴)를 찾아서 은유를 추출하는 일은 매우 의미 있는 일이며, Deignan (2005, 2006), Hilpert (2006), Choi(2016a)와 같은 학자들은 코퍼스 기반의 매우 제한된 규모의 은유적 표현만 추출하는 연구를 현재까지 하고 있다. 그러나 다양한 웹데이터 내에서 체계적으로 은유가 가진 구조적 패턴을 밝힌 바는 없다. Dependency Parsing 등 자연어처리의 의미 분석을 통해 문맥 정보를

추출하고 그 안에서 개념적 은유의 구조적 패턴을 체계적으로 밝히는 연구가 필요하다. 또한, 근원 영역과 목표 영역을 구분하여 매핑하는 개념적 은유 자동화 연구도 미흡하다. 개념적 은유 인식 연구는 우리 인식에 중요한 역할을 하기에, 자연어처리와 인공지능 분야에서 은유 연구가 필수이다.

본 연구에서 제안하는 개념적 은유 후보 선정을 위한 근원-목표 영역 벡터 관계 모델은 개념적 은유 이론의 기초 위에, 하나의 도메인(근원 도메인)에서 다른 도메인(목표 도메인)으로의 벡터화를 통해 매핑을 표현하고 단어 간의 의미적 유사성과 각 영역의 단어 간의 개념적 관계를 표현하고자 한다. 은유 데이터베이스 생성, 저장, 말뭉치 텍스트에서 은유를 추출하고 의미적인 어휘 관계를 분석하고 문장의 구조 분석 및 패턴을 추출하며, 추출된 데이터의 분석 및 성능평가를 통해 개념적 은유 후보 선정이 가능해지도록 하고자 한다.

B. 연구 내용 및 구성

본 연구는 개념적 은유 후보 선정을 위한 근원-목표 영역 벡터 관계 모델은 개념적 은유 이론의 기초 위에, 하나의 도메인(근원 도메인)에서 다른 도메인(목표 도메인)으로의 벡터화를 통해 매핑을 표현하고 단어 간의 의미적 유사성과 각 영역의 단어 간의 개념적 관계를 표현하고자 한다. 텍스트 코퍼스에서 은유를 추출하고, 의미적인 어휘 관계를 분석하고 문장의 구조 분석 및 패턴을 추출하며, 추출된 데이터의 분석 및 성능평가를 통해 개념적 은유 후보 선정이 가능한 모델을 제안한다.

본 논문의 주요 내용 및 구성은 다음과 같다.

제1장은 서론으로 본 연구의 배경 및 목적에 관해 설명하고 연구 내용 및 구성에 관한 내용을 설명한다.

제2장에서는 본 연구의 이론적 배경이 되는 개념적 은유에 대한 정의, 개념적 은유 사상과 개념적 은유의 유형에 대해 알아보고, 본 연구 수행에 필요한 토픽 모델링에 대한 개념 및 특징을 기술해 연구 내용의 이해를 돕는다.

제3장에서는 개념적 은유 인식을 위한 프레임워크 및 개념적 은유 후보 선정을 위한 근원-목표 영역 개념 추출 방법 및 개념적 은유 인식을 위한 관계 추출 방법을 제시한다.

제4장에서는 개념적 은유의 근원-목표 영역 분류 및 분석으로 CNN 뉴스데이터에서 사용된 개념적 은유 표현들을 중심으로 개념적 은유 근원-목표 영역을 분석하고 그 분석에 사용된 모델에 대한 정량적 비교평가를 수행한다.

제5장에서는 본 논문에서 제안한 방법론의 실험 데이터를 정확도와 재현률, F-measure 평과 결과에 관해 기술하고, 결론과 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

A. 개념적 은유(Conceptual Metaphor)

은유는 동서고금을 막론하고 우리의 일상 언어 속에 체계적이면서 일관성 있게 존재한다고 주장한다. 즉, 은유는 다양하면서도 공유된 신체적, 사회적, 문화적 경험을 토대로 우리 인간의 무의식(unconscious) 속에서 개념화되어 언어로 체계적으로 명시되거나 발현된다는 것인데 이를 두고 개념적 은유라 칭하였다. 그는 개념적 은유 이론을 빌려 인간의 기본 감정인 화, 사랑, 기쁨, 슬픔, 두려움 등의 추상적인 개념을 이론적으로 설명하려 노력하기도 하였다[1].

Lakoff & Johnson(1980/2006)은 “은유의 본질은 한 종류의 사물을 다른 종류의 사물의 관점에서(in terms of) 이해하고 경험하는 것이며, 은유는 언어의 문제, 즉 낱말들의 문제가 아니라 오히려 인간의 사고 과정을 구성하고 개념체계를 규정하는 것”이라고 했다[2,3].

은유란 전통적으로 시(詩)나 수사에 적합한 장식적 표현으로 단순한 언어 현상으로 간주해온 것이다. 이런 은유가 다른 관점으로 다루어진 것은 『삶으로서의 은유(1980)』에서 시작되었다. 은유가 단순히 언어뿐만 아니라 인간의 사고와 행위에 아주 중심적 역할을 하고 있다고 주장하였다[2]. 은유의 전통적인 관점의 견해와 현대 인지언어학적 관점의 개념적 은유 이론의 차이점을 정리하면 표 1과 같다[4].

표 1. 은유의 전통적 견해와 개념적 은유 이론

전통적 견해	개념적 은유 이론
• 은유는 언어적 현상이다.	• 은유는 개념적 속성이다.
• 은유는 미적·수사적 목적을 달성하기 위해 사용한다.	• 은유는 예술적·미적 목적뿐 아니라 어떤 개념을 더 잘 이해하기 위한 것이다.
• 은유는 비교되고 동일시되는 두 개체 사이의 닮음에 기초한다.	• 은유는 종종 유사성에 기초하지 않는다.
• 은유란 낱말의 의식적·고의적 사용이며, 특별한 능력을 지녀야 잘 사용할 수 있다.	• 은유는 평범한 사람들도 일상생활에서 별다른 노력 없이 사용할 수 있다.
• 은유란 없어도 살 수 있는 비유적 표현이다.	• 은유는 인간의 사고와 추론이 불가피한 과정이다.

이 표 1과 같이 은유의 전통적 견해와 인지언어학적 관점에서의 개념적 은유의 속성은 큰 차이를 보인다. 개념적 은유 이론에서 주목할 만한 점은 은유의 본질을 언어 자체로 보는 것이 아니라, 인간사고 능력의 문제로 본다는 것에 있다. 즉 우리의 사고나 개념 자체가 본질에 있어서 은유적이라는 것이다[4].

Lakeoff(1994)는 이를 두고 은유는 단순 언어 표현의 차원이 아니라 “한 정신적 영역을 다른 정신적 영역에 의해서 개념화하는 방식”이라 말한다[5]. 인지언어학에서는 이러한 개념화 과정에 바탕을 둔 은유를 ‘개념적 은유’라고 정의하게 된다. 즉, 추상적인 개념의 사물이나 대상을 또 다른 구체적인 개념에 비유하여 표현하는 것이라 이해한다는 것이다. 다시 말해 “인간의 은유는 우리 인간의 경험적 기초에 의해 인간과 세계 사이의 ‘일상적인 상호작용’의 본질에 둔다”는 것이다[6].

Lakoff & Johnson(1980)은 이 은유를 하나의 경험 영역에서 다른 하나의 경험 영역으로의 체계적인 ‘인지적 사상’(cognitive mapping)으로 규정하고 이를 ‘개념적 은유’(conceptual metaphor)라 지칭하였다[7]. 이 개념적 은유는 「A는 B」로 표시할 수 있는데, 여기서 A는 개념적 은유의 목표영역으로, 이는 우리가 이해하고자 하는 추상적인 개념을 가리키고, B는 개념적 은유의 근원 영역으로, 이는 목표영

역의 추상적인 개념을 이해하기 위해 동원되는 구체적인 개념을 가리킨다[7].

Kövecses(2002)에 의하면 이 근원영역과 목표영역이 가지는 유사성이 서로 체계적으로 일대일 대응되어 사상(mapping)이 이루어 질 때 비로서 'A는 B이다'라고 대응되는 관계에서 이해할 수 있다고 하였다[1]. 따라서 개념적 은유의 기본 속성을 정리해보면 표 2와 같다.

표 2. 근원 영역과 목표 영역의 속성

근원 영역(source domain)	목표 영역(target domain)
• 우리의 일상 경험으로부터 나온 것으로 대체로 구체적이고 명확하게 윤곽이 주어며 직접 경험하고 지각할 수 있는 개념	• 표현하려는 영역으로 더 추상적이고 주관적인 그리고 심리적인 경험과 관련된 개념들로써 그 윤곽이 불명확하고 구조화되지 않은 경험

표 2에서 보듯이, 목표 영역은 구체적인 근원 영역을 사용하여 우리가 이해하고자 하는 영역이며, 근원 영역은 언어적 은유에서 추상적인 목표영역을 이해하기 위해 구체적인 단어나 언어 표현을 빌려오는 영역이다. 이 근원 영역과 목표 영역의 사상관계를 도식화하여 나타내면 그림 1과 같다.

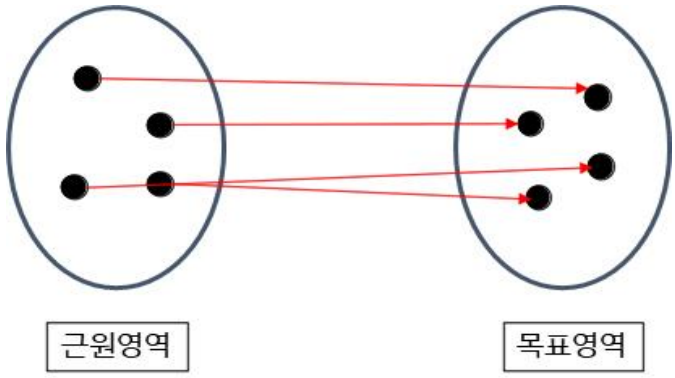


그림 1. 근원 영역과 목표영역 간의 사상

개념적 은유는 우리가 일상 언어생활에서 사용하는 '은유적 언어표현(metaphorical linguistic expressions)'과 일정한 차이가 있다. 은유적 언어 표현은 더 구체적인 개념영역에 대한 언어나 용어로 이루어진 어휘 또는 여타의 표현을 말하며, 개념적 은유는 모든 은유적 표현들의 개념적 기초이다. 개념적 은유는 구체적이고 친숙한 경험을 바탕으로 하는 근원영역으로서 추상적이고 모호하며 새롭게 다가가는 목표영역을 명확히 개념화하는 사고의 방식이며, 은유적 언어 표현과 구별한다[8]. 아래의 표 3을 통해 살펴보자.

표 3. 개념적 은유와 은유적 언어 표현 예

개념적 은유	은유적 표현	해 석
ARGUMENT IS WAR (논쟁은 전쟁이다)	a. Your claims are indefensible.	• "ARGUMENT IS WAR(논쟁은 전쟁이다)"는 사고하는 방식을 나타내는 '개념적 은유 • (a ~ h)는 개념적 은유가 언어로 구체화된 '은유적 언어표현들' • Lakoff & Johnson(1980)에 의하면 이런 전쟁의 관점에서 논쟁을 묘사하는 것은 논쟁의 전투적 측면을 더 부각시키려 함 • 비록 논쟁 또는 전쟁이라는 개념이 소유한 '협동적인 측면'등과 같은 다른 요소들을 배제시켜 나의 주장을 강하게 설득함으로써 본연의 목적에 상응하는 결과를 도출해 낼 수 있다는 것
	b. He attacked every weak point in my argument.	
	c. His criticisms were right on target.	
	d. I demolished his argument.	
	e. I've never won an argument with him.	
	f. You disagree? Okay, shoot!	
	g. If you use that strategy, he'll wipe you out.	
	h. He shot down all of my argument.	
ARGUMENT IS BOULDING (논쟁은 건물이다)	a. She constructed a solid argument.	• (a-d)는 추상적인 개념인 '논쟁'을 마치 건물의 '토대(foundation)'와 '골격(framework)'에 철저희 서로 일 대 일 대응 관계로 형성이 되어 구체적인 개념들이 우리 머릿속에 사상(mapping)됨으로써 목표영역이 이해 가능하다는 것을 보여줌.
	b. He have got a good foundation for the argument.	
	c. She laid the foundation for het argument.	
	d. His argument collapsed.	

TIME IS MONEY (시간은 돈이다)	a. He spent the weekend profitably.	• 기본적을 돈을 중심으로 하는 자본주의적 가치관에 바탕을 둠. • (a)에서는 시간의 사용에 대한 가치를 부여 • (b), (c)는 시간을 철저히 '물질적인 개념'으로 인식하여 조금이라도 낭비해서는 안된다는 중요성의 의미를 부각시키기 위해 사용된 예 • 이처럼 시간을 돈이나 어떤 물질을 계량화시켜 목적 없이 시간을 낭비하는 행위가 부정적인 행위임을 부각시켜려는 의도
	b. Don't waste your time.	
	c. You need to budget your time for your trip to Europe.	

[출처 (Lakoff & Johnson 1980: 4)]

B. 토픽 모델링(Topic Modeling)

토픽 모델링은 확률적 문서 군집화(probabilistic document clustering)의 대표적인 기법의 하나로 비정형화된 방대한 텍스트 문서로부터 잠재된 주제를 추출하여 숨은 의미 구조를 발견하기 위한 확률 모델이다[9].

토픽 모델링에서 하나의 문서는 토픽들의 혼합체이고, 토픽들은 해당 토픽을 구성하는 단어들의 확률 분포로 표시할 수 있다. 토픽 모델링을 이용해 문서 내에 잠재되어있는 토픽을 반영하는 단어 분포를 찾아내어 토픽에 따라 문서를 군집화하는 것이 가능하다. 맥락과 관련된 단서들을 이용해 유사한 의미를 가진 단어들을 클러스터링하는 방식으로 주제를 추론하는 모델[16]이며, 주제들이 서로 어떠한 연관이 있는지, 시간 흐름에 따라 어떻게 변화되는지 알 수 있는 모델이다.

토픽 모델링은 의미를 단어의 관계로 규정하고, 단어들이 소속된 군집을 통해 의미를 찾고자 한다. 이때 군집은 순서가 없는 단어의 자루이며 따라서 단어가 동시에 등장하는 빈도가 주제를 결정하는 요건이 된다. 이러한 방식으로 의미를 찾는다는 특성으로 인하여 토픽 모델링 기법은 동음이의어, 다의어와 방언 등의 문제를 해소해준다는 장점을 지닌다[10]. 또한 텍스트에 드러나는 여러 주제를 기계적으로 분석하여 전체 내용 중에 나타나는 토픽을 추출할 수 있다는 점에서 다른 연구자에 의해 재현 가능성이 크고 사전 지식에 크게 의존하지 않고 의미의 관계성을 추출할 수 있다는 큰 장점이 있다[11].

김남규 외(2017) 연구에 따르면 토픽 모델링은 방대한 양의 문서를 주제에 대응하는 문서로 군집화하여 제공한다는 측면에서는 문서 군집화(Document Clustering)와 유사하지만, 전통적인 경성 군집화(Hard Clustering)와 달리 하나의 문서가 여러 토픽에 동시에 대응될 수 있다는 점에서 현실 세계의 모델링에 보다 적합한 것으로 평가받고 있다[10].

토픽 모델링의 기법으로는 단어의 잠재된(Latent) 의미를 이끌어내는 LSA(Latent Semantic Analysis)와 확률적 잠재 의미 분석인 pLSA(probabilistic Latent Semantic Analysis), 토픽의 단어 분포와 문서의 토픽 분포를 결합한 LDA(Late nt Dirichlet Allocation)가 있다[9]. LDA 기법은 일반적으로 알고리즘이 단순하고 자료의 축소가 쉬우며, 일관성 있는 주제를 생산할 때 유용하다. 다량의 문서의 주제를 찾는 연구에서 많이 사용되고 있다. 또한, 방대한 자료에서 해석 가능성이 큰 토픽들을 추출해주기 때문에 시간의 흐름에 따른 주제의 변화를 살펴볼 수 있으며 그 분야의 연구동향을 알 수 있다는 장점이 있다[12]. 최근까지도 많이 사용되고 있음에도 불구하고 몇 가지 약점을 가지고 있다. 예를 들어 종종 알려진 주제의 수와 텍스트 전처리 단계가 필요하다. 더 나은 일관성과 혼란을 위해 모델의 하이퍼파라미터가 아니며 이상적으로는 알려진 주제의 수를 가정한다. 또한, 불용어의 제거가 필요하다. 그렇지 않으면 주제-단어 분포가 불용어로 오염될 가능성이 높다. 그리고 의미를 무시하고 문서의 BOW(bag of word) 표현에서 작동하며, lemmetization, 형태소 분석과 같은 사전 처리 단계는 도움이 될 수 있지만 측면과 가장자리가 동일 하다는 것을 이해하는 데 도움이 될 수는 없다.

토픽 벡터를 찾기 위해 공동 문서와 단어 의미론적 임베딩을 활용할 수 있는 모델이 Top2Vec이다. 이 모델은 불용어 목록, 형태소 분석 또는 표제어가 필요하지 않으며 자동으로 주제 수를 찾으며, 그 결과 토픽 벡터는 의미론적 유사성을 나타내는 거리와 함께 문서 및 단어 벡터와 함께 포함됩니다. 우리의 실험은 Top2Vec모델이 확률적 생성 모델보다 훨씬 더 유익하고 훈련된 코퍼스를 대표하는 주제를 찾는 것을 보여준다.

Dimo Angelov에 의해 2020년 3월에 개발된 Top2Vec모델은 2020년 8월에 ArXiv에 논문으로 발표되었다. 텍스트에 있는 주제를 자동으로 감지하는 토픽 모델링 및 의미 검색을 위한 알고리즘으로 좋은 토픽 벡터를 학습하기 위해 불용어 제거, 형식화, 형태소 분석 및 토픽 수에 대한 사전 지식이 필요하지 않다. Top2Vec모델에서 사용하는 알고리즘은 Doc2Vec, UMAP, HDBSCAN으로 잘 확립되어 있으며 Top2Vec모델은 Universal Sentence Encoder 및 BERT와 같은 임베딩 모델의 사용을 지원한다. Top2Vec모델은 의미적으로 유사한 많은 문서가

기본 주제를 표시한다는 가정을 기반으로 작동한다.

Top2Vec모델은 시맨틱 공간에서 토픽을 지속적으로 표현하기 때문에 토픽의 수를 원하는 수로 줄일 수도 있으며, 다. 이것은 가장 작은 토픽의 토픽 벡터와 가장 가까운 토픽 벡터의 가중 산술 평균을 취함으로써 수행되며, 각각 토픽 크기에 따라 가중된다. 각각 병합 된 후 토픽 크기(주제가 속한 문서 수를 의미)는 각 토픽에 대해 다시 계산된다. LDA와 같은 전통적인 토픽 모델링 방법에 비해 몇 가지 장점이 있습니다. 예를 들어 불용어 목록이나 lemmatization/stemming이 필요하지 않고, 토픽 수를 자동으로 찾을 수 있으며, 토픽 수에 대한 사전 지식이 필요하지 않으므로 구현이 쉽고 빠르다. 또한, 짧은 텍스트에서도 실행이 가능하며, 함께 포함된 주제, 문서 및 단어 벡터를 만들고 검색 기능이 내장되어 있다. 그리하여 본 논문에서는 근원-목표 영역 개념 추출을 위해 Top2Vec 모델을 사용하였다.

본 논문에서는 Doc2Vec 임베딩 모델을 사용하여 Top2Vec 알고리즘을 실행하였다. Doc2Vec은 문장, 단락, 문서와 같은 가변 길이의 텍스트(이하 문서로 표기)로부터 고정 길이의 벡터 표현을 학습하는 비지도 학습 알고리즘이다[13]. 알고리즘은 각각의 문서를 하나의 고유한(unique) 벡터로 표현하며, 이 벡터는 문서 내 단어들을 예측하기 위해 훈련된다. 인공 신경망을 통해 훈련되며, 훈련을 마친 후 인공 신경망의 가중치(weight)들을 문서의 벡터 표현으로 사용하게 된다. 훈련 시 사용되는 네트워크는 하나의 은닉층을 가지는 얕은 인공 신경망으로 오류 역전파 알고리즘에 의해서 훈련된다[14].

Doc2Vec은 Word2Vec을 확장한 알고리즘으로 문서에 포함된 토큰의 개수와 상없이 문서 자체를 문서의 의미가 반영된 유사도에 기반하여 고정된 크기의 하나의 벡터로 표현하는 알고리즘으로 기사, 웹 문서, 댓글 분석 등에 활용되고 있다[15].

Doc2Vec 이전의 텍스트에 대한 가장 일반적인 고정 길이 벡터 표현은 Bag-of-Words(BOW)이다. Doc2Vec은 단어의 순서와 의미를 무시하는 BOW의 단점을

극복하고, 단어의 순서와 의미를 내포한 벡터 표현을 생성해낸다[13]. 즉, 문서에 포함된 어휘는 동일하나 순서가 다르면 다른 벡터를 생성하고, 의미가 유사한 문서들은 벡터 공간상에 가까이 위치하도록 벡터를 생성하게 되는 것이다[16].

이렇게 생성된 벡터를 이용하여 문서의 분류에 적용 시 Bag-of-Words모델을 적용했을 때보다 성능이 향상됨은 이미 실험을 통해 증명되었다[16]. Doc2Vec에는 Word2Vec의 CBOW 모델과 SG 모델에 대응하는 DM(distri butedmemory, 이하 DM으로 표기)모델과 DBOW(distributedbag ofwords, 이하 DBOW로 표기) 모델이 있다[14]. 이 두 모델은 모두 문서에 포함된 단어들을 예측하는 모델이다. 그림 2는 두 가지 모델에 대한 구조를 나타낸 것이다[17].

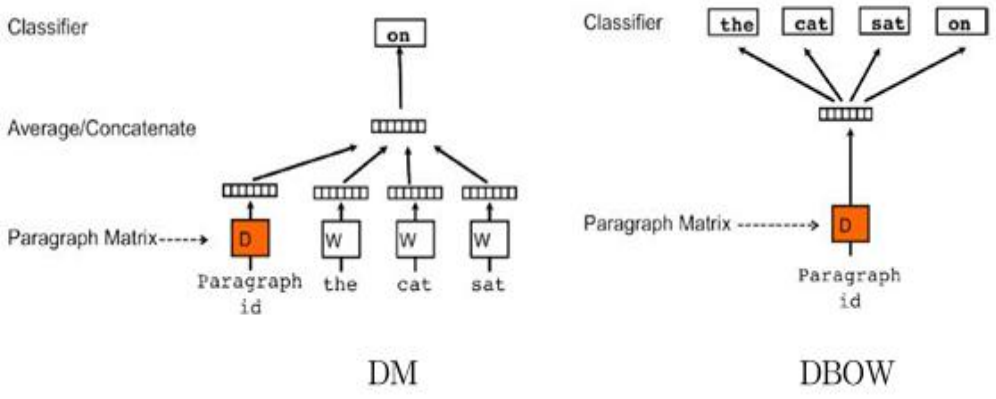


그림 2. Doc2Vec의 두 가지 모델

Ⅲ. 개념적 은유의 근원-목표 영역 벡터 관계 모델

A. 근원-목표 영역 벡터 관계 모델 프레임워크

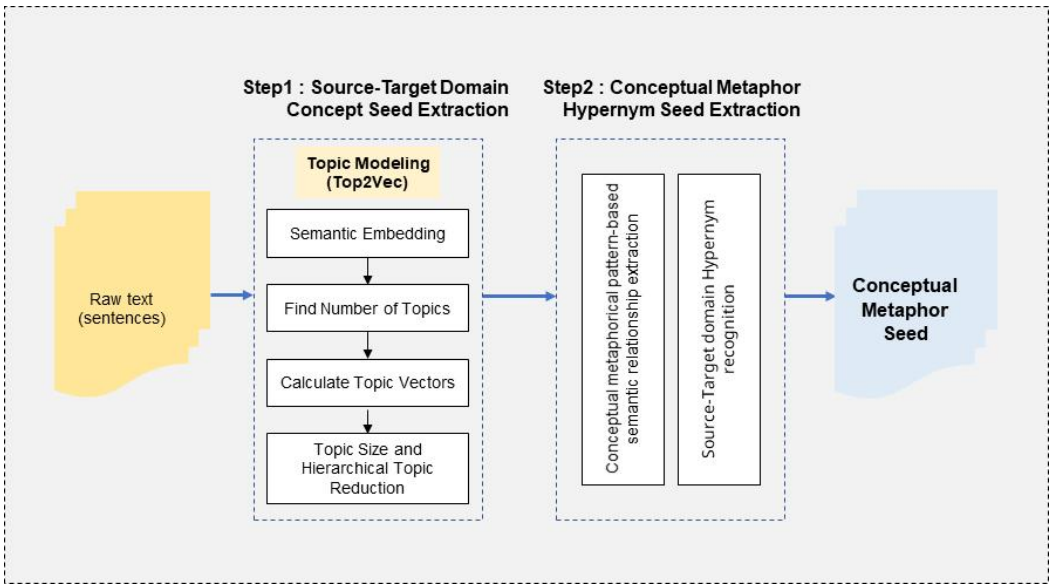


그림 3. 전체 구성도

그림 3은 개념적 은유 후보 선정을 위한 근원-목표 영역 벡터 관계 모델에 대한 전체 구성도이다.

개념적 은유의 근원-목표영역 개념과 관계 추출을 위해 토픽 모델링 방법 중 Top2Vec모델을 기반으로 개념 후보를 추출하고, 개념적 은유 패턴을 기반 의미 관계를 정의하고, 이를 모델링 한다. 또한, 근원-목표영역의 개념 후보 선정을 위해 상위어 인식을 위한 Hearst Pattern의 확장된 패턴을 정의하여 개념 후보와 의미 관계 추출 방법을 이용하여 상위어 후보 추출을 수행한다.

제안된 모델에서 근원-목표영역의 개념 추출 과정은 다음과 같다. 첫째, 토픽

모델 중 하나인 Top2Vec 모델을 적용하여 문서 내 단어들의 핵심 개념 어휘와 의미상으로 유사한 관련 어휘를 추출한다. 둘째, 근원-목표영역의 개념 후보 선정을 위해 핵심 개념 어휘를 중심으로 상위어를 선정한다. 셋째, 상위어 후보 어휘를 중심으로 'is-a' 관계를 추출한다. 넷째, 추출된 근원-목표영역을 기반으로 문서 내 개념적 은유를 선정하고, 이에 대한 정량적 비교평가를 수행한다.

B. 근원-목표 영역 개념 후보 추출

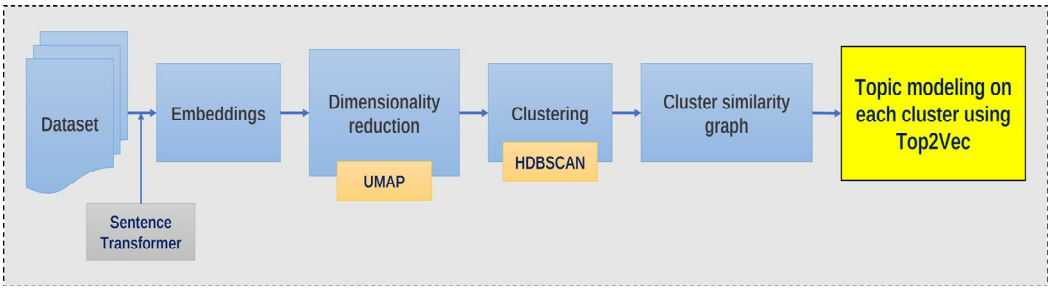


그림 4. 개념 후보 추출 과정

1. 문서 내 핵심 개념 어휘와 관련 어휘 추출

개념적 은유는 하나의 개념영역을 들어 다른 개념영역을 이해하는 것이다. 예를 들어, 우리가 인생이나 사랑을 여행으로 이해하거나 논쟁을 전쟁으로 이해할 때 인생, 사랑, 논쟁은 표현하려는 영역으로서 '목표영역(Target Domain)'이라고 하고, 반면에 여행, 전쟁은 목표영역을 이해하기 위해 수단이 되는 영역으로서 '근원 영역(Source Domain)'이라고 한다.

개념적 은유란 우리에게 익숙한 근원 영역으로써 낯선 목표 영역을 개념화하는 인지 전략이다. 개념적 은유는 우리가 일상 언어생활에서 사용하는 '은유적 언어 표현(metaphorical linguistic expressions)'과 일정한 차이가 있다. 개념적 은유는 구체적이고 친숙한 경험을 바탕으로 하는 근원 영역으로서 추상적이고

모호하며 새롭게 다가가는 목표영역을 명확히 개념화하는 사고의 방식이며, 은유적 언어 표현이란 개념적 은유가 구체적 언어로 표현된 것을 지칭한다[8].

본 논문에서는 근원-목표영역의 개념 추출을 위해 그림 3과 같이 개념 후보 추출 과정을 실행한다. 우선, 개념 추출을 위한 Top2Vec 모델을 적용하기 위한 데이터를 수집한 후, 수집된 데이터에 대해 속성을 선택하고 불용어 처리 등 전처리를 수행하고, 전처리된 데이터를 기반으로 근원-목표영역의 개념 추출과 개념 확장을 위한 토픽 모델과 단어 임베딩 모델 기반으로 Top2Vec 모델을 생성한 후, 생성된 모델을 통해 근원-목표영역의 개념을 선택한다. 또한, 추출된 근원-목표영역의 개념에 대해 Perplexity 지표를 중심으로 제안한 모델을 평가하여 제안한 모델의 성능을 확인하였다.

개념적 은유는 일반적으로 목표영역의 어휘는 추상적인 개념을 사용하고, 근원영역은 유형적이거나 물리적인 개념을 사용한다. 개념적 은유의 한 예인 'THE ARGUMENT IS WAR'의 은유적 표현은 아래와 같다.

1. Your claims are indefensible.
2. He attacked every weak point in my argument.
3. His criticisms were right on target.
4. I demolished his argument.

[출처 : Tendahl, Markus, 2009]

위의 문장(1-4)에서 밑줄 친 부분은 논쟁과 관련된 문장에서 관례로 사용되는 은유적 표현이다. 문장에서 은유적 표현은 'ARGUMENT'의 개념에 대해 더욱 상세하고, 이해하기 쉬운 설명을 제공하는 데 도움을 준다.

표 4. 개념적 은유 개념

	A	IS	B
도메인	목표영역		근원 영역
개념 특성	추상적		물리적, 구체적
예	ARGUMENT	IS	WAR

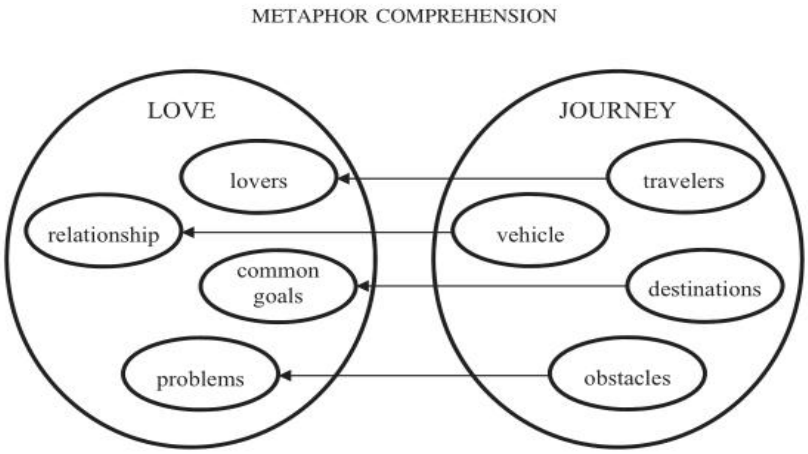


그림 5. “LOVE IS A JOURNEY”에 대한 가상의 개념적 매핑

개념적 은유의 근원 영역과 목표영역은 일련의 체계적인 대응이 있다. 이러한 개념적 대응을 매핑(Mapping)이라고 한다. 개념적 은유 인식을 위해서는 근원-목표영역의 매핑 세트를 추출하는 것이 필수적이다. 개념적 은유를 인식하기 위해서는 개념적 은유와 은유적 표현을 잘 구별해야 한다. 은유적 표현은 구체적인 개념 영역인 근원 영역의 어휘에서 사용되는 어휘의 표현이다. 표 5와 같이 추상적인 ‘ARGUMENT’의 목표 도메인을 이해하기 위해 근원 영역의 전쟁 관련 용어인 ‘won, shoot, demolished, shot down’ 등이 사용된다. 따라서, 은유적 표현은 근원 영역과 관련된 어휘 집합이다.

표 5. 근원 영역 어휘에서 사용되는 은유적 표현

개념적 은유	은유적 표현
ARGUMENT IS WAR	1. I've never won an argument with him. 2. You disagree? Okay, shoot! 3. I demolished his argument. 4. He shot down all of my arguments.

a. 문장 임베딩

개념적 은유의 근원 영역과 목표영역의 핵심 어휘 추출을 위해서 문서 내 잠재된 의미 구조를 발견할 수 있는 토픽 모델링을 수행한다. 토픽 모델링은 문서 내 주제를 가장 잘 나타내는 단어의 가중치 집합으로 토픽을 찾는다. 일반적으로 많이 사용되는 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)과 확률적 잠재 의미 분석(Probabilistic Latent Semantic Analysis, pLSA) 방법은 문서 내 단어의 순서와 의미를 고려하지 않은 문서의 단어 집합에 의존한다. 이는 개념적 은유 인식에서 중요한 역할을 하는 형태소 분석, 전치사, 표제어 등의 요소를 필터링하기 때문에 본 논문의 방법론으로 적합하지 않다. 또한, LDA 및 pLSA는 문장 내 단어의 의미나 순서를 무시하는 BOW(Bag-of-Words) 표현을 사용하기 때문에, 의미적으로 유사하더라도 다른 단어로 인식한다. 따라서, 본 논문에서는 문서 내 주제를 추출하면서, 불용어 제거, 형태소 분석, 표제어와 같은 텍스트 사전 처리를 수행하지 않은 Top2Vec 모델을 기반으로 근원-목표영역의 핵심 어휘를 추출한다. Top2Vec 모델은 조인트 문서와 단어 의미 임베딩을 활용하는 문서 클러스터링을 기반으로 동작한다. 수행된 결과인 토픽 벡터는 문서 및 단어 벡터가 함께 포함되고, 그 사이의 거리는 의미적인 유사성을 나타낸다. 이를 위해, 문서 코퍼스에 대해 Doc2vec를 사용하여 의미 공간(문서 벡터)을 생성한다. 의미 공간은 벡터 간의 거리가 의미 유사성을 나타내는 공간이다[18].

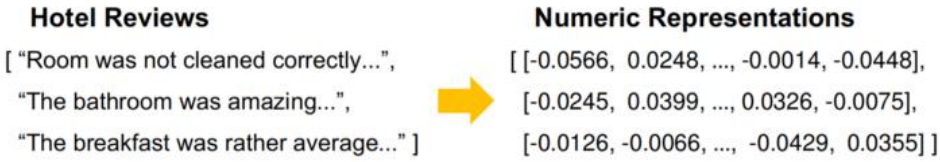


그림 6. 문서 벡터화

그림 7은 의미공간을 보여주는 그림으로 보라색 점은 문서이고, 녹색 점은 단어입니다. 단어는 가장 잘 나타내는 문서에 가장 가깝고 유사한 문서는 서로 가깝습니다[18].

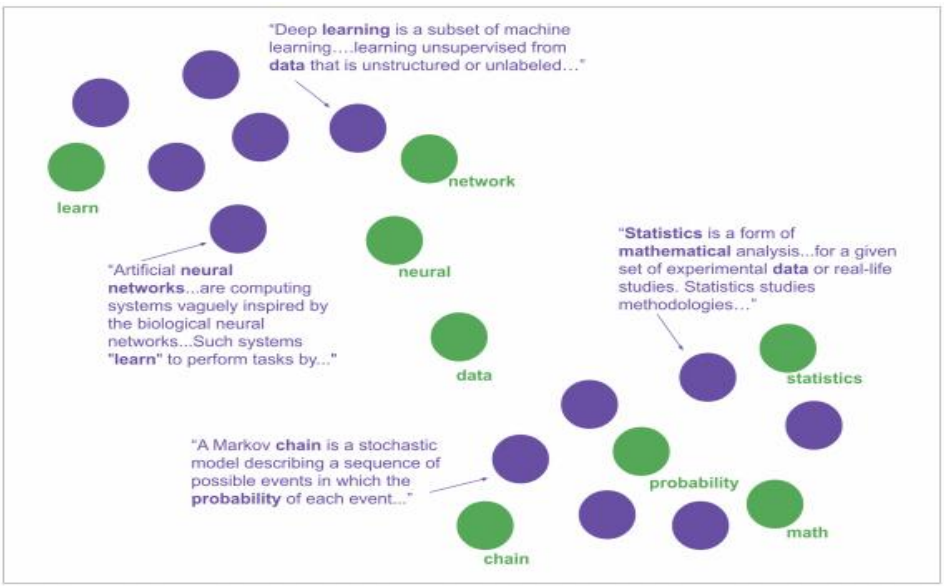


그림 7. 의미 공간의 예

b. 차원 축소

임베딩된 의미 공간(문서 벡터)은 UMAP(Uniform Manifold Approximation and Projection) 축소 방법을 사용하여 축소한다. UMAP은 데이터의 전역 및 로컬 구조를 보존하는 반면, t-SNE(t-Stochastic Neighbor Embedding)는 데이터의 로컬 구조만 보존하기 때문에, t-SNE 대신 UMAP 축소 방법을 선택하였다. 또한, UMAP은 t-SNE에 비해 대용량의 문서 코퍼스에서 좋은 성능을 나타낸다[18].

본 논문에서 차원 축소를 위한 평균 최적값으로 n-Neighbours 매개변수 값을 15로 설정한다.

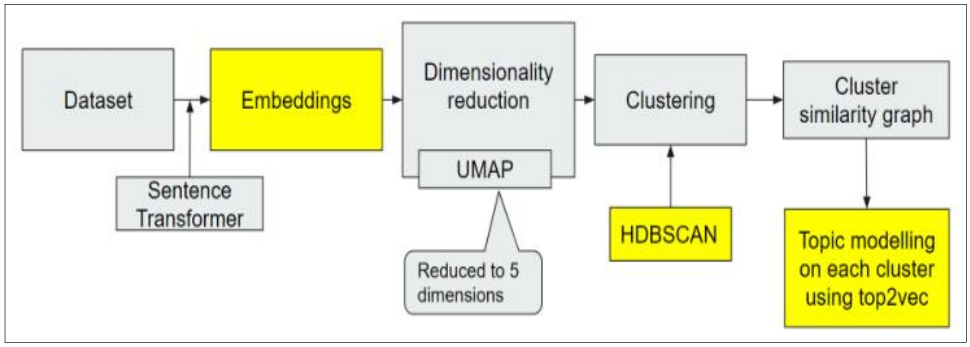


그림 8. 임베딩된 문서 차원 축소 순서도

c. 클러스터링

차원이 축소된 임베딩은 HDBSCAN((Hierarchical Density-Based Spatial Clustering of Applications with Noise) 클러스터링 알고리즘을 사용하여 분류한다. 분류된 군집은 이들 사이의 코사인 거리를 기반으로 연결된 그래프로 표시된다. 선택된 클러스터에 대해 Top2Vec 기반 토픽 모델링을 수행한다. HDBSCAN은 UMAP이 저차원 공간에서도 로컬 구조를 유지하기 때문에 UMAP에 어울리는 밀도 기반 알고리즘이다. HDBSCAN에는 다음과 같은 특징을 가지고 있다. 첫째, HDBSCAN은 모든 데이터 포인트를 클러스터로 강제하지 않지만, 데이터 포인트가 이상값으로 간주되도록 허용한다. 둘째, HDBSCAN은 LDA, K-Means와 같은 클러스터 수를 설정할 필요가 없다. 셋째, HDBSCAN은 계층 구조를 제어할 수 있는 계층적 클러스터링을 수행한다. 또한, HDBSCAN은 의미 있는 클러스터만 생성하고, 노이즈를 클러스터 하지 않는다. 따라서 HDBSCAN 기반 클러스터링 품질은 다른 클러스터링 알고리즘과 비교할 때 좋은 성능을 나타낸다. 클러스터를 계층적으로 결합하고 분리하는 클러스터 수를 제어할 수 있다. 이는 개념적 은유의 근원 영역과 목표영역의 핵심 어휘에 대해 특정 클러스터 내에서 더 정밀한 주제를 찾는 데 효과적이다.

토픽 벡터를 인식하기 위해서 의미 공간의 밀도가 높은 영역을 찾기 위해 밀도 기반 클러스터링 알고리즘인 HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise) 알고리즘을 사용한다. HDBSCAN 알고리즘 적용 시, 다차원 희소 공간 문제를 해결하기 위해 차원 축소를 수행한다. 문서 벡터의 차원을 줄이기 위해 UMAP(Uniform Manifold Approximation and Projection for Dimension Reduction)을 사용한다. 의미 공간을 저차원 공간으로 압축한 후, HDBSCAN을 사용하여 문서의 밀집 영역을 찾는다.

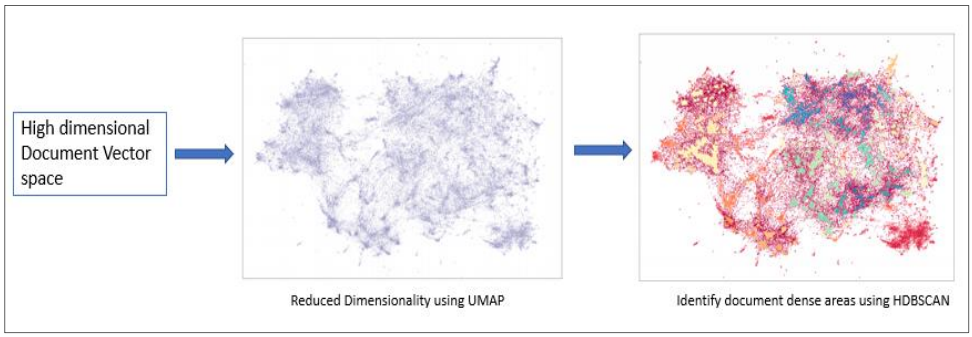


그림 9. 문서 밀집 영역 식별

그림 9에서 각 점은 문서를 나타내고 문서의 밀집된 영역은 색상이 지정되어 있고, 각 영역은 주제를 나타낸다. 빨간색 점은 클러스터에 속하지 않는 값이다[18]. 각 밀집 영역에 대해 동일한 클러스터에 있는 모든 문서 벡터의 산술 평균을 이용하여 주제 벡터를 구한다. 각 문서에는 해당 문서 벡터에 가장 가까운 토픽 벡터를 기반으로 하는 토픽 번호가 할당된다.

d. 토픽 생성

토픽을 생성하는 문서 코퍼스에 대해 TF-IDF를 적용할 때 문서 간의 단어 중요성을 비교한다. TF-IDF 점수는 토픽에서 중요한 단어를 강조한다. 이는 토픽을 생성하는 문서 클러스터의 각 단어에 대해 중요도를 측정한다. 단어와 문서 벡터로 구성된 의미공간은 이산 공간에서 토픽을 샘플링하는 LDA와 다르게 토픽을 연속적으로 표현한다. 단어와 의미공간을 이용하여 문서의 높은 밀집 영역

은 유사한 주제를 갖고, 의미적으로 유사한 단어가 주변에 분포한다.

이와 같은 방법을 통해 본 장에서는 개념적 은유의 근원 영역과 목표영역의 핵심 어휘 추출을 위해 미국의 18개 미국 간행물에서 204,135개의 뉴스 데이터 중 CNN 뉴스를 활용한다. 날짜, 제목, 출판, 기사 텍스트, 출판 이름, 연도, 월 및 URL 등이 포함된 데이터로, 2013년부터 2018년 초까지의 뉴스 기사이고, 14,300개의 레코드를 포함하고 있다. 또한, 샘플 데이터는 총 8,302,822개의 단어와 130,061개의 중복되지 않은 고유 단어를 포함하고 있다.

id	title	author	date	content	year	month	publicat
50358	Istanbul attack: Dozens killed a	Euan McKirdy	2016-12-31	[Istanbul (CNN)At least 39 people were killed and at least	2016	12	CNN
50359	Alabama, Clemson back in nati	Jill Martin	2016-12-31	[Atlanta (CNN)This season's College Football Playoff cent	2016	12	CNN
50360	New year celebrations ring in 2(	Ray Sanchez	2016-12-31	[(CNN)Revelers on the United States' west coast cheered	2016	12	CNN
50361	Trump says he has inside inforn	Kevin Liptak	2017-01-01	[West Palm Beach, Florida (CNN)President-elect Donald T	2017	1	CNN
50362	3 dead in Texas plane crash col	Tony Marco	2017-01-01	[(CNN)Two small planes collided in Texas on Saturday, k	2017	1	CNN
50363	21 rescued from California theme park ride		2016-12-31	[(CNN)A ride up to see the sights of southern California	2016	12	CNN
50364	Trump wishes Happy New Year	Eugene Scott	2016-12-31	[(CNN)President-elect Donald Trump is not quite ready t	2016	12	CNN
50365	Timeline: Turkey's bloody year	Greg Botelho	2016-12-11	[(CNN)In one of the most terror-scarred countries in the	2016	12	CNN
50366	US utility: Alleged Russian malv	Evan Perez	2016-12-31	[(CNN)The indicators from the malicious software found	2016	12	CNN
50367	NBC and Charter extend talks before deadline			[]			CNN
50368	Roberts praises lower court jud	Ariane de Vogu	2016-12-31	[(CNN)Chief Justice John Roberts devoted his annual rep	2016	12	CNN
50369	Former student charged in teac	Shachar Peled	2016-12-28	[(CNN)A 23-year-old man who as a high school student	2016	12	CNN
50370	Liverpool beats Manchester City, Chelsea equal		2016-12-31	[(CNN)Chelsea registered a record-equaling 13th success	2016	12	CNN
50371	Greek ambassador to Brazil kill	Marilia Brocche	2016-12-31	[(CNN)Police contend Greece's ambassador to Brazil was	2016	12	CNN
50372	Police kill suspect in PA troope	Azadeh Ansari	2016-12-31	[(CNN)A 32-year-old man suspected of killing a Pennsylv	2016	12	CNN
50373	Parents of American journalist missing in Syria 4 years hold (			[]			CNN
50374	Rebel warning over Syrian ceas	Laura Smith-Sp	2016-12-31	[(CNN)Rebels in the Free Syrian Army have warned they	2016	12	CNN
50375	Trump ditches press pool to pla	Eugene Scott	2016-12-31	[(CNN)President-elect Donald Trump ditched his press p	2016	12	CNN
50376	Skakel murder case: Court reinstates conviction		2016-12-30	[(CNN)Michael Skakel could be headed back to prison a	2016	12	CNN

그림 10. CNN 뉴스 샘플 데이터

해당 뉴스 데이터에서 뉴스 기사에 해당하는 'content' 칼럼의 텍스트를 이용하여 핵심적인 개념 어휘를 추출한다. 이를 위해, 입력된 뉴스 기사를 Top2Vec 임베딩하고, 진행 과정에서 벡터화 품질과 학습 시간을 고려했다. 벡터화를 진행한 결과, 총 151개의 토픽이 추출되었다.

각 토픽은 해당 토픽 클러스터에 속한 원본 문서의 중심(평균점)인 토픽 벡터를 구한다. 이를 이용해 키워드 세트를 사용하여 주제에 레이블을 지정하기 위해 주제 중심 벡터에 대해 가장 가까운 단어를 계산한다.

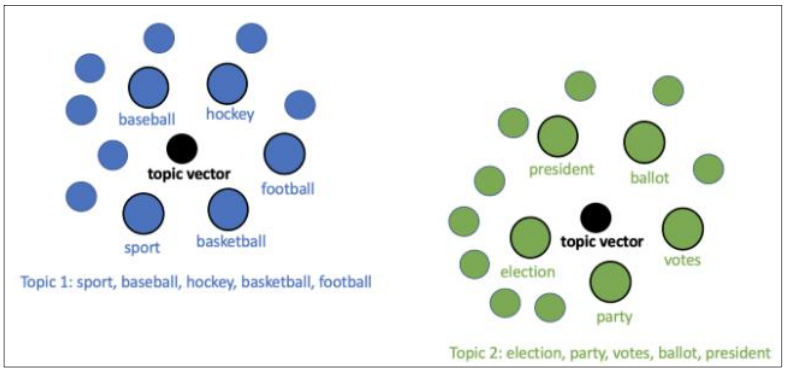


그림 11. 추출된 topic vector

그림 11과 같이 추출된 토픽 151개 그룹 중 임의의 토픽 5개를 워드 클라우드로 표현하면 그림 12와 같다. 각 토픽은 주제별로 '예술', '영화', '음악' 등의 그룹으로 나타났다.

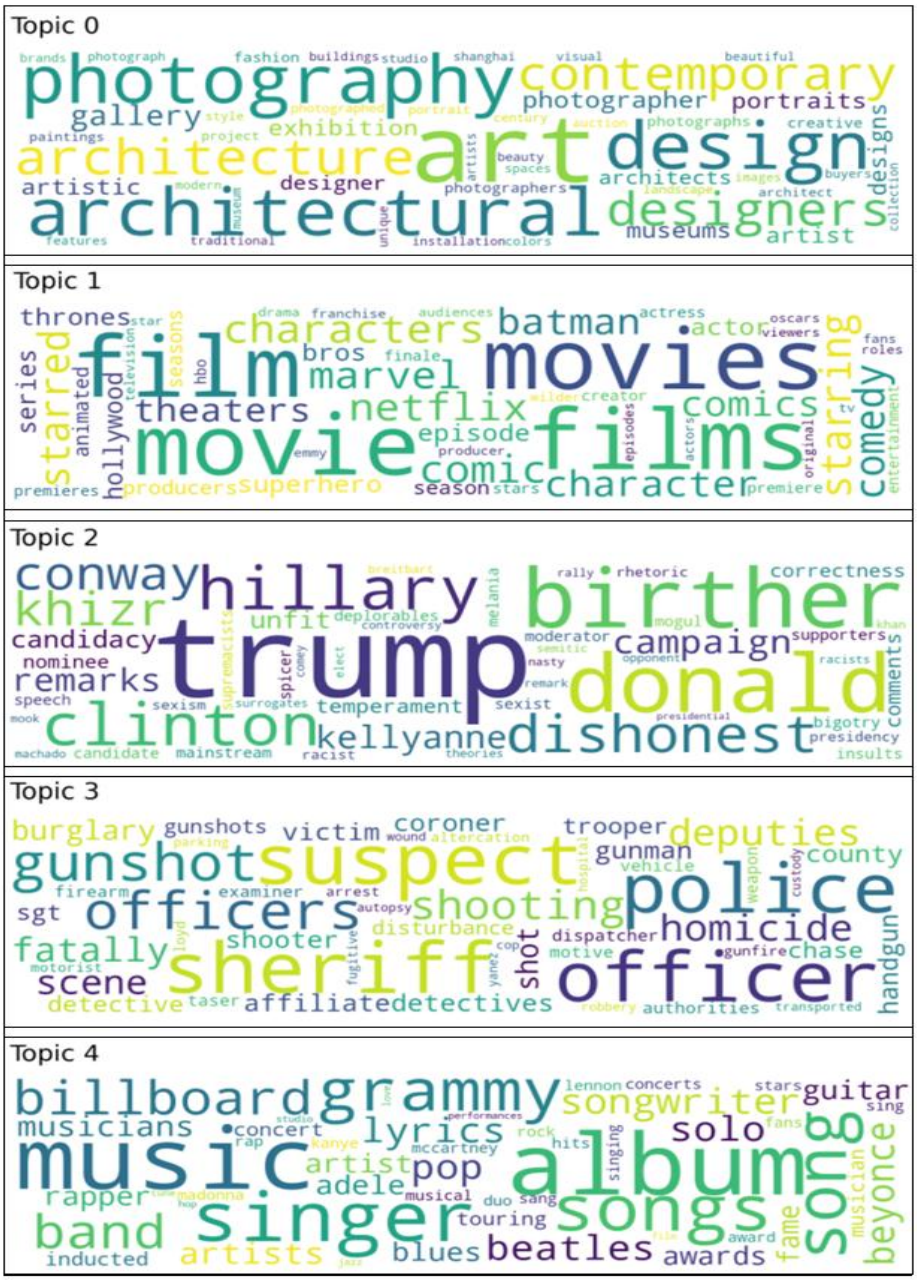


그림 12. 임의의 토픽 5개의 워드 클라우드 결과

개념적 은유의 근원 영역과 목표영역의 핵심 어휘는 추출된 토픽 중 특정 토픽에 속하는 문서에서 추출해야 하므로 토픽별 문서 검색을 수행한다. 토픽별 문서

검색은 토픽에 대한 문서의 의미적 유사성을 기반으로 수행된다. 이를 위해 문서 및 토픽 벡터의 코사인 유사도를 사용한다.

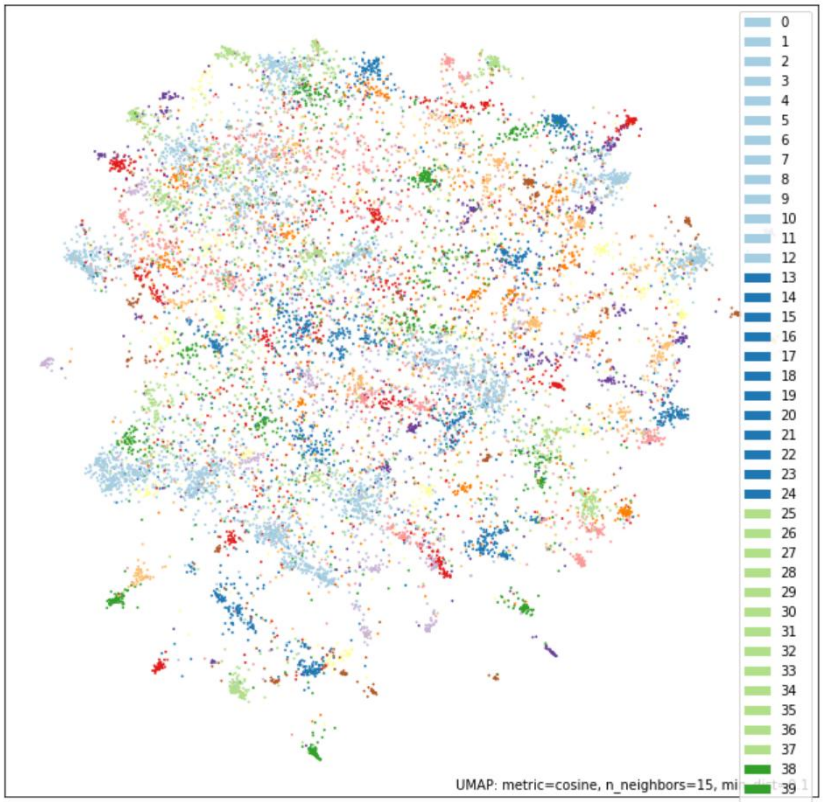


그림 13. UMAP을 이용한 151개 토픽 차원 축소 분포

그림 13은 임베딩 공간에 대해 비선형 차원 축소 알고리즘인 UMAP을 사용하여 차원이 축소된 분포도이다. n_neighbors 파라미터는 15로 설정하고, 토픽 사이의 최소 거리를 정의하기 위해 min_dist는 0.1로 설정하였다.

본 논문에서는 근원 영역과 목표영역의 핵심 어휘가 모두 관련된 문서를 동시에 검색해야 하므로, 두 개의 키워드를 검색하는 방법을 고려한다. 표 6은 특정 단어('police')를 중심으로 의미적으로 유사한 문서를 검색한 결과를 출력한 예이다. 검색 결과는 3개의 결과만 출력되도록 조정하였다. 'Document'는 문서 번호이고, 'Score'는 코사인 유사도 측정값이다.

표 6. 특정 단어('police')를 중심으로 의미적으로 유사한 문서 검색한 결과 출력한 예

Document: 6431, Score: 0.45776331424713135

[(CNN)At least 64 law enforcement officers have been shot and killed this year, the most in five years, according to the National Law Enforcement Officers Memorial Fund. , The 2016 shootings have spanned the nation, from California to Massachusetts. , They've exceeded the annual average of police shooting deaths over the past 10 years, 53. And this year's total is higher than the number of firearms-related police deaths in 2015. According to the fund, firearms were responsible for 41 of 123 officer fatalities in 2015, one of the safest years for officers on record. , The decade's highest total came in 2011, with 73 officers shot dead., This year's shooting deaths include an officer on her first day on the job and a sheriff's corporal who was about to retire....]

Document: 8684, Score: 0.44441866874694824

[(CNN)Police are on high alert -- and in at least one city, mourning -- after four officers were shot Sunday in incidents around the country. , Perhaps most startling to law enforcement is that in three of the shootings, it appears the alleged perpetrators specifically targeted police officers, according to authorities and local media reports....]

Document: 6987, Score: 0.4078167676925659

[(CNN)A police officer was killed when gunfire erupted as he pursued a suspect on foot in southern New Mexico, authorities said., Officer Clint Corvinus, 33, of the Alamogordo Police Department was fatally shot Friday while he and another officer chased down the suspect, who was also shot dead during the incident....]

[출처 : CNN]

또한, 개념적 은유의 근원 영역과 목표영역에 해당되는 두 개의 키워드('police', 'olympic')를 이용하여 문서를 검색한 결과는 표 7과 같다[19].

표 7. 두 개의 키워드('police', 'olympic')를 이용하여 문서 검색한 결과

Document: 6170, Score: 0.43347132205963135

[(CNN)The athletes have arrived, the venues are ready and expectation is palpable as the sporting world awaits the commencement of the 31st Olympic Games, the first ever Olympiad to be staged in South America., Although there is little by the way of sporting action Friday, there's still plenty happening in host city Rio de Janeiro., Here are five things to look out for on day zero of the 2016 Olympic Games., Opening ceremony, Rio 2016 will officially get underway Friday when IOC president Thomas Bach opens the Games in the Maracana Stadium. Before that an estimated television audience of billions will tune in to view an opening ceremony that is expected to showcase the best of Brazilian culture and sport -- follow CNN's blog coverage here....]

Document: 5905, Score: 0.4278363585472107

[(CNN)The governing body of world aquatics has banned seven Russian swimmers from the Rio Olympics because of doping violations, the group said Monday., The Federation Internationale de Natation (FINA), however, ruled the Russian synchronized swimming team, divers and the women's water polo team are still eligible to participate because they have not been implicated in the ongoing doping scandal....]

Document: 6399, Score: 0.4155385494232178

[(CNN)Sunday was a golden day for some of the biggest names in Olympic sports, with runners and gymnasts harvesting those precious medals at the Rio 2016 Games., And in the event of the evening, the fastest man in history, Usain Bolt, stayed right there in front, taking the gold medal in the men's 100 meters in 9.81 seconds. , It was also the chance for the world's best gymnasts -- both men and women -- to show their incredible specialist skills on the apparatus....]

[출처 : CNN]

일반적으로 토픽 모델링 관련 연구에서는 지정된 주제에 대한 상위 단어를 생성하는 단계가 있지만, 본 논문에서는 개념적 은유의 근원 영역과 목표영역의 핵심 어휘와 연관된 토픽을 추출해야 하므로, 특정 토픽에 대해 같은 주제의 단어를 검색할 필요가 있다. 다음 표 8은 본 논문에서 제시한 방법을 이용하여 특정 단어 한 개를 입력했을 때, 관련 단어 상위 6개와 유사도를 측정한 결과이다[19].

표 8. 특정 단어와 의미적으로 유사한 상위 6개 단어와 유사도

검색 단어	유사 단어 및 유사도
“police”	officers 0.8260947052559264 officer 0.7335886764989438 suspect 0.7149735073713546 shooting 0.6892047899497351 arrested 0.6781121006154538 shot 0.6628881502355699
“olympic”	rio 0.8646687603913213 olympics 0.8369427128812612 medals 0.7571148174236726 gold 0.7437170561728466 medalist 0.7213711583620623 games 0.7040126607960367
“god”	faith 0.5605402592091314 church 0.550112244684918 love 0.5457138704620763 my 0.5397571635460767 theology 0.5182957824890648 jesus 0.5018108992444168
“journey”	faith 0.5605402592091314 church 0.550112244684918 love 0.5457138704620763 my 0.5397571635460767 theology 0.5182957824890648 jesus 0.5018108992444168
“art”	artist 0.6995137636448607 contemporary 0.6869768921441608 exhibition 0.6360569770071887 gallery 0.6103803694631849 artistic 0.6044104370254337 photography 0.6007255710104091
“movie”	film 0.7796233324237345 movies 0.775671178893141 films 0.6654072774617057 starring 0.6482709643410052 hollywood 0.6198818314807624 character 0.5974882575449706

개념적 은유는 문장 내에서 근원 영역과 목표영역 간의 의미적 매핑이 나타난다. 일반적으로 근원 영역의 구조는 목표영역 구조에 매핑 된다. 예를 들어, 'POVERTY IS A DISEASE' 이라는 개념적 은유에서 '빈곤' 영역의 가난한 사람은 '질병' 영역에서 '질병을 경험하는 환자'로 표현된다. 따라서, '질병' 영역의 역할은 '빈곤' 영역의 해당 영역에 매핑된다. 이는 한 문장 내에서 개념적 은유 표현이 사용됐다면, 두 영역의 어휘가 동시에 출현해야 함을 의미한다. 특정 단어 한 개에 대한 검색뿐만 아니라, 근원 영역과 목표영역에 속한 단어를 추출할 필요가 있다. 이는 두 영역에서 사용되는 단어가 동시에 의미적으로 표현되는 문장을 인식하는 방법과 연관단어를 추출하기 위해 사용된다. 표 9는 두 단어를 동시에 검색하여 유사도를 측정한 결과이다.

표 9. 두 단어와 의미적으로 유사한 상위 6개 단어와 유사도

검색 단어	유사 단어 및 유사도
"movie" + "art"	film 0.6951681054371743 movies 0.6403788344525585 artist 0.6390603853975375 films 0.5995843435298476 contemporary 0.5682952832500399 artistic 0.5679172100510507
"olympic" + "police"	rio 0.6583118780921509 officers 0.6125457916589019 olympics 0.6053703218271207 gold 0.5588616457101875 shooting 0.547810659799587 shot 0.5427592231296535
"disease" + "food"	diseases 0.6320081656960014 symptoms 0.5881881631031565 prevention 0.5834017625613557 healthy 0.5669591165516459 infections 0.5619370403428978 patients 0.5535925388414771
"music" + "jazz"	songs 0.6580094159640537 singer 0.6224925207923211 musicians 0.6025112003277235 song 0.582242699928014 album 0.572554095564078 band 0.5708015202682132

C. 개념적 은유 인식을 위한 관계 추출

1. 개념적 은유 패턴 기반 의미 관계 추출

텍스트에서 개념적 은유를 식별하는 것은 문장 레벨, 문법적 관계 레벨, 단어 레벨 별로 수행할 수 있다. 문장 레벨에서는 개념적 은유적 표현에 대해 특정 주석이 없이 개념적 은유가 사용된 단어 혹은 표현을 포함하여 전체 문장을 개념적 은유 그대로 분류한다. 문법적 관계 레벨의 개념적 은유 식별은 근원 영역과 목표영역의 단어가 모두 개념적 은유 표현으로 분류되는 단어 쌍을 기반으로 특정한 문법적 관계를 활용한다. 문장이 다양한 구문 구조를 가진 하위 구문으로 분할되는 방식을 이용하기 때문에 구문 중심의 개념적 은유 식별이라고도 한다. 일반적으로 사용되는 문법 관계는 동사 또는 형용사의 은유가 명사와의 연관성을 고려하여 식별되는 동사-명사 및 형용사-명사 관계이다. 단어 레벨의 개념적 은유 식별은 토큰 수준에서 문장의 각 단어를 문맥에 따라 은유적으로 사용되는지를 결정한다. 이 접근방식은 시퀀스 라벨링 또는 단일 단어로 분류하여 처리한다. 이때, 근원 영역 단어만 은유적 레이블로 지정한다. 단어 수준에서 다양한 구문 유형의 개념적 은유를 식별하지만, 동사를 중심으로 접근방식이 일반적이다.

개념적 은유는 특정한 규칙적인 패턴으로 표현된다. 예를 들어, 명사구 ‘poverty trap’은 목표영역 단어 ‘poverty’에 의해 사용되는 근원 영역의 단어 ‘trap’이 있다. 이러한 명사-명사 패턴의 개념적 은유 구조는 이러한 종속 관계에서 일관되게 나타난다. 하지만, 근원 영역의 단어가 목표영역을 표현하기 위해 사용하는 역방향은 관찰되지 않는다. 따라서, 개념적 은유 표현에서 자주 발생하는 여러 유형의 사용 패턴을 정의할 필요가 있다. 이를 위해 종속 구문 분석(Dependency Parsing)을 이용하여 관계를 정의한 후, 개념적 은유 구조 패턴을 추출한다. 종속 구문 분석은 문장 내 단어 간의 의존성을 기반으로 문장의 문법적 구조를 분석하는 방법이다. 이는 문장의 구문은 단어 간의 종속 그래프를 사용하여 단어 사이의 방향이 지정된 에지로 표현한다. 종속성 구문 분석에서는 이를 다양한 태그로 정의하여 문장에서 두 단어 간의 관계를 나타낸다. 예를 들어, ‘rainy weather’라

는 구에서 단어 'rainy' 는 명사 'weather'의 의미를 변형시킨다. 'weather'는 헤드 역할을 하고, 'rainy '는 종속(혹은 자식) 역할을 한다. 이러한 종속성은 그림 14와 같이 형용사 수식어를 나타내는 amod 태그로 표시된다.

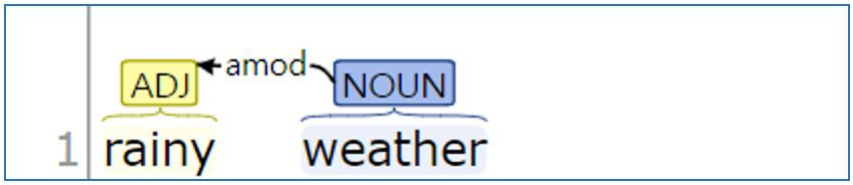


그림 14. 종속 구문 분석

문장 내에는 단어 사이에 많은 의존 관계가 존재하지만, 의존 관계는 오직 두 단어만 포함되어 하나는 헤드 역할을 하고, 다른 하나는 자식 역할을 한다. 현재 Universal Dependency에서 37개의 종속 관계를 정의하고 있다. 종속 관계 목록은 표 10과 같다[20].

표 10. 종속 구문 관계 태그

	Nominals	Clauses	Modifier words	Function words
Core arguments	nsubj (주어) obj (목적어) iobj (간접목적어)	csubj (주어절) ccomp (보문) xcomp(개방보문)		
Non-core dependents	obl (사격어) vocative (호격) expl (허사) dislocated (전위)	advcl (부사절)	advmod (부사어) discourse (담화표지)	aux (조동사) cop (계사) mark (절접속표지)
Nominal dependents	nmod (명사관형어) appos (동격어) nummod (수관형어)	acl (관형절)	amod (관형어)	det (한정사) clf (분류사) case (격표지)
Coordination	MWE(Multi word expression)	Loose	Special	Other
conj (병렬접속) cc (등위접속)	fixed (관용구) flat (명사구) compound (합성어구)	list (목록) parataxis (무표지접속)	orphan (생략) goeswith (분할어절) reparandum(발화수정)	punct (구두점) root(최상위지배소) dep (주석불가)

예를 들어, “Kaggle is the largest community of data scientists and provides best resources for understanding data and analytics.” 문장을 구문 분석해 보면 각 토큰의 속성별로 다음 표 11과 같은 결과가 나타난다.

표 11. 텍스트 구문 분석

텍스트	기본형	품사	종속 구문 관계	ALPHA	불용어
Kaggle	Kaggle	PROPN	nsubj	True	False
is	be	AUX	ROOT	True	True
the	the	DET	det	True	True
largest	large	ADJ	amod	True	False
community	community	NOUN	attr	True	False
of	of	ADP	prep	True	True
data	datum	NOUN	compound	True	False
scientists	scientist	NOUN	pobj	True	False
and	and	CCONJ	cc	True	True
provides	provide	VERB	conj	True	False
best	good	ADJ	amod	True	False
resources	resource	NOUN	dobj	True	False
for	for	ADP	prep	True	True
understanding	understand	VERB	pcomp	True	False
data	datum	NOUN	dobj	True	False
and	and	CCONJ	cc	True	True
analytics	analytic	NOUN	conj	True	False
.	.	PUNCT	punct	False	False

또한, 예문을 종속 구문 분석을 통해 분석하면 다음과 같은 종속 그래프로 표현할 수 있다.

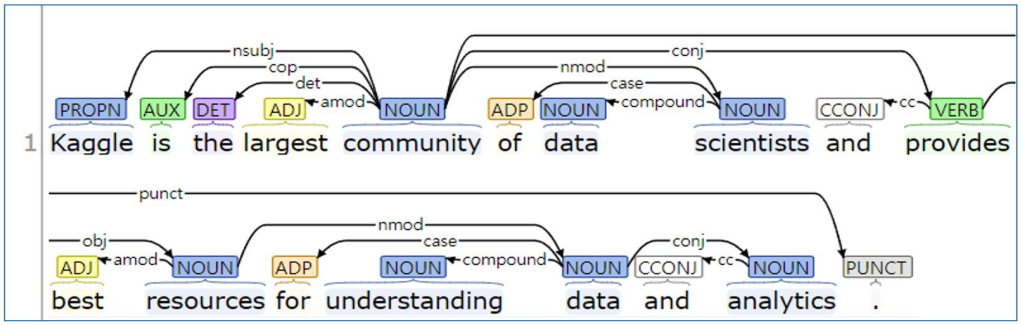


그림 15. 종속 구문 트리 예

그림 15 종속 구문 트리에서 핵심 단어는 “community”이고, 품사 태깅으로 명사 역할을 하고 있으며, 이 문장의 다른 모든 단어는 목적어, 주어, 수식어 등과 같은 종속 관계로 “community”에 연결된다. 이러한 관계는 문장에서 각 단어의 역할과 기능 및 여러 단어가 함께 연결되는 방식을 정의할 수 있다. 따라서, 위 문장에서 주체-관계-종속객체 관계는 <Kaggle><is><the largest community of data scientists>와 같은 패턴을 추출할 수 있다.

문장 내 모든 개념은 명사구에서 추출된다. 따라서, 명사구는 개념 간의 관계를 결정하는데 사용할 수 있다. 관계를 분석하는 과정을 우선, 개념으로서의 모든 명사구를 찾고, 명사구의 수식어를 찾는다. 또한, 종속 구문 관계 중 ‘nsubj(주어)’, ‘acl(관형절)’, ‘obj(목적어)’이 포함된 동사의 주어를 찾는다. 본 논문에서 사용된 동사의 주어를 인식하기 위한 패턴은 다음 표 12와 같다.

표 12. 동사의 주어를 인식하기 위한 패턴

패턴	문장 추출 예
dep(nsubj)	A [woman] is [playing] the piano.
dep(acl)	A [woman] [playing] the piano.
dep(pobj) - head_dep(agent) - head_pos(VERB)	The piano is [played] by a [woman].

이를 기반으로 개념 사이의 관계를 추출하기 위한 패턴을 표 13과 같이 정의한다.

표 13. 관계 인식을 위한 종속 구문 패턴 정의

관계 추출 패턴	문장 예
dep(dobj) - head in relation_subj	A woman is [playing] the [piano]. The woman [is] a [pianist].
dep(pobj) - dep(agent)	The piano is played [by] a [woman].
head_pos(VERB) - phrasal_verb	A [woman] is playing with the piano in the [room].
head_pos(VERB) - dep(acl)	A [woman] is playing the piano in the [room]. A [woman] playing the piano in the [room].
dep(pobj) - phrasal_prep	A [woman] in front of a [piano].
head_pos(NOUN)	A [piano] in the [room].
dep(amod, advmod) - head_pos(NOUN)	A [piano] next to a [woman].
dep(amod, advmod) - head_pos(VERB)	A [woman] standing next to a [piano].
head_pos(VERB) - head in relation_subj	A [woman] is playing the [piano] in the room.
dep(nsubjpass) - head in relation_subj	The [piano] is played by a [woman]. #수동태 (subjpass -> obj, objpass -> subj)

종속 구문 패턴을 통해 표 14와 같은 예문을 통해 관계를 추출한 결과는 그림 16과 같다.

표 14. 관계 패턴 추출을 위한 샘플 문장

A woman is playing the piano in the room.
A piano is played by a woman in the room.
A woman is playing the space craft at NASA.
A woman is playing with a space craft at NASA.
A woman standing next to a piano.
The woman is a pianist.
A giraffe grazing a tree in the wildness with other wildlife.
Cow standing on sidewalk in city area near shops.

표 15. 관계 패턴 추출을 위한 종속 구문 분석

{'entities': [{'head': 'woman',
'lemma_head': 'woman',
'lemma_span': 'a woman',
'modifiers': [{'dep': 'det', 'lemma_span': 'a', 'span': 'A'}],
'span': 'A woman',
'span_bounds': (0, 2),
'type': 'unknown'},
{'head': 'space craft',
'lemma_head': 'space craft',
'lemma_span': 'a space craft',
'modifiers': [{'dep': 'det', 'lemma_span': 'a', 'span': 'a'}],
'span': 'a space craft',
'span_bounds': (5, 8),
'type': 'unknown'},
{'head': 'NASA',
'lemma_head': 'NASA',
'lemma_span': 'NASA',
'modifiers': [],
'span': 'NASA',
'span_bounds': (9, 10),
'type': 'unknown'}],
'relations': [{'lemma_relation': 'play with',
'object': 1,
'relation': 'playing with',
'subject': 0},
{'lemma_relation': 'at',
'object': 2,
'relation': 'at',
'subject': 1}]}

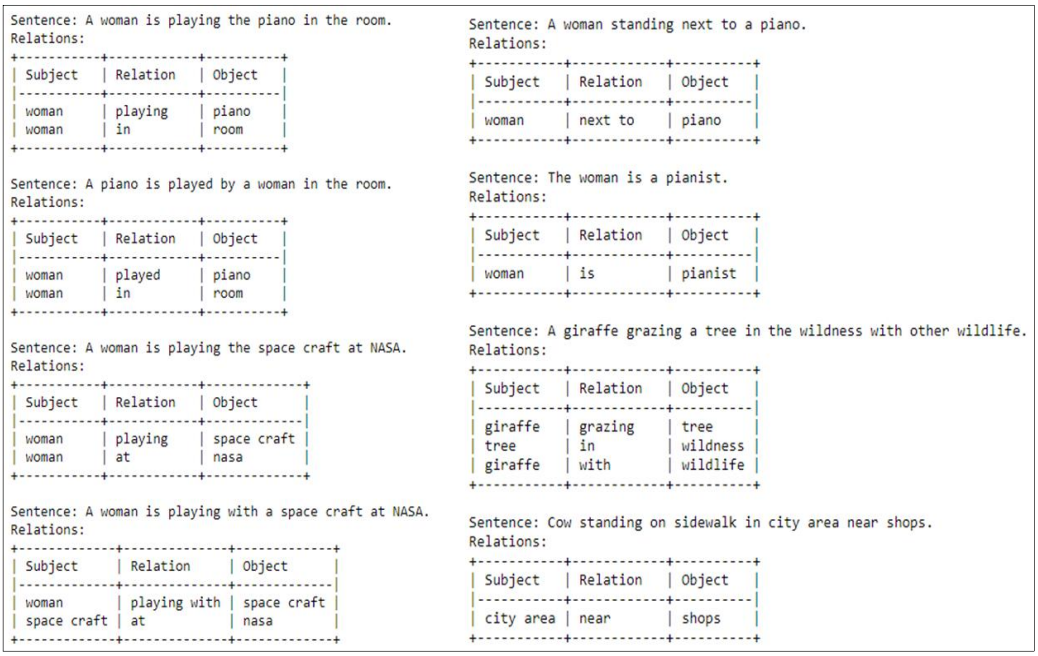


그림 16. 종속 구문 패턴을 이용한 관계 추출 결과

2. 근원-목표 영역의 개념적 은유를 위한 상위어(Hypernym) 인식

개념적 은유는 개념적 영역에 걸친 일반적인 사고의 대응이다. 이 대응은 근원 영역, 목표영역, 그리고, 근원-목표영역 사이의 관계와 같은 공통적인 구조로 되어있다. 근원 영역의 개념은 목표영역의 개념에 해당한다. 개념적 은유 대응은 목표영역의 고유한 구조와 일치하는 방식으로 근원 영역의 인지적인 구조를 보존해야 한다. 이는 개념적 은유를 사용할 때 근원 영역에 대한 지식을 사용하여 목표영역의 도메인에 대해 추론하기 때문이다.

문장 내 어휘들의 의미 관계는 일반적으로 반의 및 동의 관계, 상위-하위 관계, 부분-전체 관계 등으로 나눌 수 있다. 특히, 상-하위 관계와 부분-전체 관계는 의미 관계가 계층 구조를 이룬다는 점에서 계층적 관계로 분류한다. 단어 그룹에서 상위어 인식을 위해서는 문장 내에서 의미적으로 계층적 관계를 이루고 있는 추출하는 방법이 요구된다. 예를 들어, ‘such X as {Y1, Y2, Y3…}’의 패턴을

이루는 문장을 살펴보면, '……works by such authors as Herrick, Goldsmith, and Shakespeare.'에서 'authors'는 'Herrick', 'Goldsmith', 'Shakespear'의 상위어로 판별할 수 있다. 이는 Hearst가 제안한 방법으로 Hearst Pattern이라고 부른다. 텍스트에서 상-하위 의미 관계를 추출하는 방법으로, 패턴 인식과 의미 관계를 이용하는 것이 구문 분석하는 것보다 정확하고 효과적인 방법이라고 제안하였다[21]. Hearst Pattern은 다음 표 16과 같이 6개의 패턴을 정의하고 있다[22].

표 16. Hearst 상-하위 관계 패턴

① NP_0 such as {NP_1 ,NP_2 ...,(and | or)} NP_n
② such NP as {NP,}* {(or|and)} NP
③ NP {,NP}* {,} or other NP
④ NP {,NP}* {,} and other NP
⑤ NP {,} including {NP,}* {or | and} NP
⑥ NP {,} especially {NP,}* {or | and} NP

이러한 패턴을 이용하여 문장 내에서 추출된 상위 개념에 대한 하위 단어를 하위어(Hyponym)라고 하고, 하위 개념을 포함하는 상위 개념의 단어를 상위어(Hypernym)라고 한다. 예를 들어, "GDP in developing countries such as Vietnam will continue growing at a high rate." 문장에 대한 품사 태깅과 종속 구문 분석을 수행하면 다음 표 17과 같이 나타난다.

표 17. 예제 문장에 대한 품사 태깅

토큰	종속 구문 태그	POS 태그
GDP	nsubj	PROPN
in	prep	ADP
developing	amod	VERB
countries	pobj	NOUN
such	amod	ADJ
as	prep	ADP
Vietnam	pobj	PROPN
will	aux	AUX
continue	ROOT	VERB
growing	xcomp	VERB
at	prep	ADP
a	det	DET
high	amod	ADJ
rate	pobj	NOUN
.	punct	PUNCT

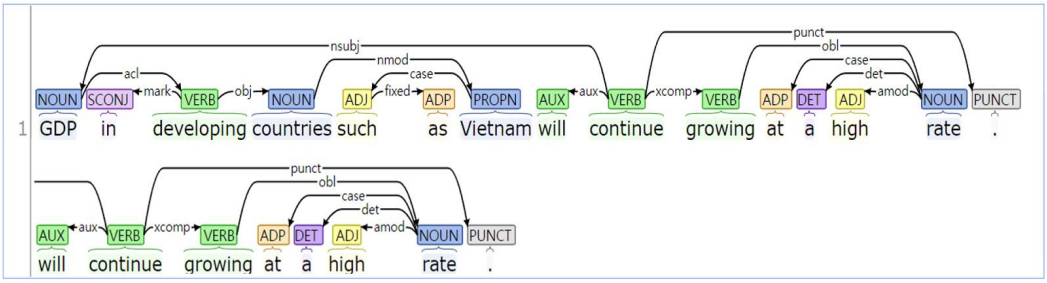


그림 17. 예제 문장의 종속 구문 트리 분석

구문 분석을 수행 후, Hearst Pattern 중 'X such as Y' 패턴을 추출하기 위해서 다음 표 18과 같은 패턴 구문을 정의한다. 패턴 구문을 이용하여 추출된 결과는 'countries such as Vietnam'이고, 'Vietnam'의 상위 개념은 'countries' 임을 알 수 있다.

표 18. 'X such as Y' 패턴을 추출하기 위한 패턴 구문

{'POS' : 'NOUN'} - {'LOWER' : 'such'} - {'LOWER' : 'as'} - {'POS' : 'PROPN'}

본 논문에서는 근원 영역과 목표영역의 상위어 인식을 위해 POS 태깅으로 구성된 Hearst Pattern와 종속 구문 트리 패턴을 표 19와 같이 정의한다. 본 논문에서는 종속 구문 분석을 통해 관계를 인식하기 때문에, 종속 구문 트리 패턴에 의한 Hearst Pattern 추출이 필요하다.

표 19. Hearst Pattern에 대한 종속 구문 트리 패턴 정의

Hearst Patterns	Dependency Patterns
NP_x *such as* NP_y	case(NP_yHead, such) nmod:such as(NP_xHead, NP_yHead)
Such NP_x *as* NP_y	amod(NP_xHead, such) case(NP_yHead, as) nmod: as(NP_xHead, NP_yHead)
NP_x *including* NP_y	case(NP_yHead, including) nmod:including(NP_xHead, NP_yHead)
NP_y *and* | *or other* NP_x	cc(NP_yHead, and | or) amod(NP_xHead, other) conj(NP_yHead, NP_xHead)
NP_x *especially* NP_y	advmod(NP_xHead, especially) dep(NP_xHead, NP_yHead)
NP_y *is a* NP_x	nsubj(NP_xHead, NP_yHead) cop(NP_xHead, was | were | is | are)

일반적으로 문장은 NP(Noun Phrase, 명사구)에서 의미적 관계가 형성된다. 명사구는 구문 분석을 기반으로 하는 구문 패턴으로 식별된다. 종속 관계는 단어 간의 구문 관계를 나타내지만, 위에서 정의한 종속 구문 트리 패턴은 NPxHead의 개념을 사용하여 명사구 사이의 상위어 관계를 식별할 수 있다. 종속 구문 트리 패턴이 문장과 일치하면 상위어(NPxHead)와 하위어(NPyHead)를 식별합니다.

그런 다음, 상위어 단어와 연관된 NP와 하위어 단어와 연관된 명사구 사이의 상위어 관계가 식별됩니다. 단어와 NPHead가 동일한 단어인 경우, 명사구는 하의어 또는 상위어와 연관된다.

개념적 은유는 감정, 아이디어, 개념 등을 표현하고 전달하는 데 중요한 역할을 한다. 이는 효과적인 의사소통을 위해 일상에서 은유를 사용한다. 은유적 표현을 이해하기 위해서는 다양한 개념에 대한 지식이 필요하다. 개념적 은유가 의미가 있으려면 근원 영역과 목표영역의 두 개념 중 하나가 명시적으로 정의되어야 한다. 근원-목표 쌍을 식별하기 위해서는 은유적 문장을 인식해야 한다. 이를 위해, 우선 개념적 은유의 경우, Is-A 관계의 개념을 파악하는 것이 중요하다. Is-A 패턴이 있는 문장은 잠재적으로 은유 문장이 될 가능성이 크다. Is-A 문장의 첫 번째 명사는 목표영역의 대상이고, 두 번째 명사는 근원 영역의 대상이다. 목표영역 어휘와 근원 영역 어휘를 집합으로 묶은 후, 문장에서 해당 집합에 대한 지식이 있는지 확인한다. 본 논문에서는 6가지 패턴으로 정의된 Hearst Pattern을 확장하여 더욱 풍부한 개념적 은유 패턴을 찾는 방법을 고려하였다. 표 20은 확장된 Hearst Pattern 목록이다.

표 20. Hearst Pattern의 확장된 패턴 정의

분류	패턴	패턴 검색식
기존 패턴	NP_x *such as* NP_y	(NP_\\w+ (,)?such as (NP_\\w+ ?(,)?(and |or)?)+)
	Such NP_x *as* NP_y	(such NP_\\w+ (,)?as (NP_\\w+ ?(,)?(and |or)?)+)
	NP_x *including* NP_y	(NP_\\w+ (,)?include (NP_\\w+ ?(,)?(and |or)?)+)
	NP_y *and | or other* NP_x	((NP_\\w+ ?(,)?)+(and |or)?other NP_\\w+)
	NP_x *especially* NP_y	(NP_\\w+ (,)?especially (NP_\\w+ ?(,)?(and |or)?)+)
	NP_y *is a* NP_x	(NP_\\w+ (,)?is a (NP_\\w+ ?(,)?(and |or)?)+)

확장 패턴	NP_y *and* | *or any other* NP_x	((NP_\\w+ ?(,)?)+(and |or)?any other NP_\\w+)
	NP_y *and* | *or some other* NP_x	((NP_\\w+ ?(,)?)+(and |or)?some other NP_\\w+)
	NP_y *and* | *or be a* NP_x	((NP_\\w+ ?(,)?)+(and |or)?be a NP_\\w+)
	NP_y *and* | *or* NP_x	(NP_\\w+ (,)?like (NP_\\w+ ? (,)?(and |or)?)+)
	such NP_y *as* NP_x *and* | *or*	such (NP_\\w+ (,)?as (NP_\\w+ ? (,)?(and |or)?)+)
	NP_y *and* | *or like other* NP_x	((NP_\\w+ ?(,)?)+(and |or)?like other NP_\\w+)
	NP_y *and* | *or one of the* NP_x	((NP_\\w+ ?(,)?)+(and |or)?one of the NP_\\w+)
	NP_y *and* | *or one of these* NP_x	((NP_\\w+ ?(,)?)+(and |or)?one of these NP_\\w+)
	NP_y *and* | *or one of those* NP_x	((NP_\\w+ ?(,)?)+(and |or)?one of those NP_\\w+)
	example of NP_y *be* NP_x *and* | *or*	example of (NP_\\w+ (,)?be (NP_\\w+ ? (,)?(and |or)?)+)
확장 패턴	NP_y *and* | *or be example of* NP_x	((NP_\\w+ ?(,)?)+(and |or)?be example of NP_\\w+)
	NP_y *for example* NP_x *and* | *or*	(NP_\\w+ (,)?for example (,)?(NP_\\w+ ?(,)?(and |or)?)+)
	NP_y *and* | *or one which be call* NP_x	((NP_\\w+ ?(,)?)+(and |or)?which be call NP_\\w+)
	NP_y *principally* NP_x *and* | *or*	(NP_\\w+ (,)?principally (NP_\\w+ ? (,)?(and |or)?)+)
	NP_y *type* NP_x *and* | *or one*	(NP_\\w+ (,)?type (NP_\\w+ ? (,)?(and |or)?)+)
	NP_y *and* | *or* NP_x *type*	((NP_\\w+ ?(,)?)+(and |or)? NP_\\w+ type)
	NP_y *whether* NP_x *and* | *or*	(NP_\\w+ (,)?whether (NP_\\w+ ? (,)?(and |or)?)+)
	compare NP_y *and* | *or one with* NP_x	(compare (NP_\\w+ ?(,)?)+(and |or)?with NP_\\w+)
	NP_y *compare to* NP_x *and* | *or*	(NP_\\w+ (,)?compare to (NP_\\w+ ? (,)?(and |or)?)+)
	NP_y NP_x *and* | *or for instance*	(NP_\\w+ (,)? (NP_\\w+ ? (,)?(and |or)?)for instance)

본 장에서는 개념적 은유의 근원-목표영역 개념과 관계 추출을 위해 Top2Vec 모델을 기반으로 개념 후보를 추출하고, 개념적 은유 패턴을 기반 의미 관계를 정의하고, 이를 모델링 하였다. 또한, 근원-목표영역의 개념 후보 선정을 위해 상위어 인식을 위한 Hearst Pattern의 확장된 패턴을 정의하였다. 개념 후보와 의미 관계 추출 방법을 이용하여 상위어 후보 추출을 수행하였다. CNN 뉴스 데이터를 Top2Vec 알고리즘을 통해 학습을 시킨 후, 해당 모델을 이용하여 'Money'와 연관된 단어는 다음 표 21과 같이 분류되었다.

표 21. 'Money' 연관단어 순위

money, tax, price, pay, income, taxis, dollar, bank, currency, profit
go, say, get, tell, friend, ask, girl, woman, love, meet
leave, return, back, go, send, decide, eventually, bring, soon, come
make, although, even, though, much, however, still, less, remain, often
take, give, place, player, hand, position, order, name, play, allow
letter, deliver, express, scandal, trial, public, accuse, private, dismiss, publicly
company, corporation, system, agency, service, local, government, business, management, public
political, support, policy, issue, concern, public, campaign, make, criticism, seek
game, player, run, deal, baseball, card, home, pitcher, season, hit
attack, kill, police, murder, victim, crime, death, cause, arrest, result
film, good, award, receive, win, awards, academy, critic, review, actor
win, world, cup, play, championship, olympic, team, event, club, since
write, john, guitar, band, author, ray, mark, book, drummer, guitarist
however, attempt, result, despite, fail, successful, prove, due, fall, continue
law, court, case, legal, act, right, jurisdiction, amendment, supreme, judge
car, company, sell, manufacturer, model, brand, market, sale, product, ford

그림 18. 'Money' 연관단어 그룹

각 임베딩 차원에 대한 관련어이고, 그림 18과 같이 "Money'와 의미적으로 가장 가까운 그룹이 가장 높은 순으로 나타났다. 다음 그림 19는 'Book'과 연관된 유사 단어 그룹핑 결과이다.

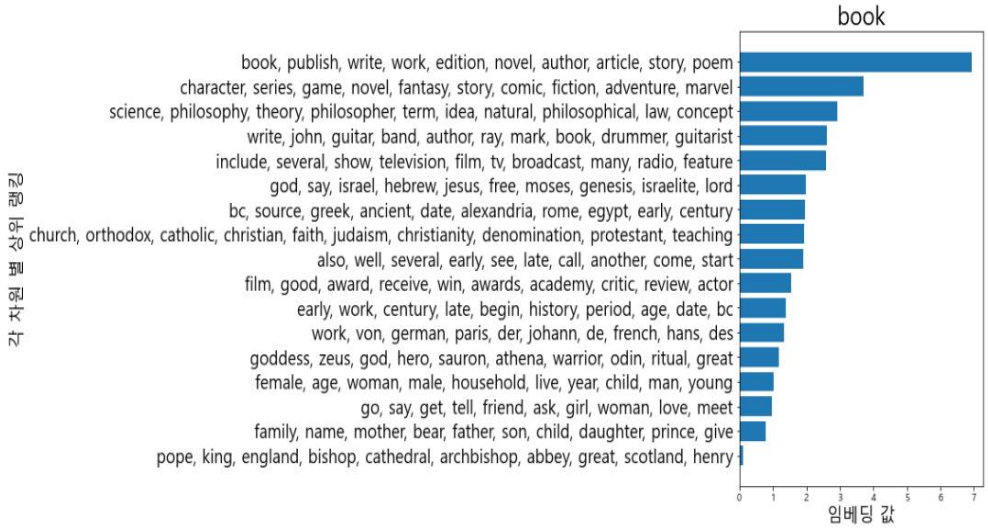


그림 19. 'Book' 연관단어 그룹

개념적 은유의 근원-목표영역의 파악하기 위해 두 영역에 동시에 출현하는 상위어 관련 어휘를 그룹핑 할 필요가 있다. 예를 들어, “Love”과 “Waepon”이 동시에 출현했을 때, 관련 어휘를 그룹한 결과는 다음 그림 20과 같다.

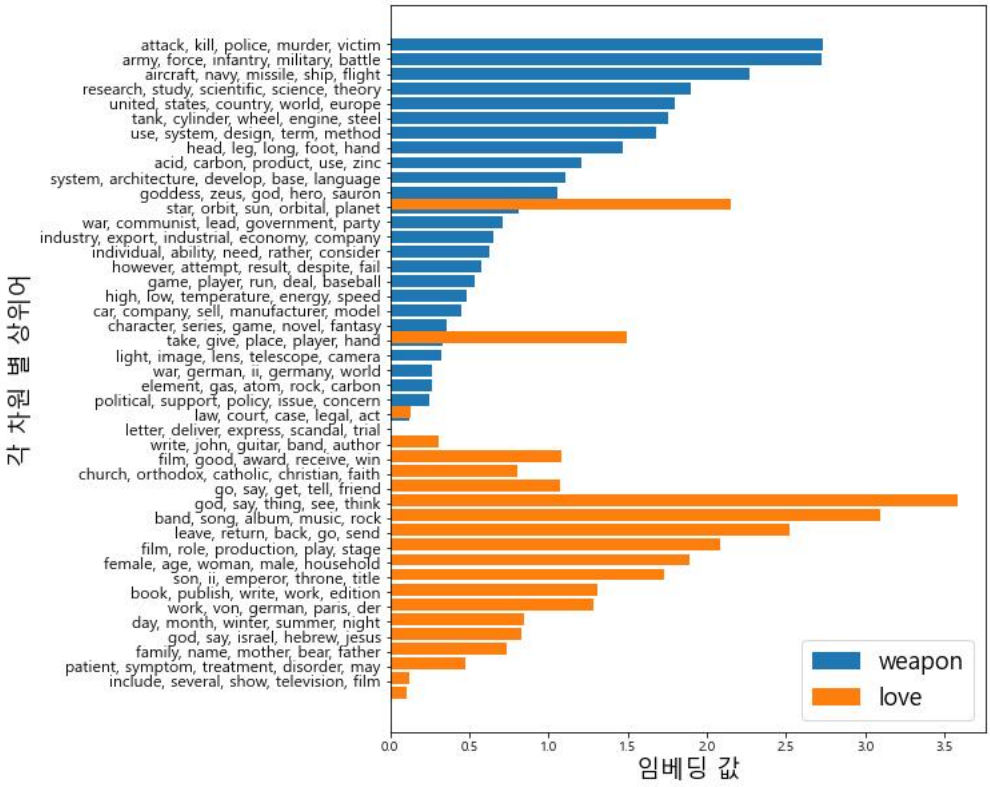


그림 20. “Love” & “Waepon” 관련 어휘 그룹

Ⅳ. 실험 및 결과

A. 실험 환경

본 연구에서 제안하는 개념적 은유에 대한 근원-목표 영역 개념 추출 방법을 성능 평가 하기 위해 CNN 뉴스 코퍼스를 사용하였다. 미국의 18개 미국 간행물에서 204,135개의 뉴스 데이터 중 CNN 뉴스를 활용하였으며, 날짜, 제목, 출판, 기사 텍스트, 출판 이름, 연도, 월 및 URL 등이 포함된 데이터로, 2013년부터 2018년 초까지의 뉴스 기사이고, 14,300개의 레코드를 포함하고 있다. 본 연구에서 사용한 Top2Vec 임베딩 모델의 성능 비교를 위해 Word2Vec 임베딩 모델을 함께 사용하였다. Word2vec과 Top2Vec 모델은 성능평가를 위해 학습 파라미터를 표 22와 같이 설정하였다. 벡터 차원(vector_size)은 300, 전체 단어 빈도수(min_count)는 10, 윈도우 크기(window)는 15로 설정한다.

표 22. 임베딩 모델 학습 파라미터

	Top2Vec	Word2Vec
vector_size	300	300
min_count	10	10
window	15	15
dbow_words	1	

뉴스 코퍼스를 전처리한 후, Top2Vec 모델을 이용하여 근원-목표 영역 개념 추출 실험을 진행했고, 정확도와 정밀도, 재현율, F-measure의 3가지 성과지표 통해 개념 추출 성능을 확인하였다.

B. 실험 평가 및 분석

1. 개념적 은유의 근원-목표 영역 분류 및 분석

개념적 은유 표현은 개념적 은유 이론에 의해 근원-목표 영역으로 분류하였다. 근원-목표 영역의 분석 결과는 다음과 같다. CNN 뉴스 데이터에서 근원-목표 영역에 대해 은유적으로 사용되는 관련 어휘의 연관 유사도와 인식된 개념적 은유 표현을 보여준다. 사용된 개념적 은유는 “IDEAS ARE FOOD”, “TIME IS A RESOURCE”, “PEOPLE ARE PLANTS” 등이다.

표 23. “IDEAS ARE FOOD”의 개념적 은유 근원-목표 영역 분석

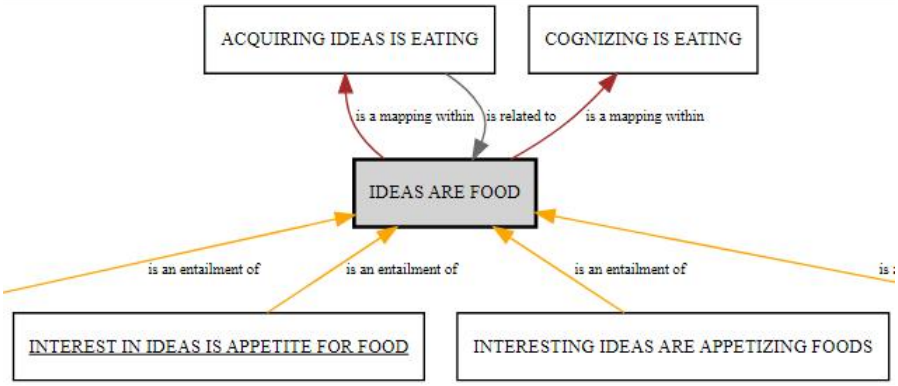
개념적 은유	은유적 표현 관련 어휘	개념적 은유 표현
IDEAS ARE FOOD	get 0.5962967662993937 it 0.5836464414359757 find 0.5763943333090229 how 0.5725077339379737 things 0.5720267378208094 like 0.571893371535049 is 0.5694106928625331 to 0.5655388468771991 says 0.5612815471211399 can 0.5606834242057072	His idea was half-baked. Let me chew on that for a while. They ate the lesson up. They gobbled up the ideas. He has an appetite for learning. The teacher spoon-fed them the information. I'm tired of warmed-over theories.
근원영역 : FOOD	eat 0.6143244436428037 foods 0.5634943962298745 eating 0.5249083969410939 vegetables 0.5157378944019795 products 0.5029946439826574 supplies 0.500249221582902 fruits 0.48907718619433455 nutrition 0.4859972590577767 agriculture 0.485996432452639 milk 0.46975952540917715	
목표영역 : IDEAS	idea 0.5107648271883313 own 0.45010869646279383 views 0.44031604390155055 create 0.43499572617730325 speech 0.4346443519317601 ways 0.4340766375197263 different 0.42947683301157924 think 0.42498797024871726 free 0.4240241007680632 rather 0.42110138837808914	
		

표 24. “TIME IS A RESOURCE”의 개념적 은유 근원-목표 영역 분석

개념적 은유	은유적 표현 관련 어휘	개념적 은유 표현
TIME IS A RESOURCE	and 0.6314585619611659 of 0.6261495601430365 for 0.6236782078447346 to 0.6196632497081049 is 0.610812527945912 in 0.5998618958304833 it 0.5969547275345501 an 0.5950180526748379 the 0.58817311683795 much 0.5864446951067903	We’re almost out of time. Don’t waste time. Find a better use for your time.
근원영역 : RESOURCE	precious 0.3494246351464284 development 0.29500059084991626 create 0.2946568502472901 resources 0.2906898503030428 projects 0.28960083850838086 processes 0.2841606342016361 project 0.28216184878514666 important 0.2819566836049729 counseling 0.277879558291306 educate 0.27500924835863066	
목표영역 : TIME	on 0.7485106242357856 for 0.7450370096581368 the 0.7442813069240894 to 0.7429406893365893 and 0.7407082150211379 when 0.7384875072710683 in 0.7305656318509227 it 0.7258407184058531 at 0.7233439887733562 with 0.7195666967058245	

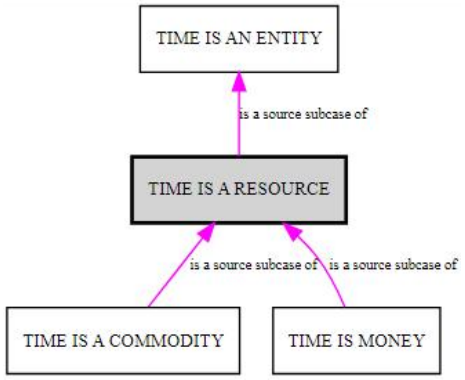
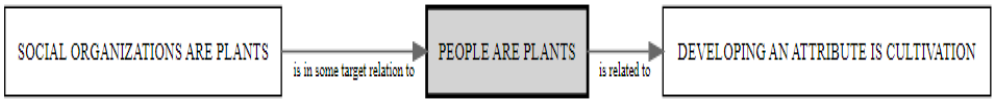


표 25. “PEOPLE ARE PLANTS”의 개념적 은유 근원-목표 영역 분석

<table>
<tr><th>개념적 은유</th><th>은유적 표현 관련 어휘</th><th>개념적 은유 표현</th></tr>
<tr><td>PEOPLE ARE PLANTS</td><td>more 0.6095884636975485
are 0.5959431371759449
of 0.5786572086911452
and 0.5680438756382545
many 0.5664513768861393
this 0.5626138858227537
than 0.5534329590693381
others 0.5514679208972217
these 0.5469491181032613
some 0.5457680059324179</td><td>She’s a late bloomer.
She is in the flower of youth.
She’s past her bloom.
She’s let herself go to seed.</td></tr>
<tr><td>근원영역 : PLANTS</td><td colspan="2">plant 0.5733670220433066
epa 0.47069534759439424
environmental 0.4558289692555519
water 0.4499712850524724
pollution 0.4325642241592258
chromium 0.4217264492097332
natural 0.4164462556732369
clean 0.41359380201311235
emissions 0.4111989497943047
greenhouse 0.4110972569298462</td></tr>
<tr><td>목표영역 : PEOPLE</td><td colspan="2">others 0.7177884819710876
are 0.6974556830427747
many 0.6956668427741624
those 0.6774201735034223
of 0.6693584888514164
there 0.6613428593197559
this 0.6612733573213913
more 0.6578044208880153
and 0.6566965562165259
where 0.6448989922472396</td></tr>
<tr><td colspan="3"> 
</td></tr>
</table>

본 연구에서는 CNN 뉴스 코퍼스의 개념적 은유에 대한 근원-목표 영역 개념을 추출하기 위해 사용한 Top2Vec 임베딩 모델의 성능 비교를 위해 Word2Vec

임베딩 모델을 함께 사용하였다. 뉴스 코퍼스를 전처리한 후, Top2Vec 모델을 이용하여 근원-목표 영역 개념 추출 실험을 진행했고, 정확도와 정밀도, 재현율, F-measure의 3가지 성과지표 통해 개념 추출 성능을 확인하였다. 표 22는 Top2 Vec 모델을 이용한 개념 추출 검증 결과를 보여주고 있다. 사용된 개념적 은유는 “IDEAS ARE FOOD”, “LOVE IS A JOURNEY”, “TIME IS A RESOURCE”, “PEOPLE ARE PLANTS”, “CHANGE IS MOTION” 등이다. 정확도를 평가하는 방법으로는 정확률, 재현율, 정확률과 재현율을 합한 단위인 신뢰도를 이용한다. 정확률은 전체 추출된 값 중에서 정확하게 판별한 비율을 나타내며, 재현율은 모 집단의 모든 데이터 중 정확히 판별한 비율을 나타낸다. F-measure는 추출 결과에 대한 신뢰도를 의미하며 수식 1과 같이 정의한다.

$$F-measure = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \qquad (1)$$

표 26. 목표 영역 개념 관련 추출 검증 결과

개념 영역	임베딩 모델	성능지표		
		Precision	Recall	F-measure
IDEAS	Word2Vec	0.68134	0.69013	0.68571
	Top2Vec	0.75123	0.76456	0.75784
LOVE	Word2Vec	0.69134	0.78534	0.73535
	Top2Vec	0.70213	0.71846	0.71020
TIME	Word2Vec	0.68321	0.71345	0.69800
	Top2Vec	0.69013	0.73145	0.72019
PEOPLE	Word2Vec	0.69067	0.70451	0.69752
	Top2Vec	0.72037	0.78312	0.75044
CHANGE	Word2Vec	0.69845	0.70984	0.70410
	Top2Vec	0.71863	0.73459	0.72652
평균	Word2Vec	0.68900	0.72065	0.70414
	Top2Vec	0.71650	0.74644	0.73104

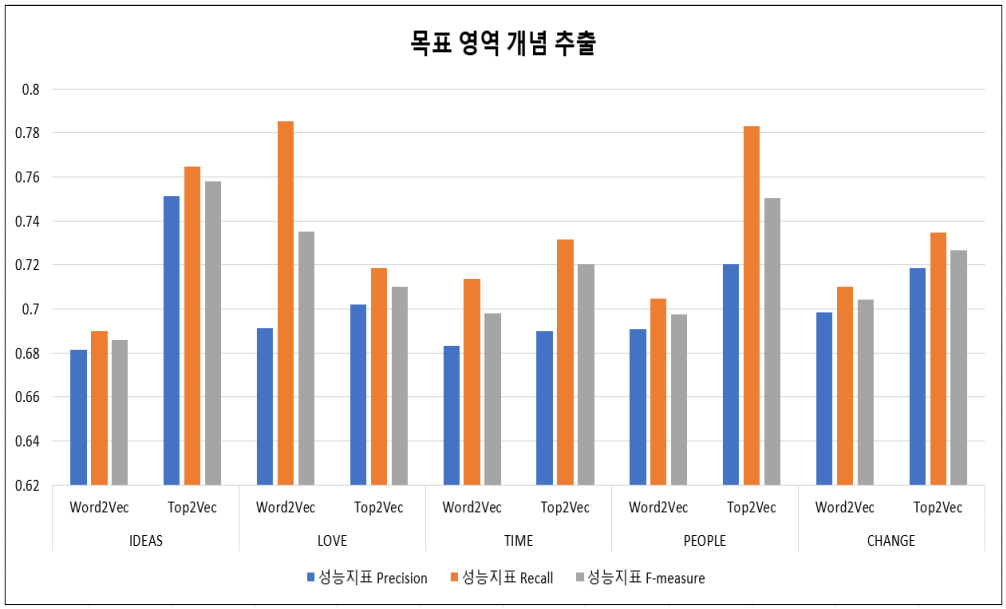


그림 21. 목표 영역 개념 추출 성능 평가

표 27. 근원 영역 개념 관련 추출 검증 결과

개념 영역	임베딩 모델	성능지표		
		Precision	Recall	F1-Score
FOOD	Word2Vec	0.56451	0.57453	0.56948
	Top2Vec	0.58340	0.59013	0.58675
JOURNEY	Word2Vec	0.62459	0.63412	0.62932
	Top2Vec	0.61934	0.65345	0.63594
RESOURCE	Word2Vec	0.59084	0.61634	0.60332
	Top2Vec	0.59532	0.61450	0.60476
PLANTS	Word2Vec	0.58742	0.59176	0.58958
	Top2Vec	0.61745	0.64368	0.63029
MOTION	Word2Vec	0.59851	0.61873	0.60845
	Top2Vec	0.63012	0.65715	0.64335
평균	Word2Vec	0.59317	0.60710	0.60003
	Top2Vec	0.60913	0.63178	0.62022

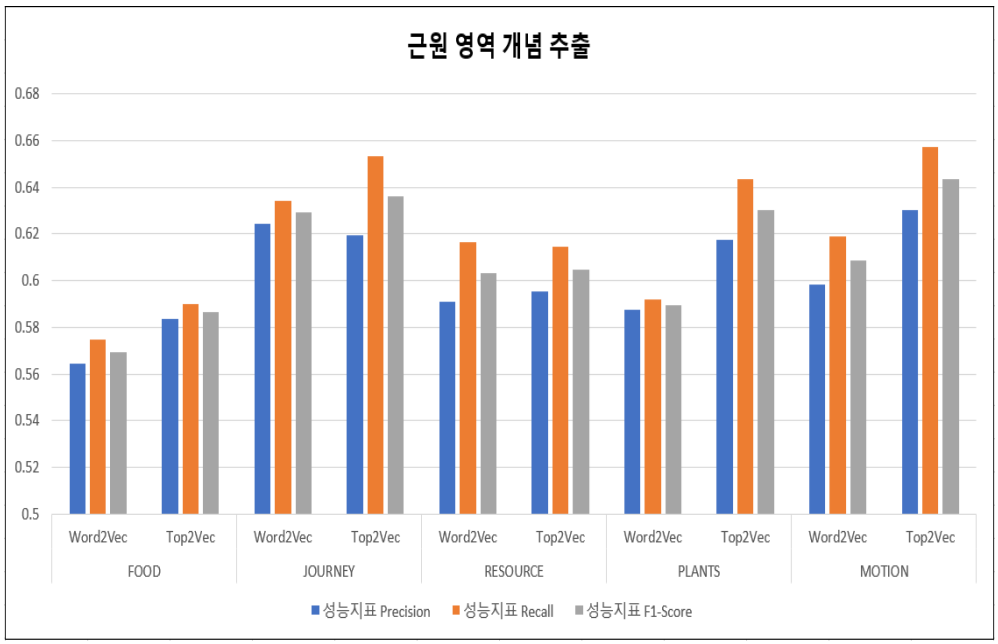


그림 22. 근원 영역 개념 추출 성능평가

표 26, 27에서 CNN 뉴스 코퍼스를 대상으로 목표 영역 개념 추출을 수행했을 때 Word2vec 모델의 F-measure는 0.70414, Top2Vec 모델의 정확도는 0.73104로 To2Vec 모델 성능이 높은 것을 확인할 수 있다. F1-score 기준으로 근원 영역을 테스트했을 때, Word2vec을 적용한 모델에서는 0.60003, Top2Vec 모델은 0.62022로 나타났다. 또한, 표 28과 같이 개념적 은유에 대한 근원-목표 영역 개념을 동시에 추출을 수행했을 때도 Word2vec 모델의 F1-score는 0.61041, Top2Vec 모델은 0.71332로 Top2Vec 모델이 개념적 은유의 근원-목표 영역의 개념 추출에 있어서 더 좋은 성능을 보인다.

표 28. 근원-목표 영역 개념 동시 추출 결과

개념적 은유	임베딩 모델	성능지표		
		Precision	Recall	F1-Score
IDEAS ARE FOOD	Word2Vec	0.58134	0.59873	0.58991
	Top2Vec	0.68132	0.70845	0.69462
LOVE IS A JOURNEY	Word2Vec	0.60341	0.62421	0.61363
	Top2Vec	0.71276	0.73564	0.72402
TIME IS A RESOURCE	Word2Vec	0.57345	0.59834	0.58563
	Top2Vec	0.69913	0.70789	0.70.348
PEOPLE ARE PLANTS	Word2Vec	0.63445	0.64216	0.63828
	Top2Vec	0.70345	0.71456	0.70896
CHANGE IS MOTION	Word2Vec	0.61388	0.63566	0.62458
	Top2Vec	0.72567	0.74561	0.73550
평균	Word2Vec	0.60131	0.61982	0.61041
	Top2Vec	0.70447	0.72243	0.71332

Ⅴ. 결론 및 향후 연구

본 연구는 개념적 은유의 후보군을 선정하기 위해 근원-목표영역 벡터 관계 모델링 방법을 제안하였다. 개념과 관계 추출을 위해 토픽 모델링 방법으로 Top2Vec모델을 사용하여 개념 후보를 추출하고, 개념적 은유 패턴을 기반 의미 관계를 정의하고, 이를 모델링 하였다. 또한, 근원-목표영역의 개념 후보 선정을 위해 상위어 인식을 위한 Hearst Pattern의 확장된 패턴을 정의하여 개념 후보와 의미 관계 추출 방법을 이용하여 상위어 후보 추출을 수행하였다.

제안된 모델에서 근원-목표영역의 개념 추출 과정은 다음과 같다. 첫째, 토픽 모델 중 하나인 Top2Vec 모델을 적용하여 문서 내 단어들의 핵심 개념 어휘와 의미상으로 유사한 관련 어휘를 추출한다. 둘째, 근원-목표영역의 개념 후보 선정을 위해 핵심 개념 어휘를 중심으로 상위어를 선정한다. 셋째, 상위어 후보 어휘를 중심으로 'is-a' 관계를 추출한다. 넷째, 추출된 근원-목표영역을 기반으로 문서 내 개념적 은유를 선정하고, 이에 대한 정량적 비교평가를 수행한다.

본 연구에서 제안한 개념적 은유에 대한 근원-목표 영역 개념 추출 방법의 성능 평가를 Word2Vec 임베딩 모델을 함께 사용하였다. CNN 뉴스 코퍼스를 전처리한 후, Top2Vec 모델을 이용하여 근원-목표 영역 개념 추출 실험을 진행하였고, 정확도와 정밀도, 재현율, F-measure의 3가지 성과지표 통해 개념 추출 성능을 비교하였다.

본 연구 결과 CNN 뉴스 코퍼스를 대상으로 목표 영역 개념 추출, 근원 영역 개념 추출 및 개념적 은유에 대한 근원-목표 영역 개념을 동시에 추출을 수행했을 때 Word2Vec을 적용한 모델보다 본 논문에서 사용한 Top2Vec 모델의 성능이 더 좋은 것을 확인할 수 있었다.

본 논문은 개념적 은유의 후보 선정을 위해 근원-목표영역의 핵심 어휘와 관련된

어휘를 추출하고 은유적 표현의 패턴을 추출하여 개념적 은유 후보군을 선정하였다. 개념적 은유 및 은유적 표현의 패턴 추출의 체계가 확립된다면 일반적인 어휘가 가지고 있는 중의성 문제를 해결하는 것보다 훨씬 높은 수준의 중의성 문제를 해결할 수 있을 것으로 판단된다.

이는 챗봇(ChatBot), 자동 번역 등과 같은 인공지능 기술의 응용 분야에 활용하여 자연스러운 대화체 구현을 가능케 하는 기반 기술이라고 판단된다.

참 고 문 헌

[1] 김상진. “정치담론에 나타난 개념적 은유 연구.” 부산대학교 석사학위논문, 2018.

[2] Lakoff, George·Mark, Johnson(1980), 『삶으로서의 은유』, 노양진 · 나익주 역 (2006),□박이정.

[3] Lakoff, George(2006), 『프레임 전쟁 : 보수에 맞서는 진보의 성공전략』, 나익주 역(2007), 창비.

[4] 김고은. “길의 문화적 의미 연구 - 개념적 은유 이론을 중심으로.” 전남대학교 석사학위논문, 2012.

[5] 임지룡. “개념적 은유에 대하여. 한국어 의미학,” 20, pp. 29-60, 2006.

[6] 권연진. “미국 대통령 취임사에 나타난 정치는 여행이다 은유에 관한 연구.” 코기토(85), pp. 291-318, 2018.

[7] 김용. “개념적 은유 기반의 한국어 동사 의미 교육 연구. ” 서울대학교 박사학위논문, 2015.

[8] 정상원. "사물 명칭의 환유적 확장 연구." 조선대학교 석사학위논문, 2020.

[9] 이상윤. "토픽모델링 적용 학술 검색 엔진 검색 품질 향상." 서강대학교 석사학위논문, 2020.

[10] 백시온. "한국어 토픽모델링을 위한 단어 임베딩 활용 가능성 탐색." 서울대학교 석사학위논문, 2019.

[11] 남춘호. "일기자료 연구에서 토픽모델링 기법의 활용가능성 검토." 비교문화연구 22(1), pp. 89-135, 2016.

[12] 남현동, 남태우. “한국 플랫폼 정부의 방향성 모색 : 공공기관 연구보고서에 대한 토픽 모델링과 네트워크 분석.” 디지털융복합연구, 18(2), pp. 139-149, 2020.

[13] 이치욱. "CNN-LSTM 모델을 이용한 이슈 관리 시스템의 모듈 분류 연구." 서강대학교 석사학위논문, 2017.

[14] 김도우. "Doc2Vec을 활용한 CNN 기반 한국어 신문 기사 분류에 관한 연구."

서강대학교 석사학위논문, 2017.

[15] 조희석. "Doc2Vec 기반 질의응답 검색 시스템 개발." 한국방송통신대학교 석사학위논문, 2019.

[16] 이지성. "텍스트 내용 기반 추천을 위한 토픽추출의 영향에 대한 연구." 한양대학교 석사학위논문, 2020.

[17] Le, Q., & Mikolov, T. "Distributed representations of sentences and documents", In International conference on machine learning, pp. 1188-1196, 2014.

[18] Dimo Angelov, "Top2Vec : Distributed Representations of Topics", arXiv:2008.09470v1 [cs.CL], pp. 1-25, 2020.

[19] https://edition.cnn.com

[20] 이찬영, 김진웅, 김한샘. "Universal Dependency 관계 태그셋의 한국어 적용." 제30회 한글 및 한국어 정보처리 학술대회 논문집, pp. 334-339, 2018.

[21] 방찬성, & 이해윤. (2008). 코퍼스를 이용한 상하위어 추출 연구. 인지과학, 19(2), pp. 143-161.

[22] Hassan, J., & Munib, M. (2010). Detecting Missing IS-A Relations in Ontologies.

[23] 권연진. "정치담론 상에 나타난 은유의 대조 분석 연구." 인문연구 82, pp. 33-62, 2018.

[24] 김종수. "독일 신문에 투영된 정치언어의 은유." 독어교육, 72, pp. 95-118, 2018.

[25] 김진해. "개념적 은유의 보편성과 특수성." 한국어 의미학, 46, pp. 331-349, 2014.

[26] 김태헌. "개념적 은유와 문화적 맥락: 영어와 한국어의 개념적 사랑 은유를 중심으로." 언어학, 16(1), pp. 129-150, 2008.

[27] 류영미. "은유적 표현의 유형과 보편성." 국내석사학위논문 경북대학교 교육대학원, 2006.

[28] 박현규. “뉴로 심볼릭 기반 규칙 추론을 통한 자동 지식 완성.” 숭실대학교 박사학위논문, 2021.

[29] 배승호. “정치와 관련된 은유 표현.” 한국어 의미학, 8, pp. 261-277, 2011.

[30] 백미현. “개념적 은유 [마음은 몸] 연구.” 담화와인지, 27(2), pp. 51-72, 2020.

[31] 심보준, 고영중, 김학수, & 서정연. “자연어 질의응답 시스템을 위한 is-a 관계 패턴의 구축과 활용”. 한국정보과학회 언어공학연구회 학술발표 논문집, 16(1), pp. 181-188, 2004.

[32] 서미지. “텍스트 데이터를 활용한 감정 패턴 분석에 관한 연구 - 대통령 담화문을 중심으로.” 한양대학교 석사학위논문, 2019.

[33] 양나영. “한중 음식관련어 개념적 은유 대조 연구.” 한국외국어대학교 석사학위논문, 2011.

[34] 염철호. “시편 23장의 개념적 은유.” 가톨릭신학, 20, pp. 5-34, 2012.

[35] 오은비. “정치 은유와 환유 연구 - 정치 기사를 중심으로.” 가천대학교 석사학위논문, 2019.

[36] 우창우. ″토픽 모델과 단어 임베딩 모델을 사용한 문헌의 핵심 토픽 및 키워드 탐색 프레임워크 설계.″ 충북대학교 박사학위논문, 2021.

[37] 유군. “날씨 관련 은유 표현의 한중 대조 연구.” 한양대학교 박사학위논문, 2018.

[38] 윤희근. “지식 그래프 확장을 위한 파스 트리 기반 트리플 추출과 논리 속성 보존 임베딩.” 경북대학교 박사학위논문, 2021.

[39] 이민우. “개념적 은유의 특수성.” 한국어 의미학, 57, pp. 1-19, 2017.

[40] 이영민. “리뷰 텍스트를 활용한 토픽별 키워드 기반 시맨틱 POI 검색.” 서울대학교 박사학위논문, 2021.

[41] 이유교. “경제 기사에 나타난 개념적 은유와 환유 연구.” 전남대학교 석사학위논문, 2019.

[42] 임지룡, 임혜원. “연결 도식과 그 은유적 확장.” 한글(276), pp. 105-132, 2007.

[43] 전혜영. “한국어 표현에 나타나는 여행 은유.” 이화어문논집, 33, pp. 75-103, 2014.

[44] 조아연. “애니메이션에 있어서 시적표현의 확장으로서 은유연구 - 개념적 은유를 중심으로.” 홍익대학교 석사학위논문, 2004.

[45] 조유진. “한국과 미국의 정치담론에 나타난 개념적 은유 연구.” 부산대학교 석사학위논문, 2019.

[46] 주려빈. “한·중 개념적 은유 표현 대조 연구.” 경희대학교 박사학위논문, 2019.

[47] 진쌍쌍. “경제 기사문에 나타난 은유 표현의 중한 번역 사례 연구-개념적은유를 중심으로.” 한국외국어대학 석사학위논문, 2020.

[48] 최윤영. “성경에서의 개념적 은유와 환유.” 신학과 목회, 45, pp. 329-351, 2016.

[49] 최현희. “The Great Gatsby 영한번역에 나타난 개념적 은유의 사상 대응에 관한 연구.” 부산대학교 석사학위논문, 2014.

[50] 최재영, 김태호. “개념적 은유의 영역(domain)과 사상(mapping)에 대한 소고.” 담화인지언어학회 학술대회 발표 논문집, pp. 73-80, 2016.

[51] 한혜원, 문아름. “소셜 네트워크 서비스의 은유적 특성 연구.” 디지털콘텐츠학회 논문지, 15(5), pp. 621-630, 2014.

[52] MA CONGCONG. “한·중 외교 연설문에 나타난 개념적 은유 대비 연구.” 이화여자대학교 석사학위논문, 2019.

[53] QIU YUNHONG. “한·중 공익광고에 나타난 개념적 은유 양상에 대한 대비 연구.” 이화여자대학교 석사학위논문, 2017.

[54] Batselem Jagvaral. “A Deep Learning Approach for Predicting Missing Links in Knowledge Graphs.” 숭실대학교 박사학위논문, 2021.

[55] Bracewell, D. B., Tomlinson, M. T., Mohler, M., & Rink, B., “A tiered approach to the recognition of metaphor. In International Conference on Intelligent Text Processing and Computational Linguistics”, Springer, Berlin, Heidelberg, pp. 403-414, 2014.

[56] Ellen Dodge, JisupHong, Elise Stickles. “MetaNet: Deep semantic automatic metaphor analysis”, Proceedings of the Third Workshop on Metaphor in NLP, pp. 40-49, 2015.

[57] Eog han Macg uire and Steve Visser. “7 Russian swimmers banned from Rio amid doping scandal”, CNN, 2016.

[58] Eric P. S. Baumer, James P. White, Bill Tomlinson. “Comparing semantic role labeling with typed dependency parsing in computational metaphor identification”, Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity, pp. 14 - 22, 2010.

[59] Farah, T., Afida, M. A., Rafik-Galea, S., & Zalina, M. K. “Getting Physical with the Market: A Study of Metaphors in the Business Times”, Editorial Board, 881, 2012.

[60] George Lakoff, Jane Espenson, and Adele Goldberg. August. “Master Metaphor List”, First Edition Compiled By. Second Edition Compiled By, 1989.

[61] Golshaie, R. “Searching for Cross-Domain Mappings in the Corpus: an Analysis of Conceptual Metaphors’ Usage Patterns in Farsi”, The International Journal of Humanities, 26(2), pp. 14-28, 2019.

[62] Heintz, I., Gabbard, R., Srivastava, M., Barner, D., Black, D., Friedman, M., & Weischedel, R. “Automatic extraction of linguistic metaphors with lda topic modeling”, In Proceedings of the First Workshop on Metaphor in NLP, pp. 58-66, June 2013.

[63] Kevin Stowe, TuhinChakrabarty, NanyunPeng, SmaramdaMuresan, Iryna Gurevych. “Metaphor Generation with Conceptual Mapping”, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 6724 - 6736, 2021.

[64] Li, X. “Unsupervised Extraction and Clustering of Key Phrases from Scientific Publications”, 2020.

[65] Michael Mohler, Bryan Rink, David Bracewell, Marc Tomlinson. “A Novel Distributional Approach to Multilingual Conceptual Metaphor Recognition”, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1752-1763, 2014.

[66] Mohler, M., Rink, B., Bracewell, D., & Tomlinson, M. “A novel distributional approach to multilingual conceptual metaphor recognition”, In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1752-1763, August, 2014.

[67] Mohler, M., Brunson, M., Rink, B., & Tomlinson, M. “Introducing the lcc metaphor datasets”, In Proceedings of the Tenth International Conference on Language Resources and Evaluation(LREC'16), pp. 4221-4227, May, 2016.

[68] Naicker, S. “A cognitive linguistic analysis of conceptual metaphors in Hindu religious discourse with reference to Swami Vivekananda’s complete works”, PhD diss., University of South Africa. http://hdl. handle. net/10500/22281, 2016.

[69] Nayak, T. “Deep neural networks for relation extraction”, arXiv preprint arXiv:2104.01799, 2021.

[70] Parde, N., & Nielsen, R. “A corpus of metaphor novelty scores for syntactically-related word pairs”, In Proceedings of the Eleventh International Conference on Language Resources and Evaluation(LREC 2018), May 2018.

[71] Pathak, J., & Shah, P. “Markov Logic Network for Metaphor Set Expansion”, In ICAART (2), pp. 621-628, 2021.

[72] Shaikh, S., Strzalkowski, T., Cho, K., Liu, T., Broadwell, G. A., Feldman, L., & Elliot, K. “Discovering Conceptual Metaphors using Source

Domain Spaces", State University of New York University at Albany Albany United States, 2014.

[73] Shanavas, N., Wang, H., Lin, Z., & Hawe, G. "Knowledge-driven graph similarity for text classification", International Journal of Machine Learning and Cybernetics, 12(4), pp. 1067-1081, 2021.

[74] Shutova, E., Sun, L., Gutiérrez, E. D., Lichtenstein, P., & Narayanan, S. "Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning", Computational Linguistics, 43(1), pp. 71-123, 2017.

[75] Shutova, E., & Teufel, S. "Metaphor Corpus Annotated for Source-Target Domain Mappings", In LREC, 2(2), pp. 2-2, May, 2010.

[76] Stowe, K., Chakrabarty, T., Peng, N., Muresan, S., & Gurevych, I. "Metaphor Generation with Conceptual Mappings", arXiv preprint arXiv:2106.01228, 2021.

[77] Steve Almasy. "Rio 2016 : Sunday was big day for athletic and g ymnast g reats", CNN, 2016.

[78] Tendahl, Markus. "A Hybrid Theory of Metaphor Volume 13 ||", Palgrave Macmillan, 2009.

[79] Tretjakova, J. "AN INSIGHT INTO CONTEMPORARY THEORY OF METAPHOR".

[80] Yuxun, L. "Improving Knowledge Graph Completion Models by Unsupervised Type Constraint Inference", 2019.

[81] Zayed, O., McCrae, J. P., & Buitelaar, P. "Contextual modulation for relation-level metaphor identification", arXiv preprint arXiv: 2010.05633, 2020.

[82] Zayed, O. "Metaphor Processing in Tweets", Doctoral dissertation, College of Science and Engineering, National University of Ireland Galway, 2021.

[83] Zohrabi, M., & Layegh, N. “Bidirectionality of Metaphor in Fiction: A Study of English Novels”, Applied Linguistics Research Journal, 4(4), pp. 88-99, 2020.

[84] Zhu, G., & Iglesias, C. A. “Computing semantic similarity of concepts in knowledge graphs”, IEEE Transactions on Knowledge and Data Engineering, 29(1), pp. 72-85, 2016.

[85] https://www.arxiv-vanity.com/papers/2008.09470/

[86] https://components.one/datasets/all-the-news-articles-dataset/

감사의 글

돌아보면 지난 시간들은 '학문에 대한 열의만큼이나 뒤엉킴의 시간'이었습니다. 어디서부터가 시작인지 무엇을 어떻게 해 나아가야 할지에 대한 고민의 시간이었고, 학자로 거듭나기 위한 준비의 시간이었고, 지도교수님을 비롯한 많은 교수님과 박사님들 그리고 선행연구자들이 걸었던 길을 따라 한걸음 한걸음 나아가는 배움의 시간이었습니다.

그런 소중한 시간들이 모여 결실의 열매를 보게 되어 감사하고 또 혼자서는 결코 올 수 없었던 여정이었기에 감사함을 마음 깊이 새겨 잊지 않겠습니다. 박사 학위과정을 마친 것에 대해 후련함과 감격스러움보다는 박사로서 갖추어야 할 품성과 자질 그리고 자기 발전을 위한 끊임 없는 노력으로 감사함에 보답하겠습니다.

유연(柔軟)함과 포용력(包容力)을 갖고 연구자의 자세를 견지하며, 무지(無知)함에 부끄러워하기보다는 더 많은 것을 알아가기 위해 노력하겠습니다.

제자 사랑의 열정으로, 보잘것없는 저를 학문적 소양과 연구 그리고 논문까지 어느 하나 부족함이 없이 지도해주신 김판구 지도교수님, 좋은 논문을 만들어 주시기 위하여 아낌없는 조언과 격려로 세심하게 지도해주신 양희덕 교수님, 학문하는 자가 갖춰야 할 도전, 노력, 인내와 더불어 연구 방법을 함께 고민해 주시고 지도해 주신 최준호 교수님, 연구 방법과 학문적 원칙을 바탕으로 연구자가 편향적사고(偏向的思考)를 갖지 않도록 아낌없는 조언과 가르침을 주신 황명권 교수님, 바쁘신 일정에도 불구하고 먼 길 마다하지 않으시며 귀중한 시간과 정성으로 논문을 다듬고 또 다듬어 부족함을 채워주신 최창 교수님께 진심으로 감사드립니다.

저의 박사과정 동안 연구실 생활에 잘 적응하며 연구 수행 및 프로젝트 수행을 어렵지 않게 하도록 도와준 연구실원들이 있습니다. 같이 졸업논문을 준비하며 연구가 잘 풀리지 않고 연구 및 프로젝트 수행에 어려운 상황이 오면 늘 상의하고 문제를 같이 해결하며 서로 의지한 홍택은, 연구실의 막내이자 박사과정으로 앞으로 연구실의 다양한 업무를 수행해야 할 후배 유경호에게 고마움을 전합니다.

나이 사십이 넘은 딸이 다시 박사 학위과정을 하고자 한다고 말씀드렸을 때 얼마나 큰 걱정과 근심이 가슴 가득하실까? 라는 걱정이 앞섰지만, 의연(依然)히 큰 나무가 되어주시며 아낌없이 지원해주시고 격려해 주신 아버님, 항상 노심초사 하시며 학위취득을 위해 기도해 주시고, 행여 힘들세라 사소한 집안일조차 시키지 않으시고 묵묵히 기다려주신 어머님, 어린 시절부터 늘 곁에서 친구처럼 때로는 언니처럼 많은 추억을 공유하고 저를 걱정해주고 챙겨주는 따뜻함을 가진 두 여동생(김수현, 김정현), 저를 빼놓지 않고 아내와 같이 곁에서 든든하게 지켜주고 챙겨주는 두 제부(김영훈, 김령태), 너무나도 사랑스럽고 제 삶의 원동력인 두 조카(김태균, 김혜영)에게 감사의 말씀 드립니다.

그리고 저의 인생에 큰 사랑과 관심을 두시는 지인분들이 있습니다. 때로는 따끔한 훈계와 애정으로 저의 석사 박사과정 및 인생의 큰 길잡이를 주고 늘 정신적으로 의지하고 있는 멘토이자 늦었지만, 박사과정을 다시 시작하고 마칠 수 있게 큰 도움을 주며 지지해준 박세익, 따뜻한 마음으로 항상 저를 대해 주시며 늘 언니이자 엄마처럼 곁에서 챙겨주신 송은미 박사, 지금의 지도교수님을 만나고 무사히 박사과정을 마칠 수 있게 늘 지지해주신 김성환 교수님, 오랜 기간 늘 제 편에서 든든하게 저를 지지하고 응원해주고 힘들 때는 격려를 아끼지 않고 해준 친구 김태연, 박사과정 동안 늘 든든하고 힘들 때마다 의지하고 응원해주시며 익살스러움과 진지함을 함께 겸비하고 계신 두 박사님이신 고훈 박사와 임광철 박사님께 감사의 말씀을 전합니다.

이 글을 통해 표현하지 못하였지만, 저의 삶에 활력이 되는 친구들, 교수님들, 선배님들, 후배들께 따로 표기하지 않았지만 늘 감사하다는 말씀을 전합니다.

마지막으로 위의 모든 분과 인연을 맺으며 인생의 즐거움과 행복을 느낄 수 있고 늘 감사하는 마음으로 모든 일에 최선을 다할 수 있도록 키워주시고 진로에 대한 저의 결정을 믿고 적극적으로 지지해주신 부모님께 다시 한번 감사하다는 말씀드립니다.