http://dx.doi.org/10.16945/2020222179

『인문언어』(*Lingua Humanitatis*) 22권2호 (2020): 179-200

분산 표현을 통한 서법 조동사 유사도 분석

윤 태 진
(성신여자대학교, 교수)

1. 서론

본 논문의 목적은 단어의 분포 가설을 바탕으로 원어민의 말뭉치와 영어학습자의 말뭉치를 각각 모델링한 후, 원어민의 단어 분포에서 관찰되는 조동사들의 유사성과 영어학습자들이 작성한 에세이에서 관찰되는 조동사들의 유사성을 각각 분석하고, 또한 두 집단에서 관찰되는 조동사들 간의 차이점을 살펴보는 것을 목적으로 한다.

어휘 의미에 관한 일반적인 연구 방법에는 사전적인 정의에 기반을 둔 연구 방법이나 어휘의 속성에 기반을 둔 연구 방법이 있으나, 이러한 연구 방법은 연구자들의 직관(intuition)에 의존해야 하는 근본적인 문제점을 내포하고 있다(McRae, de Sa, and Seidenberg, 1997). 이에 반해 어휘가 쓰이는 환경(environment) 혹은 분포(distribution)에서 의미 내용을 규정하려는 연구 방법도 있지만, 이는 상당한 양의 말뭉치를 관찰해야 한다는 문제점을 내포하고 있다. 본 논문에서는 분포 가설을 주장한 언어학자들의 제안을 살펴보고, 실제 원어민 말뭉치와 영어학습자 말뭉치를 사용하여, 분포 가설을 통해 조동사

의 쓰임에 대한 분석을 시도하고자 한다.

분포 가설과 관련하여 가장 잘 알려진 언어학자는 London 학파의 J.R. Firth를 들 수 있다. Firth의 유명한 인용구인 “You shall know a word by the company it keeps. (Firth, J. R. 1957:11)”는 문맥을 통해 단어의 의미를 유추할 수 있다는 주장을 단적으로 보여준다고 할 수 있다. 구체적인 연구 방법론과 관련하여 Firth는 그의 논문 “Modes of meaning”에서 문맥을 통한 의미를 조사하는 한 가지 방법은 연어 테스트(collocation test)라고 하였다.

> In order to understand these terms, the modern scholar has to resort to the linguistic device of meaning by collocation; the great advantage of this method, of course, is that linguistic methods and devices are being used to establish the meaning of these linguistic terms. (Firth, 1957; 신익성 1974 재인용)

단어 의미의 일부는 이러한 연어 테스트를 통해 기술할 수 있는데, 예를 들면, 젖소의 부분적인 의미는 ‘they are milking cows,’ 혹은 ‘cows give milk’와 같은 구에서 관찰되는 연어(collocation)의 도움을 받아 확립할 수 있다 (Firth 1967: 13; 이민우, 2016).

어휘 의미에 대한 문법적 접근법은 다른 언어학자들에 의해서도 제안되었다. 예를 들어, Deese(1965: 43-46)에서는 “The distribution of responses evoked by a particular word as stimulus defines the meaning of that word. (4.3)”라고 하였다. 또한 Cruse(1986)는 그의 저서 어휘 의미론(Lexical Semantics)에서 “the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts.”라고 하여, 한 어휘의 의미가 실질 혹은 잠정적인 문맥에서 추출된다는 가정에 기반을 두고 있다고 하였다.

그렇다면 문맥에서 어휘의 의미를 관찰할 수 있는 접근법은 어떤 것이 있을까? Harris (1954: 156)는 “In certain important cases it will even prove possible to state certain aspects of meaning as functions of measurable

distributional relation."이라고 제안하여, 측정 가능한 분포적인 관계를 통해 어휘의 의미를 관찰할 수 있다고 하였다.

Harris가 제안한 문맥을 통한 의미 연구의 가능성은 실제 말뭉치를 통한 분산 표현 방법론을 통해 구현되고 있다. 분산 표현(distributed representation) 방법은 기본적으로 분포 가설(distributional hypothesis)이라는 가정 하에 만들어진 표현 방법이다. 분포 가정은 '비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다'라는 가정이다. 강아지란 단어는 귀엽다, 예쁘다, 애교 등의 단어가 주로 함께 등장하는데 분포 가설에 따라서 이러한 내용을 가진 텍스트를 벡터화한다면 이러한 단어들은 의미적으로 가까운 단어가 된다. 분포에 기반한 의미 분석 방식을 분포 의미론(distributional semantics)라고 하며, 분포 의미론은 컴퓨터로 구현할 수 있으며, 인간의 유사성 판단과 마찬가지고 단어 간의 유사성을 잘 모델링할 수 있는 것으로 파악하고 있다.

> "Distributional semantics is a theory of meaning which is computationally implementable and very, very good at modeling what humans do when they make similarity judgments. ... This approach to meaning is in no way the only one, but has come from a particular philosophical tradition involving linguists and philosophers such as Leonard Bloomfield, Zellig Harris, J.R. Firth or again Lugwig Wittgenstein (in his later work) and Margaret Masterman[1] (최재웅, 2018에서 재인용).

언어학의 분석 방법 중 하나로 자리잡고 있던 분포의미론이 근래에 분산 의미론이 대중의 눈길을 끈 사건이 있었다. 2013년 구글의 토마스 미코로프(Tomas Mikolov)는 word2vec을 발표하였다 (Mikolov et al. 2013). 단순한 구조의 신경망을 사용하여 효과적으로 단어들을 잠재 공간(latent space)에 성공적으로 투사시킴으로써, 딥러닝을 활용하여 자연어 처리 및 활용을 할 수 있는 가능성을 보여주었다 (지인영, 김희동, 2020). 비슷한 의미의 단어일수록 저차원의 잠재 공간에서 가깝게 위치하는 것을 tSNE라는 시각적인 장치로 보

1) http://aurelieherbelot.net/research/distributional-semantics-intro/

여 주었으며, 컴퓨터가 할 수 없을 것이라고 믿고 있었던 단어 간의 관계도 보여 주었다. 예를 들면, man과 boy의 관계를 보여주고, woman과 관련된 단어를 물었을 때, girl이라는 결과를 보여주었다. 또한 London, Seoul, New York, Japan과 같은 단어들 중 관련이 가장 적은 단어를 골라내는 문제에서도 Japan이 가장 관련성이 적다는 것을 정확히 보여 주었다.

이와 같이 최근 분포 가설을 기반으로 한 어휘 의미의 연구가 상당한 진척을 보여주고 있다. 이러한 분포 가설에 입각한 전통적인 연구 방법을 통하여, 본 논문에서는 원어민의 말뭉치에서 추출한 조동사들과 유사성을 가지는 단어와 외국어로서 영어를 학습하는 비원어민의 말뭉치를 사용하여 추출한 조동사들이 유사하다고 판단한 단어들을 비교·분석하고자 한다. 이를 통해, 일반적으로 원어민 말뭉치에서만 적용되던 분산 표현이 영어학습자 말뭉치에서는 어떤 결과를 보여주는 지 살펴보고자 한다.

조동사는 분포 상 명사 뒤와 본동사 앞에 위치하고 있으며 시제를 지닌다는 특징을 가지고 있다. 이러한 특징은 분포 상 조동사들이 서로 유사성을 가진다고 가정할 수 있다. 따라서 본 논문에서 자주 사용되는 조동사의 분포를 보면 타 조동사와는 유사성을 가지며, 그렇지 않은 품사들과는 유사성을 덜 띄게 될 것이라고 가정할 수 있다.

또한 서법 조동사의 쓰임에 있어서 원어민과 영어학습자 사이에 차이가 있을 것으로 가정하고 있다. 서법 조동사에 대한 선행연구들은 영어학습자들이 양태동사를 올바르게 사용하는 데 어려움이 있으며, 특정 양태의 의미나 형태들을 모국어화자에 비해 너무 많이 사용하거나 너무 적게 사용하여 중간언어의 특징을 드러낸다고 알려져 있다 (Adjemian, 1976; Aijmer, 2002; Oh, 2007; Park 2012 등). 이와 더불어 영어학습자들이 서법조동사를 사용하는 담화 문맥이 원어민들과는 차이를 보이는 것으로 알려져 있다 (Deshors, 2010). 따라서 비록 원어민의 서법 조동사 사용 및 분포를 연구하는 것도 중요하지만, 영어 학습자들의 서법 조동사 사용과 관련된 양상을 살펴보는 것도 중요하다. 문맥상에서 서법 조동사가 어떻게 사용되는 지를 원어민의 말뭉치와 영어학습자의 말뭉치에서 분석하여 비교함으로써 영어학습자 및 원어민의 서법조동

사 사용과 관련한 특징을 각각 파악할 수도 있으며, 또한 두 특징을 비교함으로써 원어민과 영어학습자 사이의 차이점도 밝혀낼 수 있을 것이다.

2. 연구방법

2.1 TOEFL11 말뭉치

TOEFL11은 2014년도 미국의 공인시험을 담당하는 ETS(English Testing Service)가 LDC(Linguistic Data Consortium)를 통해 공개한 말뭉치로서, 서로 다른 모국어를 구사하는 영어학습자들이 주제별로 쓴 에세이로 구성되어 있는 말뭉치이다 (Blanchard et al. 2013). 이 말뭉치는 11개의 서로 다른 모국어를 구사하는 영어학습자들이 각각 영어로 작성한 1,100개의 에세이를 구성하여, 총 12,100개의 에세이로 구성되어 있다. TOEFL11의 구성 내용과 관련하여서는, 각 텍스트 파일이 에세이를 작성한 영어학습자의 모국어(L1)에 대한 정보와 전문적인 훈련을 받은 에세이 채점자들의 채점결과인 상중하의 영어 구사능력을 제외하고는 아무런 정보가 담겨있지 않는 원시 말뭉치(raw corpus)이다.

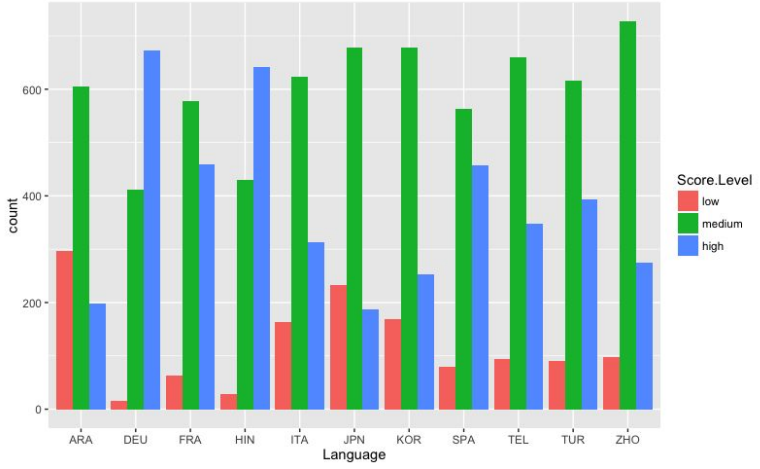
TOEFL11은 2014년도 공개되었지만, 2006년에서 2007년 사이에 TOEFL iBT® 시험을 보는 동안 작성한 독립형 에세이들로 구성되어 있다. 독립형 글쓰기 과업에서는 8개의 프롬프터 가운데 하나가 주어지고, 이 주어진 주제에 대한 자신의 의견을 구체적이고 논리적으로 설명해서 독자를 설득시키는 persuasive writing으로서 30분 동안 작성한다.

독립형 에세이는 두 명의 채점자가 정해진 기준에 의해 0에서 5점까지 점수를 매겨서 그 평균값을 산출하는 방식으로 채점이 진행된다. 만약 두 채점자의 점수 차이가 1점 이상이면 제3의 채점자가 최종 점수를 확정한다.

참고로, 윤태진·이용훈 (2016)에서는 TOEFL11 말뭉치에서 영어 응시자들이 받은 등급과 모국어에 따라 관찰되는 서법 조동사들의 분포를 분석하였다.

그리고 윤태진(2018) 및 윤태진(2019)에서는 영어 응시자들이 받은 등급과 에세이 내용에 담겨진 언어정보들을 사용하여 기계학습을 이용하여 주제 프롬프터를 자동으로 분류하는 연구와 비지도 학습(unsupervised learning) 방식 중 토픽모델링을 이용하여 주제 프롬프터를 유추하는 연구를 각각 수행하였다.

<그림 1>은 전체 말뭉치를 상중하로 나눈 에세이의 수를 참고로 보여주고 있다. 본 논문에서는 TOEFL 응시자들이 제출한 에세이 중 상위 등급을 받은 에세이들을 추출하여 분석의 토대로 사용하고자 한다.



<그림 1> 11개 모국어의 영어학습자들의 성적을 상중하로 구분한 TOEFL11 말뭉치의 에세이수
(윤태진 · 이용훈 2016에서 가지고 옴)

모든 종류의 조동사를 분석한 것은 아니고, 본 TOEFL11 말뭉치에서 관찰되는 조동사들인 can, could, may, might, will, would 등을 분석하였다 (윤태진· 이용훈, 2016). 조동사의 선택은 선행 연구의 결과에 기반을 두었다. 선행 연구에서 품사 태깅된 자료에서 조동사를 추출하였는데, 현저하게 빈도수가 많은 조동사는 can(26,710)과 will(25,818)이라는 것을 볼 수 있다. 이어서 빈도수가 다소 높은 조동사는 그 순서대로 나열하면 would(10,713), should (6,588), could(5,491), may(5,357), might(3,329), 그리고 must(2,665)이다. 이에

반해, shall, ought, dare는 각각 97, 42, 9의 빈도수로 TOEFL11에는 거의 나타나지 않았다 (윤태진 · 이용훈, 2016). 본 논문에 사용된 단어들의 토큰 빈도수는 1,676,252단어이며, 타입 빈도수는 36,715 단어이다. 참고로 학습자 말뭉치에서 철자상의 오류 등은 따로 수정하지 않았다. 따라서 타입 빈도수에 오타 등이 포함되어 있어서 실제 타입 빈도수라고 보기는 힘들다는 문제점을 내포하고 있다.

2.2 PTB Corpus

PTB(Penn Treebank) 말뭉치는 Thomas Miklov의 웹페이지에서 다운로드 받은 것을 사용하였다. 이 PTB 말뭉치는 LDC에서 배포한 원래의 PTB 문장에 몇 가지 전처리(preprocessing)를 한 텍스트 파일이다. 예를 들어, 빈도수가 낮은 단어는 <unk>라는 문자로 치환하였고, 102,013과 같은 숫자들은 통칭하여 'N'으로 대체하는 등의 작업을 적용한 후 word2vec에 사용하기 위해 공개한 텍스트 데이터이다. PTB 말뭉치는 한 문장이 하나의 줄로 저장되어 있으나 본 논문에서는 전체 문장에서 나타나는 단어의 빈도수를 계산하였기 때문에, 문장의 구분을 고려하지 않았다. 본 논문에 사용된 말뭉치의 토큰 빈도수는 1,075,569단어이며, 타입 빈도수는 9,926단어이다.

2.3 모델링

본 논문에서는 원어민 말뭉치와 TOEFL11의 영어학습자 말뭉치를 이용하여 각 말뭉치에서 관찰되는 단어들에 대한 분산표현을 적용한 후, 조동사 형태에 해당하는 can, could, may, might, will, would 등이 가장 유사성을 지니는 단어를 찾는 방식으로 모델링하였다. 또한 각 말뭉치에서 연산된 조동사 간의 거리를 말뭉치 별로 특징을 살펴보았으며, 말뭉치 사이에 관찰되는 특징도 살펴보았다. 분산 표현을 위한 모델링을 위해서는 Python과 numpy, matplotlib, sklearn과 같은 여러 패키지들을 이용하였다. 그리고 시각화를 위

한 그래프는 Python의 연산 결과를 이용하여 matplotlib을 이용하기도 하고, 연산 결과를 R에서 읽어 들여 R의 패키지인 ggplot을 이용하여 만들었다.

2.3.1 동시 발생 행렬

분포 가설을 토대로 단어의 의미를 유추하기 위해서는 단어의 문맥 정보(contextual information)를 이용해 해당 단어를 벡터로 나타낼 필요가 있다. 예를 들어, I wish I can swim.이라는 단어가 있을 때, 문맥의 범위를 해당 단어의 전후 1이라고 가정하였을 때, 단어 I의 문맥은 wish라는 단어 하나와 can이라는 단어 하나로 잡을 수 있다. 이렇게 단어 I의 문맥으로써 동시에 등장하는 단어의 빈도를 벡터로 표현할 수 있으며, 예를 든 문장의 경우 해당 단어의 벡터가 처리되었으면 다음 단어로 하나하나씩 건너뛰어 이전 단어와 동일한 방법으로 문맥 정보에 따라 해당 단어의 벡터 표현을 할 수 있다. 분석 대상이 되는 모든 단어를 벡터로 표현한 후, 모든 단어에 대해 동시 발생하는 단어를 행렬의 변환시켜 동시 발생 행렬(co-occurrence matrix)을 구한다 (Manning & Schütze, 1999).

2.3.2 벡터 간 유사도

동시 발생 행렬에서 각 행과 열은 해당 단어의 전후에 나타나는 단어들의 문맥정보가 포함된 벡터들이다. 이 벡터로 표현된 단어들은 정량화된 도구를 사용하여 처리할 수 있는데, 이 중 하나가 벡터 간의 유사도를 사용하여 단어 간의 유사성을 측정할 수 있다는 것이다.

벡터 사이의 유사성을 측정할 수 있는 방법은 맨하탄 거리(Manhattan Distance), 유클리드 거리(Euclidean distance), 코사인 유사도(Cosine similarity) 등 여러 가지가 있다 (Manning & Schütze, 1999). 본 논문에서는 코사인 유사도를 이용하여 분석을 하였다. 코사인 유사도는 분자에 벡터의 원소별 곱인 벡터의 내적을 계산하고, 분모에는 각 벡터의 크기(magnitude)를 계산한다. 벡터의 크기로 벡터의 내적을 나눔으로써 유사성을 정규화(normalization)시키는 역할을 한다. 코사인 유사도의 결과가 1에 가까울수록 방향은 일치하고, 0에

가까울수록 직교이며, -1에 가까울수록 반대 방향임을 의미한다. 코사인 유사도는 크기와 방향 모두를 고려한 방법이어서, 텍스트 처리에서 가장 널리 쓰이는 유사도 측정 방법이다. 물론 단점이 없는 것은 아닌데, 벡터 차원의 크기가 클수록 연산이 부담이 된다는 단점이 있다.

2.3.3 상호정보량(mutual information)

동시발생 행렬의 원소는 두 단어가 동시에 발생한 빈도수를 나타낸다. 하지만 발생 빈도수는 the, a(n), and 등과 같은 고빈도 단어를 고려했을 때 문맥이 해당 단어의 정보량을 적절하게 전달하지 못한다는 문제점을 가지고 있다. 이러한 문제를 해결하기 위해 사용할 수 있는 방법 중 자주 사용되는 것이 상호 정보량(mutual information)이다 (Manning & Schütze, 1999). 상호 정보량은 단어 x와 단어 y가 동시에 일어날 확률을 단어 x가 일어날 확률과 단어 y가 일어날 확률의 곱으로 나누어서 계산한 후, 베이스를 2로 하는 로그 함수로 변환하여 계산한다. 본 논문에서는 말뭉치에 기반을 둔 경험적인 확률값을 계산하기 위해 실제의 구현에 있어서는 확률을 특정 단어가 나타나는 빈도수를 해당 말뭉치 전체의 단어수로 나누어서 계산하였다. 이와 같은 상호 정보량을 사용하면 분모에서 각 단어 x와 y가 나타나는 확률이 정규화의 역할을 하여, x와 y의 상호정보량의 크기를 조정하는 역할을 한다.

2.3.4 차원 감소(dimensionality reduction)

본 논문은 동시발생 행렬을 상호 정보량 행렬로 전환하였다. 상호 정보량 행렬로 전환하면 생길 수 있는 문제가 있다. 이는 말뭉치의 어휘 수가 증가함에 따라 각 단어 벡터의 차원 수도 증가한다는 문제이다. 또한 행렬을 구성하고 있는 단어들의 벡터에 해당 단어와 관계를 가지고 있지 않은 단어들이 포함하고 있어서 0으로 채워진 원소들이 많이 있게 된다. 이러한 문제에 대체하고자 자주 수행하는 기법은 벡터의 차원 감소(dimension reduction)이다. 차원 감소는 중요한 정보는 최대한 유지하면서 차원을 줄이는 방법을 일컫는다. 차원 감소를 위한 방법으로는 PCA(Principal Component Analysis)와 특이값 분

해(Singular Value Decomposition(SVD) 등 여러 방법이 있다 (Manning & Schütze, 1999). 본 논문에서는 특이값 분해를 사용하여 차원을 감소하였다. 특이값 분해는 상호정보량 행렬을 행렬의 크기에 3제곱에 비례해서 연산이 되는 것으로 알려져 있다. 이를 해결하기위해 Python의 sklearn 라이브러리에 있는 randomized_svd 함수에 있는 truncated SVD를 사용하여 연산 속도를 빨리하는 방식을 취하였다. 이와 같이 하여 특이값이 높은 값을 가지는 차원이 감소된 행렬을 사용하였다.

3. 결과 분석 및 논의

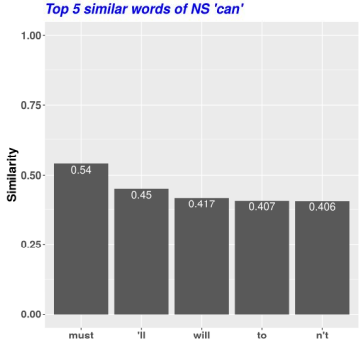
이상 말뭉치를 사용해 문맥에 속한 단어의 등장 횟수를 센 후 상호 정보량 행렬로 변환하고, 다시 특이값 분해를 이용해 차원을 감소시킴으로서 단어 벡터를 구현하였다. 말뭉치를 구성하고 있는 단어들의 유사성을 구현하면, 이를 활용해 특정 단어가 그 외의 다른 단어들과의 유사도를 검색할 수 있다. 이를 응용하면 상위 5개 정도의 가장 유사성이 높은 단어를 찾아 볼 수 있다. 이와 같은 방식을 통해 단어를 분산 표현의 형태로 구현한 후, can, could, may, might, will, would와 같은 단어들과 유사하다고 계산된 단어를 각각 상위 5개씩 추출하였다.

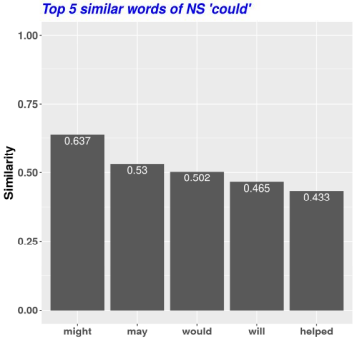
추출된 결과는 단어의 의미 혹은 문법적인 관계에서 비슷한 단어들이 가까운 벡터로 나타날 것이며, 우리의 직관 혹은 예측과 일치한 결과가 나타날 수도 있을 것이며, 그렇지 않을 수도 있을 것이다.

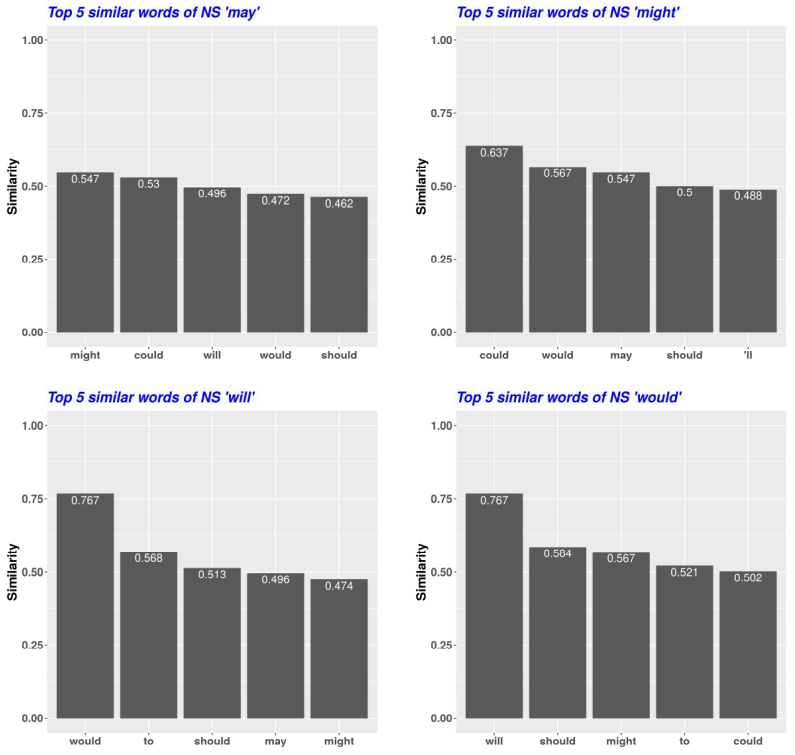
3.1 원어민 텍스트 분석의 결과

원어민의 말뭉치를 분석하여 can, could, may, might, will, would와 유사하다고 계산된 상위 5개의 단어들을 각각 막대그래프를 이용하여 <그림 2>에 제시하였다. 전반적으로 각 조동사와 유사한 단어라고 유추되어 추출된 상위

5개의 단어들은 확률로 표시되었다. 결과를 살펴보면, 일부 몇 개의 조동사들을 제외하고는 전반적으로 다른 조동사 혹은 그 축약형들이 유사한 형태의 단어들이라고 계산되어 반환된 것을 막대그래프를 통해서 살펴볼 수 있다. 구체적으로 can의 경우는 원형 부정사 to를 제외하고는 must, 'll, will, 그리고 n't가 유상 단어 중 상위 5개에 속한다고 계산되었다. 'll과 n't는 각각 will과 won't와 같은 조동사들의 축약형으로 볼 수 있다. could의 경우는 helped를 제외하고는 might, may, would, 그리고 will이 가장 유사한 상위 4개의 단어를 그 결과로 가졌다. may는 might, could, will, would, 그리고 should의 순으로 유사단어들이 추출되었으며, might의 경우는 could, would, may, should, 그리고 'll의 순으로 유사단어들이 계산되어 드러났다. will과 would의 경우도 can의 경우와 같이 to가 상위의 유사 단어로 포착되었다. to의 의미가 어떤 것인지 파악하기는 힘들지만, 부정사의 의미로 쓰인 to가 유사한 것으로 파악되었다면, 조동사와 전혀 무관한 것은 아니고, 조동사의 의무 혹은 인식론적인 의미(epistemic or denontic meaning)와 관련이 있다고 간주할 수 있을 것이다. 그 외에는 will은 조동사인 would, should, may, might가 발견되었으며, would의 경우는 will, would, might, could가 유사한 단어로 발견되었다.

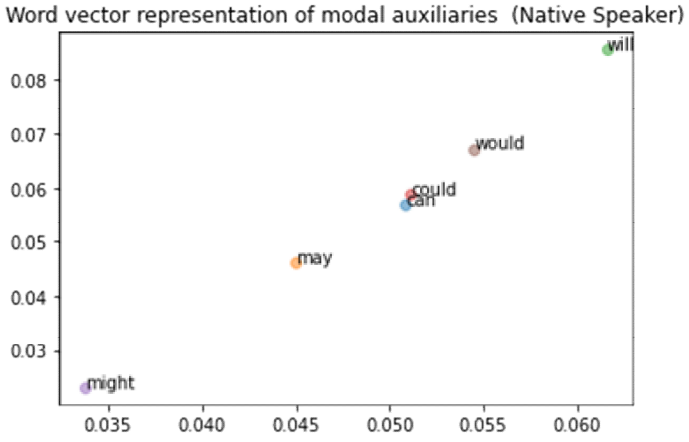






<그림 2> 원어민 말뭉치에서 서법 조동사가 가장 유사하다고 판단한 상위 5개의 단어와 그 확률을 나타낸 막대그래프

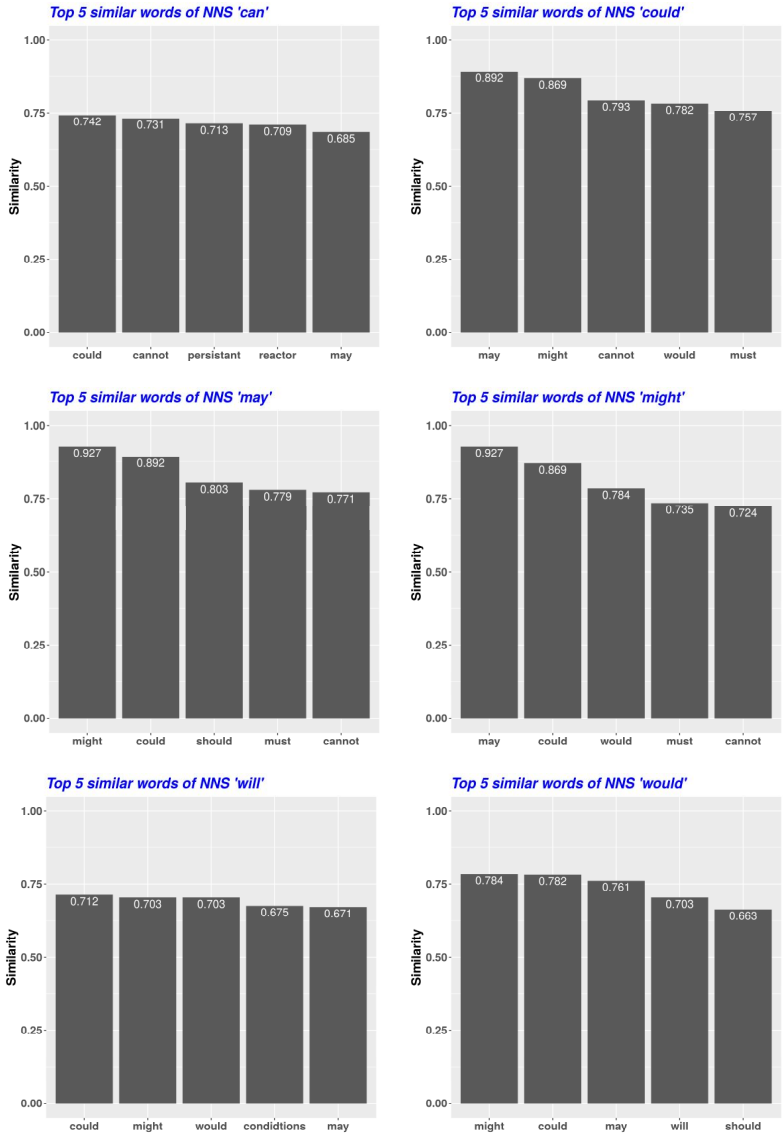
<그림 3>은 PTB 말뭉치에서 can, could, may, might, will, 및 would가 가지는 상호간의 거리를 두 차원의 벡터 상에 제시하였다. 이 6가지의 조동사 형태들은 대각선을 기준으로 나열이 되어 있는 것으로 나타난다. might와 will이 가장 거리가 먼 쌍으로 표현되어 있으며, can과 could는 그 거리가 아주 밀접하게 표시되어 있는 것을 볼 수 있다.



<그림 3> 원어민 말뭉치에서 계산된 서법조동사 간의 거리

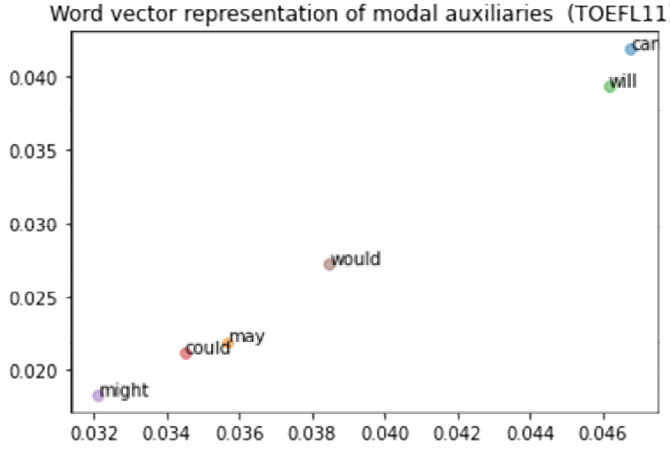
3.2 영어학습자 텍스트 분석의 결과

이상 원어민의 경우 일부 유사 단어들을 제외하고는 조동사들이 다른 조동사들과 가장 유사한 분포적인 구조를 보이는 것으로 발견되었다. 영어학습자의 경우도 일반적으로 다른 조동사들을 유사한 단어로 포착하였다는 특징을 아래의 <그림 4>에서 살펴볼 수 있다. 구체적으로 can의 경우는 (persistent의 철자상의 오류인) persistant와 reactor를 제외하고는 could, cannot, may가 유사한 문맥적 정보를 가진 단어로 포착되었으며, could의 경우는 may, might, cannot, would, 그리고 must가 유사한 단어로 포착되었다. may의 경우는 might, could, should, must, cannot이 유사한 단어로, might의 경우는 may, could, would, must, cannot이 유사한 단어로 포착되었다. will은 could, might, would, may의 조동사와 일반적으로 유사한 것으로 관찰되었으나, conditions라는 단어도 유사한 것으로 계산되었다. would의 경우는 might, could, may, will 그리고 should와 같이 모두 조동사가 유사한 단어로 포착되었다.



<그림 4> 영어학습자 말뭉치에서 서법 조동사가 가장 유사하다고 판단한 상위 5개의 단어와 그 확률을 나타낸 막대그래프

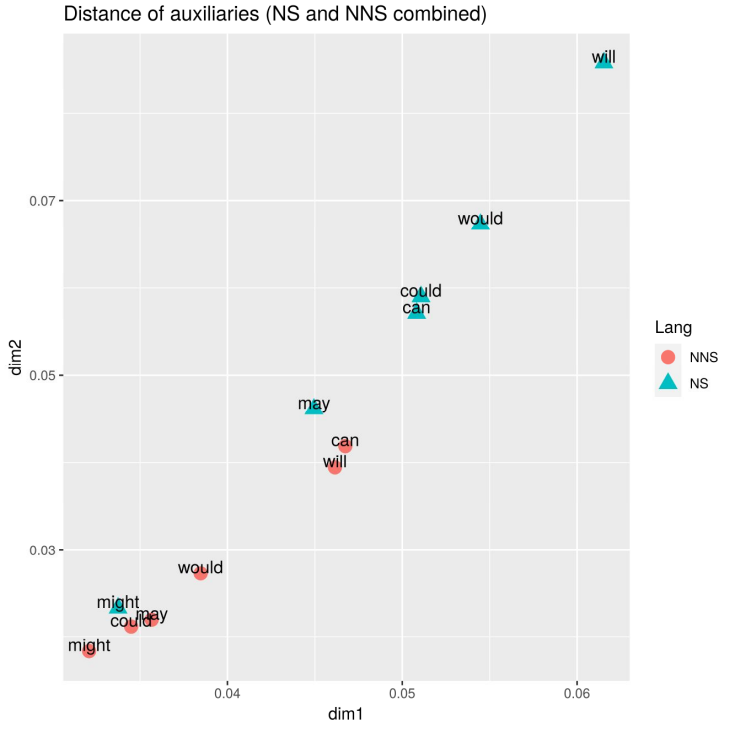
<그림 5>는 TOELF11 말뭉치에서 상위의 등급으로 분류된 영어 학습자들의 에세이를 모델링하여 2차원의 벡터로 표시한 그래프이다. 6가지의 조동사들이 대각선을 기준으로 나열되어 있는 것을 볼 수 있으며, can과 will이 아주 밀접하게, 그리고 may와 could가 아주 밀접하게 위치해 있음을 볼 수 있다. 또한 can과 might의 벡터상의 거리가 가장 먼 것으로 파악할 수 있으며, would는 중간에 위치해 있음을 그래프를 통해 분석할 수 있다.



<그림 5> 상급의 점수를 받은 영어학습자들의 서법조동사 간의 거리

<그림 6>은 개별적으로 분석한 영어학습자와 원어민들의 조동사 간의 거리를 합쳐서 제시한 것이다. 전반적으로 원어민들의 조동사간의 거리가 비원어민들의 조동사간의 거리보다 더 넓게 퍼져 있음을 관찰할 수 있다. 또한 원어민 말뭉치에서 계산된 can과 could의 거리가 다른 조동사들에 비해 매우 가깝게 위치해 있음을 관찰할 수 있다. 이에 비해 영어학습자 말뭉치의 경우 can은 will과 더 가까운 것으로, 그리고 could는 may와 더 가까운 것으로 관찰되고 있다. 또한 비원어민 말뭉치의 may와 could는 원어민 말뭉치의 might와 유사한 것으로, can과 will은 원어민 말뭉치의 may와 가까운 것으로 관찰

되었다.



<그림 6> 이차원 상에 표현된 원어민(NS)과 영어학습자(NNS)들의 서법조동사 간 거리

3. 결론

본 논문에서는 TOEFL11 말뭉치에서 상위 수준의 점수를 받은 영어학습자들의 에세이와 원어민 말뭉치인 PTB 말뭉치를 대상으로 분산 표현을 이용한 분석을 시도하였다. 구조주의 언어학의 의미 분석에 대한 제안인 측정 가능한 분포상의 관계(measurable distributional relation) 내에서 의미 관계를 분석하였다는 점에서 또한 상대적으로 적은 양의 말뭉치를 사용하여 직관(intuition)

에 근접하는 결과를 도출할 수 있었다는 점에서 본 논문의 의의가 있다 (Harris, 1951). 원어민 말뭉치와 학습자 말뭉치에서 빈도수가 높은 조동사들을 선택하여 각 조동사가 말뭉치 내에서 가장 유사한 단어라고 추정되는 단어들의 확률을 계산한 후, 확률 상 유사성이 높다고 계산된 상위 5개의 단어들을 검토해 본 결과, 약간의 예외는 있지만 조동사는 다른 조동사들과 유사성이 높다는 결론을 내릴 수 있었다. 원어민 말뭉치에서의 조동사 간의 거리와 영어 학습자 말뭉치에서의 조동사 간의 거리를 비교 분석한 결과, 원어민 말뭉치의 경우 조동사들 간의 거리가 어느 정도 간격을 두고 있었지만, 영어 학습자의 경우 조동사들이 더 밀접하게 분포되어 있었던 것을 그래프 분석을 통해 관찰할 수 있었다. 또한 원어민 말뭉치에서는 can과 could가 상당히 밀접한 분포상의 관계를 보여줌에 반해, 학습자 말뭉치에서는 can과 could는 거리가 있었으며, 대신 can의 경우 will과 유사성을 보여주었으며, could의 경우 may와 유사성을 보여주고 있다는 것을 관찰할 수 있었다.

본 연구가 가지는 이론적 문제점과 기술적 문제점을 몇 가지 고찰할 수 있다. 첫째는 말뭉치에서 분산 표현에 따라 조동사의 문맥적 정보를 이용해 다른 조동사들과 유사한 특징을 가진다는 흥미로운 사실을 발견한 것은 고무적인 일이지만, 조동사들이 가지는 의미가 하나로 한정되어 있지 않지만 현재의 접근법으로는 조동사들의 보다 세부적으로 구분된 의미와 관련한 유사성을 관찰할 수 있을 지에 대한 의문이 있음을 떨칠 수 없다는 한계점이 있다. 예를 들면, 조동사들은 기본적으로 인식론적 의미(epistemic meaning)과 의무론적 의미(deontic meaning)가 있지만 (Stephany, 1995), 이러한 구분을 본 접근법으로는 찾아낼 수 없다는 한계점을 가진다.

본 연구의 분산 표현에 사용한 방식은 전통적이고 잘 알려진 방식이지만, 기술적으로 단점이 없는 것은 아닌데, 가장 큰 기술적 단점 중의 하나는 분석 속도가 말뭉치가 많아짐에 따라 늦어지고 메모리를 많이 차지한다는 점이다. 최근 딥러닝을 이용한 자연언어처리에서 skip-gram 모델, CBOW(continuous bag of words), 혹은 글로브(Global Vectors for Word Representation, GloVe)와 같은 방식을 사용하여 분산 표현을 구현하고 있다 (위혜경, 2019; 지인영,

김희동, 2020). 이는 word2vec의 방식으로 단어를 벡터로 표현하는 워드 임베딩과 관련이 있다. CBOW는 주변에 있는 단어들을 가지고, 중간에 있는 단어들을 예측하는 방법이다. 반대로, Skip-Gram은 중간에 있는 단어로 주변 단어들을 예측하는 방법이다. Glove는 카운트 기반과 예측 기반을 모두 사용하는 방법론으로 2014년에 미국 스탠포드 대학에서 개발한 단어 임베딩 방법론이다.

이러한 최근의 접근법들은 대용량의 원어민 말뭉치를 기반으로 한 연구가 일반적이다. 이를 영어 학습자 말뭉치에 적용하는 시도는 가치가 있을 것으로 판단된다. 또한 본 연구에서 밝혀진 분산 표현의 결과는 빈도수 기반 모델로 한 것이다. 동일한 말뭉치를 사용하여 예측 기반 모델을 사용하였을 때는 어떤 결과가 나오는지, 그리고 용량을 더 늘릴 수 있다면 결과가 다르게 나올지 연구해 볼 필요가 있다.

마지막으로 본 연구에서는 TOEFL11 말뭉치에서 상위의 성적을 받은 영어 학습자의 말뭉치를 분석의 대상으로 하였다. 말뭉치의 양과 관련해서는 중위의 성적을 받은 학습자들의 말뭉치가 가장 많은 반면, 하위의 성적을 받은 학습자들의 말뭉치는 가장 적은 것으로 관찰되었다. 말뭉치의 양이 충분히 많다면 성적에 따른 조동사의 유사성을 분석해서 문맥에 따른 조동사의 쓰임에 관한 차이를 살펴보는 것도 의미있는 연구가 될 것이다.

인용문헌

신익성. J.R. Firth의 언어 이론: 의미의 문제를 중심으로.『어학연구』10(2), 160-173. 1974.

유만근. J.R. Firth의 언어 이론 연구 - Prosody 음운론에 중점을 두고.『한글』145, 71-110. 1970.

윤태진, 이용훈. 동북아시아 영어 학습자들의 서법 조동사 사용에 대한 연구.『인문언어』22(1), 195-217. 2020.

윤태진. TOEFL11 을 이용한 비지도 토픽 모델링.『언어과학』26(1), 51-70. 2019.

윤태진. TOEFL11 코퍼스에서 주제 프롬프터의 자동 분류.『인문언어』 20(1),

157-177. 2018.
윤태진. 대응 분석을 통한 서법 조동사의 범언어적 분포 연구.『인문언어』20(2), 311-330. 2018.
윤태진, 이용훈. 학습자 코퍼스를 이용한 서법조동사의 분포의 범언어적 연구.『인문언어』18(2), 137-154. 2016.
이민우. 단어 내부의 의미 관계에 대한 연구.『어문론집』66, 77-34. 2016.
위혜경. 언어 능력과 자연어 처리: 자연인가 경험인가?『인문언어』21(2), 211-249. 2019
지인영, 김희동. 사전학습 언어모델의 기술 분석 연구.『인문언어』22(1), 111-133. 2020
최재웅. 딥러닝 방식의 워드 벡터를 활용한 어휘 의미 표상. 한국영어학회 2018 가을 학술대회 발표자료. 2018.
Aijmer, K. Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 55-76, Amsterdam: John Benjamins. 2002.
Blanchard, D., J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow. TOEFL11: A Corpus of Non-native English. *ETS RR – 13-24*. Prinston, NJ: Educational Testing Service. 2013.
Cruse, D. A. *Lexical semantics*. Cambridge: Cambridge University Press. 1986.
Deese, J. *The Structure of Association in Language and Thought*. Baltimore: Johns Hopkins University Press. 1965.
Deshors, S. *Multifactorial Study of the Uses of May and Can in French-English Interlanguage*. Ph.D. dissertation, University of Sussex. 2010.
Firth, John R. A synopsis of linguistic theory 1930 – 1955. In *Studies in linguistic analysis*, 1 – 32. Oxford: Blackwell. 1957.
Firth, J. R. Modes of Meaning, In *Papers in Linguistics* 1934-51. Oxford: Oxford University Press. 1951/1967.
Harris, Z. *Methods of Structural Linguistics*, Chicago: Chicago University Press. 1951.
Jun, K. and Y. Lee. A statistical analysis of can and may in British and Indian English. *English Language and Linguistics* 21(3), 63-84. 2015.
Manning, C., & Schütze, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press. 1999.

McRae, K., de Sa, V. R., & Seidenberg, M. S. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126, 99-130. 1997.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781. 2013.

Oh, S.-Y. A corpus-based study of epistemic modality in Korean college students' writings in English. *English Teaching*, 62(2), 147-175. 2007.

Park, H. Modality in Korean Learner's Spoken Interlanguage. *English Language & Literature Teaching*, 18(1), 197-216. 2012.

Stephany, U. Function and form of modality in first and second language acquisition. In *From Pragmatics to Syntax: Modality in second language acquisition*, A. Giacolone-Ramat & G. Crocco-Galeas (eds), 105-120. Tübingen: Gunter Narr. 1995.

[Abstract]

Similarity Analyses of Modal Auxiliaries using Distributional Representation

Tae-Jin Yoon
(Sungshin Women's University, Professor)

This paper attempts to measure the similarity of frequently occurring modal auxiliaries in both L1 and L2 writings. The modal auxiliaries are known to be difficult areas of study for both L1 and L2, due to the overlapping in their use. A subset of TOEFL11 corpus and a subset of PTB(Penn Treebank) corpus were used to capture the similarities among modal auxiliaries within L1 and L2, respectively, and also across L1 and L2. Based on the hypothesis of distributional representation, similarities of modal auxiliaries were computed by applying cosine similarity and mutual information to every pair of words in the corpus, and then finally reducing dimensions with Singular Value Decomposition. The results show that the modals are found to be similar to other modals, and the distribution of modals in the reduced dimensions show that the modals in L1 and those in L2 exhibits different patters of similarities, while the distance among the modals in L1 is further away than the distance among those L2 modals.

Keywords: Distributional representation, SVD(Singular Value Decomposition), cosine similarity, TOEFL 11, PTB(Penn Treebank)

윤태진 (Yoon, Tae-Jin)
성신여자대학교 인문과학대학 영어영문학과, 교수
02844 서울특별시 성북구 보문로 34다길2
02-920-7185
tyoon@sungshin.ac.kr

접 수 일: 2020년 11월 23일
심사기간: 2020년 11월 23일~12월 20일
게재결정: 2020년 12월 20일