ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 44, No. 9, pp. 946-953, 2017. 9
https://doi.org/10.5626/JOK.2017.44.9.946

워드 임베딩과 유의어를 활용한 단어 의미 범주 할당

(Assignment Semantic Category of a Word using Word Embedding and Synonyms)

박 다 솔 †　　차 정 원 ††
(Da-Sol Park)　　(Jeong-Won Cha)

요 약 의미역 결정은 서술어와 논항들 사이의 의미 관계를 결정하는 문제이다. 의미역 결정을 위해 의미 논항 역할 정보와 의미 범주 정보를 사용해야 한다. 세종 전자사전은 의미역을 결정하는데 사용한 격틀 정보가 포함되어 있다. 본 논문에서는 워드 임베딩과 유의어를 활용하여 세종 전자사전을 확장하는 방법을 제시한다. 연관 단어가 유사한 벡터 표현을 갖도록 하기 위해 유의어 사전의 정보를 사용하여 재구성된 벡터를 생성한다. 기존의 워드 임베딩과 재구성된 벡터를 사용하여 동일한 실험을 진행한다. 워드 임베딩을 이용한 벡터로 단어의 세종 전자사전에 나타나지 않은 단어에 대해 의미 범주 할당의 시스템 성능은 32.19%이고, 확장한 의미 범주 할당의 시스템 성능은 51.14%이다. 재구성된 벡터를 이용한 단어의 세종 전자사전에 나타나지 않은 단어에 대해 의미 범주 할당의 시스템 성능은 33.33%이고, 확장한 의미 범주 할당의 시스템 성능은 53.88%이다. 의미 범주가 할당되지 않은 새로운 단어에 대해서 논문에서 제안한 방법으로 의미 범주를 할당하여 세종 전자사전의 의미 범주 단어 확장에 대해 도움이 됨을 증명하였다.
키워드: 워드 임베딩, 유의어, 의미 범주, 세종 의미사전

Abstract Semantic Role Decision defines the semantic relationship between the predicate and the arguments in natural language processing (NLP) tasks. The semantic role information and semantic category information should be used to make Semantic Role Decisions. The Sejong Electronic Dictionary contains frame information that is used to determine the semantic roles. In this paper, we propose a method to extend the Sejong electronic dictionary using word embedding and synonyms. The same experiment is performed using existing word-embedding and retrofitting vectors. The system performance of the semantic category assignment is 32.19%, and the system performance of the extended semantic category assignment is 51.14% for words that do not appear in the Sejong electronic dictionary of the word using the word embedding. The system performance of the semantic category assignment is 33.33%, and the system performance of the extended semantic category assignment is 53.88% for words that do not appear in the Sejong electronic dictionary of the vector using retrofitting. We also prove it is helpful to extend the semantic category word of the Sejong electronic dictionary by assigning the semantic categories to new words that do not have assigned semantic categories.
Keywords: word embedding, synonyms, semantic category, Sejong semantic dictionary

· 이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2017R1D1A1B03033534)
· 이 논문은 제 28회 한글 및 한국어 정보처리 학술대회에서 '워드 임베딩을 이용한 세종 전자사전 확장'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 창원대학교 친환경해양플랜트FEED공학과
glory1503@naver.com
†† 종신회원 : 창원대학교 컴퓨터공학과 교수(Changwon Nat'l Univ.)
jcha@changwon.ac.kr
(Corresponding author임)

논문접수 : 2017년 5월 18일
(Received 18 May 2017)
논문수정 : 2017년 6월 27일
(Revised 27 June 2017)
심사완료 : 2017년 7월 10일
(Accepted 10 July 2017)

1. 서 론

의미 분석은 문장을 구성하는 단어들의 의미를 구분하고, 문장 구성 성분들 사이의 의미적 관계를 논리적으로 밝혀내어 문장의 전체적 의미를 파악하는 기술을 말한다. 그리고 형태소 분석과 구문 분석의 과정을 거쳐 이루어지는 자연어 처리의 상위 단계이다[1]. 의미 분석을 하기 위해서는 의미역 결정 단계를 거쳐야 한다.

의미역 결정(Semantic Role Labling)에는 의미 논항 역할 정보(Semantic Role)와 의미 범주 정보를 사용해야 한다. 의미 논항 역할 정보와 의미 범주 정보가 포함되어 있는 세종 전자사전의 데이터는 실제 문제에서 한국어 문장을 처리하기 위해 확장할 필요가 있다. 따라서 본 논문에서는 세종 전자사전의 확장을 시도한다.

21세기 세종 계획의 전자사전은 현대 한국어 어휘 전반에 대한 종합적이고 방대한 정보를 담고 있고 한국어 자동 처리에 보편적으로 사용될 수 있으며 다양한 전산 처리에 필수적이고 실용적인 전자사전이다[2].

세종 전자사전은 표제어에 대해 다양한 통사적, 의미적 정보가 XML형태로 수록되어 있으며 의미역을 결정하는데 사용할 수 있는 격틀 정보가 포함되어 있다. 세종 전자사전은 25,458개의 명사, 15,181개의 동사, 4,398개의 형용사와 645개의 명사 의미 하위 부류와 631개의 용언 의미 하위부류로 구성되어 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 소개한다. 제안한 방법에 대해서는 3장에서 기술하고, 마지막으로 5장에서는 결론을 기술한다.

2. 관련 연구

사전 확장을 위한 연구로는 코퍼스로부터 상·하위어 추출이 있다. 상·하위어를 추출하기 위해 사용된 방법론은 크게 패턴을 이용한 추출, 벡터를 이용한 추출로 구분된다.

[3]은 텍스트에서 패턴 인식과 의미관계를 이용하여 상·하위어를 자동적으로 추출하는 방법을 처음으로 제안하였다. 여러 가지 문법적 패턴을 생성한 후, 주어진 문장의 형태가 패턴 형태와 동일하면 관계 트리플을 추출한다. 그러나 동일한 패턴이지만 수식어들이 포함되어 있는 경우에는 수식어를 인식하지 못한다는 문제점이 있다. [4]는 텍스트에서 상·하위어 관계를 자동 추출하는 데 있어 “Latent Semantic Analysis(LSA)”를 사용하였다. LSA를 사용함으로써 정확률과 재현율을 향상시켰다. [5]는 상·하위어가 공기(co-occurrence)하는 패턴을 확인하고 상·하위어 관계의 패턴 생성에 대한 연구를 하였다. 그러나 부사어나 관사가 포함된 문장은 공기하는 패턴에 적용되지 않는다는 문제점이 있다. [6]은 코퍼스의 어휘 패턴과 의존 패턴을 이용하여 상·하위어를 추출하였다. [7]은 코퍼스를 이용하여 상·하위어 관계 패턴을 추출하는 방법을 제안하였다. 그러나 패턴 추출을 목적으로 명사의 열거를 나타낼 때 다양한 조사 또는 문장부호의 변이로 인하여 고정된 패턴 포착에 대한 어려움이 있다. 그리고 문맥에 의존적인 어휘들이 나타날 때 어휘 자체만으로 상·하위어 판단이 어렵다. [8]은 자동 시소러스 구축을 위한 상위어 자동 추출을 연구했다. 문헌정보학 용어사전에 기술된 문장의 구문적 특성을 조사하였다. 그리고 표본조사를 통하여 얻은 구문정보를 이용하여 10개의 알고리즘을 개발하였으며, 89.4%의 정확도를 보였다.

[9]는 구의 분포 벡터 표현을 이용하여 형용사-명사 구조와 한정사-명사 구조의 함의를 찾는 연구를 진행하였다. 형용사-명사 구조와 한정사 또한 의미적 벡터로 표현되어 있고 분포 벡터로 SVM, classifier를 이용하여 함의를 찾을 수 있었다. [10]은 벡터로 하위어를 찾는 연구를 진행하였다. 그러나 패턴 기반의 하위어를 찾는 것은 함께 언급되는 두 단어에만 의존하기 때문에 매우 낮은 재현율을 가져온다. 지도학습 또는 패턴 구조 없이 다른 도메인과 다른 언어에도 적용할 수 있는 벡터 유사도 방법을 이용하였다. 의존성 기반 벡터 표현을 사용하여 신경 네트워크와 윈도우 기반의 모델을 사용하여 다른 벡터 유형보다 성능이 뛰어나며, 최상위 순위 결과에서 2.73%의 Mean Average Precision(MAP)와 25.41%의 정확도를 보였다.

단어 벡터를 의미적으로 수정하기 위한 연구도 있다. [11]은 의미적 단어 벡터를 수정하기 위한 연구이다. 일반적으로 단어의 벡터를 생성할 경우에 WordNet, FrameNet 등의 의미적 어휘를 사용하지 않는다. 이 연구는 의미적 어휘를 이용하여 의미적으로 유사한 단어들은 비슷한 벡터 표현을 가지도록 벡터를 조정하는 방법이다. 재구성된 벡터(retrofitting)는 후처리 단계이며 어느 벡터 표현에나 사용할 수 있다는 장점을 가진다. 단어 표현의 의미론적 측면과 통사론적 측면을 잘 나타내는 작업에서 좋은 성능을 보였다.

3. 제안 방법

본 논문에서는 세종 전자사전에 나타나지 않는 새로운 단어의 의미 범주를 할당하기 위해서 할당하고자 하는 단어의 임베딩 벡터와 유의어를 사용한다. 먼저 대용량 문서에서 기존 세종 전자사전에 있는 단어들의 임베딩을 구했다.

워드 임베딩은 구글의 Word2vec[12]을 사용하여 구했고 재구성된 벡터를 통해서 보정한 후에 비교하였다. 재구성된 벡터는 구글의 Word2vec을 사용하여 구해진 워드 임베딩 값을 유의어 사전을 이용하여 보정된 벡터

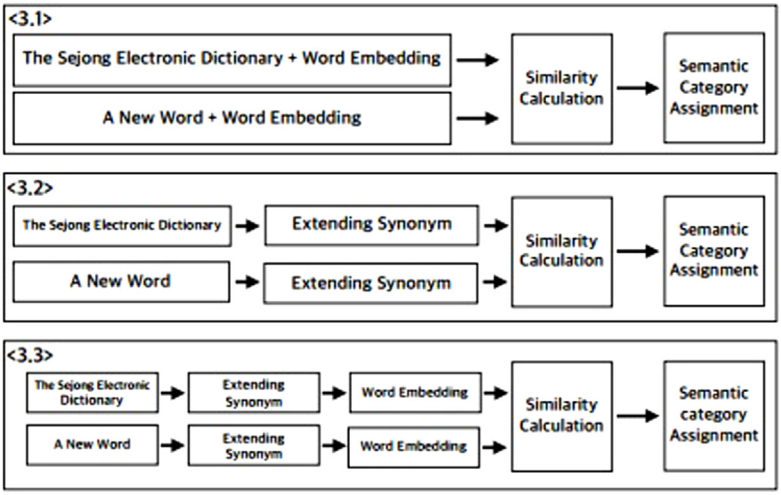

그림 1 제안 방법
Fig. 1 Proposed method

표 1 테스트 단어를 구성하는 의미 범주 수
Table 1 The number of semantic categories that make up a test word

Frequency	The number of semantic category
six or more	58
five	17
four	28
three	53
two	80
one	166

를 생성한다. 워드 임베딩 값과 세종 전자사전에 나타나지 않는 단어들의 유사도를 구하여 새로운 단어의 어휘 범주를 할당한다. 또한 세종 전자사전에 있는 단어의 유의어와 세종 전자사전에 나타나지 않는 유의어를 비교하여 새로운 단어에 의미 범주를 할당한다. 모든 실험의 cosine similarity와 pearson 상관관계는 높은 값을 가지는 단어의 의미 범주를 할당하고 euclidean distance는 낮은 값을 가지는 단어의 의미 범주를 할당한다. 선형 조합을 적용할 때 가중치 값들은 0.1 단위로 조절하며 모든 경우를 실험하여 결정하였다. 테스트에 사용된 단어는 체언이며 총 876개이다. 테스트 단어에 해당하는 의미 범주는 다중으로 부착이 가능하며 402개로 구성되어 있다. 표 1은 테스트 단어로부터 구성된 의미 범주를 보여준다.

그림 1은 본 논문에서 제안한 방법을 요약하여 보여준다. 이 세 방법에 대해서는 3.1, 3.2, 3.3장에서 자세하게 설명한다.

3.1 어휘 유사도를 이용

첫 번째로 시도한 방법은 워드 임베딩 벡터를 이용하여 의미 범주를 할당하는 방법이다. 대용량 문서를 이용하여 세종 전자사전의 단어들에 대한 워드 임베딩 값을 구한다. 또한 의미 범주를 할당하고자 하는 단어들도 같은 방법으로 워드 임베딩 값을 구한다. 이 단어들 사이의 임베딩 값의 유사도를 계산하여 의미 범주를 할당한다.

단어들의 임베딩 값을 구하기 위해서 인터넷 신문과 한국어 위키피디아에서 모은 2억 8천만 형태소 코퍼스를 사용하였다. 이 코퍼스를 창원대학교 형태소 품사 태거[13]를 사용하여 품사를 부착하였다. 워드 임베딩은 50차원이며 세종 태그셋을 기준으로 하여 41개의 소분류 형태소로 구성되어 있다. 대분류로 체언, 용언, 수식언, 독립언, 관계언, 의존형태, 기호 총 7가지로 분류된다.

유사도 계산을 위해서 cosine similarity[14,15,16], euclidean distance[17], pearson 상관관계[18]를 사용하였다. best 1은 유사도 가장 높은 하나가 정답일 때이고 best 3은 유사도 높은 상위 3개 중에 정답이 있을 때의 유사도이고 best 5는 유사도 높은 상위 5개 중에 정답이 있을 때의 정확도이다.

3.1.1 Word2vec을 적용한 어휘 유사도를 이용

표 2는 Word2vec 임베딩에 해당하는 실험 결과를 보여준다. 괄호 안의 수는 '정답 수/전체'를 나타낸다. 새로운 단어는 의미 범주 할당이 되어 있지 않은 단어를 의미한다.

3.1.2 재구성된 벡터를 적용한 어휘 유사도 이용

표 3은 재구성된 벡터를 이용하여 워드 임베딩에 해당하는 실험 결과를 보여준다. 재구성된 벡터를 생성하기 위해 유의어는 21,060개의 단어로 730개의 카테고리로 분류되어 구성된 유의어 사전을 이용하였다.

표 2 Word2vec 유사도를 이용한 실험 결과
Table 2 Experimental Results using the similarity with Word2vec

similarity	best 1	best 3	best 5
cosine similarity	25.57% (224/876)	42.57% (373/876)	49.77% (436/876)
euclidean distance	25.00% (219/876)	42.57% (373/876)	49.77% (436/876)
pearson co-relation	25.45% (223/876)	42.35% (371/876)	50.68% (444/876)

표 3 재구성된 벡터 유사도를 이용한 실험 결과
Table 3 Experimental Results using the similarity with retrofitting vector

similarity	best 1	best 3	best 5
cosine similarity	28.31% (248/876)	44.06% (386/876)	51.48% (451/876)
euclidean distance	27.63% (242/876)	43.26% (379/876)	51.83% (454/876)
pearson co-relation	28.08% (246/876)	43.72% (383/876)	52.28% (458/876)

3.1.3 실험에 대한 고찰

단어의 의미 유사도를 구하기 위해서 워드 임베딩 값을 사용하였다. 표 4는 Word2vec 워드 임베딩을 이용한 범주를 할당할 단어들을 보여준다. 표 4를 보면 워드 임베딩이 자세한 의미 구분에는 한계가 있다는 것을 보여준다. 또한 저빈도 단어에 대해서는 정확한 값을 구하지 못하는 경우도 발견할 수 있었다. 하지만 재구성된 벡터로 워드 임베딩의 실험 결과로 유의어 사전에 존재하는 단어들에 대해서는 워드 벡터들이 조정이 되었음을 알 수 있다. 하지만 유의어 사전에 존재하지 않는 단어는 기존의 Word2vec 워드 임베딩의 저빈도 단어에 대한 한계를 동일하게 가지고 있다는 것을 알 수 있었다.

표 4 범주를 할당할 단어
Table 4 Words to assign categories

Word in Sejong electronic dictionary	바톤/NNG (Baton/NNG)	삐에로/NNG (Pierrot/NNG)	각각/NNG (Each/NNG)
best 5 words	콜/NNG (Call) 큐/NNG (Cue) 하프/NNG (Half,Harp) 헤드/NNG (Head) 러브/NNG (Love)	고백/NNG (Confession) 제비족/NNG (Gigolo) 짝사랑/NNG (Crush) 진짜/NNG (Real) 무심/NNG (Indifference)	형태/NNG (Form) 방식/NNG (Way) 각자/NNG (Severalty) 순서/NNG (Sequence) 특징/NNG (Characte-ristic)

3.2 유의어 확장을 통한 매칭 이용

단어 자체의 임베딩 값의 단점을 보완하기 위해서 단어의 유의어를 이용하여 유사도를 계산한다. 여기에 사용한 유의어는 대국어사전의 유의어 정보, 개체명 사전 정보, 워드 임베딩을 이용한 유의어 정보 등이다. 유의어는 21,060개의 단어로 730개의 카테고리로 분류되어 구성된 유의어 사전을 이용하였다. 확장한 단어들의 매칭 수를 두 번째 자질로 사용한다.

$$V = \alpha \times x + \beta \times (y/z)$$

x = 어휘 유사도
y = 유의어를 확장한 단어들의 매칭 수
z = 유의어를 확장한 단어들의 전체 수
α = 어휘 유사도의 선형조합 가중치
β = 유의어를 확장한 단어들 매칭 수의 선형 조합 가중치

어휘 유사도(α)와 유의어를 확장한 단어의 매칭 수(β)를 선형 조합한다.

3.2.1 Word2vec 유사도와 유의어 확장을 통한 매칭 이용

표 5는 Word2vec 유사도와 유의어 확장을 통한 매칭 이용한 실험 결과를 보여준다. 괄호 안의 수는 '정답 수/전체'를 나타낸다. 표 6은 Word2vec 유사도와 유의어 확장을 통한 매칭 이용한 실험의 선형 조합에 적용된 가중치 값이다.

표 5 Word2vec 유사도와 확장된 유의어 매칭을 이용한 실험 결과
Table 5 Experimental results using the Word2vec similarity and matching of extended synonym

similarity	best 1	best 3	best 5
cosine similarity	31.96% (280/876)	49.09% (430/876)	56.96% (499/876)
euclidean distance	4.57% (40/876)	12.67% (111/876)	18.49% (162/876)
pearson co-relation	31.62% (277/876)	48.97% (429/876)	56.96% (499/876)

표 6 Word2vec 유사도와 확장된 유의어 매칭을 이용한 실험의 선형조합 가중치 값
Table 6 Linear combination weight of the experiment using Word2vec similarity and matching of an extended synonym

similarity	α	β
cosine similarity	0.7	0.3
euclidean distance	0.1	0.9
pearson co-relation	0.7	0.3

3.2.2 재구성된 벡터 유사도와 확장한 유의어 매칭을 이용

표 7은 재구성된 벡터 유사도와 확장한 유의어 매칭을 이용한 실험 결과를 보여준다. 표 8은 재구성된 벡터 유사도와 확장한 유의어 매칭을 이용한 실험의 선형조합에 적용된 가중치 값이다.

표 7 재구성된 벡터 유사도와 확장된 유의어 매칭을 이용한 실험 결과

Table 7 Experimental results using the retrofitting similarity and matching of the extended synonym

similarity	best 1	best 3	best 5
cosine similarity	33.68% (295/876)	51.14% (448/876)	59.36% (520/876)
euclidean distance	11.64% (102/876)	17.58% (154/876)	25.46% (223/876)
pearson co-relation	33.45% (293/876)	51.03% (447/876)	60.39% (529/876)

표 8 재구성된 벡터 유사도와 확장된 유의어 매칭을 이용한 실험의 선형조합 가중치 값

Table 8 Linear combination weight of the experiment using the retrofitting similarity and matching of an extended synonym

similarity	α	β
cosine similarity	0.7	0.3
euclidean distance	0.2	0.8
pearson co-relation	0.7	0.3

3.2.3 실험에 대한 고찰

어휘 유사도를 사용할 때 워드 임베딩 벡터 값이 존재하지 않는 단어에 대해서 평가를 할 수 없다. 또한 유사도 측정에서 상대적으로 낮은 값이었던 정답이 유의어 사전을 통해서 더 높은 값을 가지게 되어 순위 변동이 있고, 정답으로 선택되었음을 알 수 있다. 하지만 유의어의 어휘 단순 매칭만으로는 의미 범주 할당에 대한 뒷받침이 부족하다.

3.3 유의어 확장 및 워드 임베딩을 이용

유의어로 얻을 수 있는 추가 정보인 유의어의 유사도를 계산한다. 확장된 유의어의 유사도는 대상 단어의 확장된 유의어끼리의 평균 어휘 유사도를 세 번째 자질로 사용한다. 유의어의 유사도 계산의 예는 표 9와 같다.

어휘 유사도(α)와 유의어 매칭 수(β) 그리고 어휘 유사도와 유의어를 확장한 단어의 유사도(γ)를 선형 조합한다. 단, α,β,γ 의 합은 1이다.

$$V = \alpha \times x + \beta \times (y/z) + \gamma \times Average(S(w_{target}, w_{candidate}))$$

x = 어휘 유사도

y = 유의어를 확장한 단어들의 매칭 수

z = 유의어를 확장한 단어들의 전체 수

$w_{target}, w_{candidate}$ = 기준 단어로부터 확장된 유의어

S = 어휘 유사도

α = 어휘 유사도의 선형조합 가중치

β = 유의어를 확장한 단어들 매칭 수의 선형 조합 가중치

γ = 유의어를 확장한 단어들 평균 유사도의 선형 조합 가중치

표 9 유의어와 유사도 계산의 예시 단어

Table 9 Examples of synonym and similarity calculations

Target word	관공서/NNG (government office)	시청/NNG (the municipal building)
Extended synonym of target word	공관/NNG (official residence) 관청/NNG (government office) 본서/NNG (a principal office)	구청/NNG (ward office) 국세청/NNG (National Tax Service) 파출소/NNG (police substation)

3.3.1 Word2vec 유사도와 확장된 유의어 매칭 및 유사도 이용

표 10은 Word2vec 유사도와 확장된 유의어 매칭 및 유사도 이용한 실험 결과를 보여준다. 괄호 안의 수는 '정답 수/전체'를 나타낸다. 표 11은 Word2vec 유사도와 확장된 유의어 매칭 및 유사도를 이용한 실험의 선형 조합에 적용된 가중치 값이다.

표 10 Word2vec 유사도와 확장된 유의어 매칭 및 유사도를 이용한 실험 결과

Table 10 Experiment results using Word2vec similarity and matching of extended synonym and similarity

similarity	best 1	best 3	best 5
cosine similarity	32.19% (282/876)	49.43% (433/876)	57.76% (506/876)
euclidean distance	16.44% (144/876)	28.88% (253/876)	36.42% (319/876)
pearson co-relation	31.96% (280/876)	49.32% (432/876)	57.88% (507/876)

표 11 Word2vec 유사도와 확장된 유의어 매칭 및 유사도를 이용한 실험의 선형조합 가중치 값

Table 11 Linear combination weight of experiment using Word2vec similarity and matching of extended synonym and similarity

similarity	α	β	γ
cosine similarity	0.4	0.3	0.3
euclidean distance	0.1	0.4	0.5
pearson co-relation	0.4	0.3	0.3

3.3.2 재구성된 벡터 유사도와 확장된 유의어 매칭 및 유사도 이용

표 12는 재구성된 벡터 유사도와 확장한 유의어 매칭 및 유사도를 이용한 실험 결과를 보여준다. 표 13은 재구성된 벡터 유사도와 확장한 유의어 매칭 및 유사도를 이용한 실험의 선형조합에 적용된 가중치 값이다.

표 12 재구성된 벡터 유사도와 확장된 유의어 매칭 및 유사도 실험 결과

Table 12 Experimental results using the retrofitting similarity and matching of an extended synonym and similarity

similarity	best 1	best 3	best 5
cosine similarity	33.33% (292/876)	51.14% (448/876)	59.02% (517/876)
euclidean distance	18.37% (161/876)	29.79% (261/876)	39.38% (345/876)
pearson co-relation	33.33% (292/876)	50.68% (444/876)	59.25% (519/876)

표 13 재구성된 벡터 유사도와 확장된 유의어 매칭 및 유사도를 이용한 실험의 선형조합 가중치 값

Table 13 Linear combination weight of an experiment using the retrofitting similarity and matching of extended synonym and similarity

similarity	α	β	γ
cosine similarity	0.5	0.2	0.3
euclidean distance	0.2	0.4	0.4
pearson co-relation	0.4	0.2	0.4

3.3.3 실험에 대한 고찰

유의어 확장과 워드 임베딩을 이용한 실험에서 성능 향상을 확인할 수 있었다. 이는 단어의 유사성을 나타내는 벡터보다 의미를 나타내는 자질인 유의어에 의한 것이다.

표 14는 단어와 단어에 대한 의미 범주 예시이다. 표 14의 세종 전자사전 단어와 새로운 단어가 유사한 의미 범주나 동일한 의미 범주를 가져야한다고 예상된다. 그러나 실제 의미 범주는 동일하지 않다. 현재 실험은 의미 범주의 어휘가 동일할 때에만 정답으로 처리한다. 하지만 현재 단어의 의미 범주만 정답으로 간주하지 않고, 해당 의미 범주의 부모 의미 범주까지 정답으로 간주해야 한다고 판단하였다.

3.4 부모 의미 범주 이용

위의 분석 결과를 바탕으로 추가 실험을 진행하였다. 3.3 실험의 결과에 부모 의미 범주까지 적용하였다.

3.4.1 Word2vec와 부모 의미 범주 이용

표 15는 Word2vec 유사도와 부모 의미 범주까지 이용한 실험 결과이다. 괄호 안의 수는 '정답 수/전체'를 나타낸다. 표 16은 Word2vec 유사도와 부모 의미 범주까지 이용한 실험의 선형 조합을 할 때 적용한 가중치 값이다.

표 14 단어와 의미 범주의 예

Table 14 Examples of words and semantic categories

Word in Sejong electronic dictionary		New words	
Lexicon	Semantic category	Lexicon	Semantic category
의료기관 (Medical institution)	기관 (Institution)	금융기관 (Financial institution)	금융기관 (Financial institution)
재혼 (Remarriage)	만남 (Meeting)	결혼 (Marriage)	대칭적행위 (Symmetrical acting)
개발도상국 (Developing country)	상황값 (Situation value)	선진국 (Developed country)	국가 (Country)

표 15 Word2vec 유사도와 부모 의미 범주를 이용한 실험 결과

Table 15 Experimental results using the Word2vec similarity and the parent semantic category

similarity	best 1	best 3	best 5
cosine similarity	51.14% (448/876)	71.58% (627/876)	78.20% (685/876)
euclidean distance	30.37% (266/876)	49.66% (435/876)	57.88% (507/876)
pearson co-relation	51.26% (449/876)	70.55% (618/876)	78.31% (686/876)

표 16 Word2vec 유사도와 부모 의미 범주를 이용한 실험의 선형조합 가중치 값

Table 16 Linear combination weight of the experiment using the Word2vec similarity and the parent semantic category

similarity	α	β	γ
cosine similarity	0.2	0.5	0.3
euclidean distance	0.1	0.4	0.5
pearson co-relation	0.1	0.5	0.4

3.4.2 재구성된 벡터 유사도와 부모 의미 범주 이용

표 17은 재구성된 벡터 유사도와 부모 의미 범주를 이용한 실험 결과이고, 괄호 안의 수는 '정답 수/전체'를 나타낸다. 표 18은 재구성된 벡터 유사도와 부모 의미 범주를 이용한 실험의 선형 조합에 적용된 가중치 값이다.

표 8, 13, 18을 보면 재구성된 벡터를 사용한 실험에서 α의 값이 표 6, 11, 16에 비해서 상승한 것을 확인할 수 있다. 이것은 조정한 벡터 값이 정답을 결정하는데 더 많은 영향을 발휘하고 있다는 것을 증명한다. 따라서 재구성된 벡터의 효과를 알 수 있다.

표 17 재구성된 벡터 유사도와 부모 의미 범주를 이용한 실험 결과

Table 17 Experimental results using the retrofitting similarity and the parent semantic category

similarity	best 1	best 3	best 5
cosine similarity	53.88% (472/876)	72.49% (635/876)	80.37% (704/876)
euclidean distance	32.88% (288/876)	51.60% (452/876)	61.53% (539/876)
pearson co-relation	54.34% (476/876)	72.72% (637/876)	80.48% (705/876)

표 18 재구성된 벡터 유사도와 부모 의미 범주를 이용한 실험의 선형조합 가중치 값

Table 18 Linear combination weight of the experiment using a retrofitting similarity and the parent semantic category

similarity	α	β	γ
cosine similarity	0.4	0.1	0.5
euclidean distance	0.3	0.2	0.5
pearson co-relation	0.4	0.3	0.3

4. 결 론

본 논문에서는 워드 임베딩과 유의어를 활용하여 세종 전자사전을 확장하는 방법을 제안하였다. 세종 전자사전에 있는 단어와 새로운 단어의 워드 임베딩 유사도를 이용하여 새로운 단어의 어휘 범주를 할당하였다. 또한 세종 전자사전에 없는 단어의 유의어와 새로운 단어의 유의어를 비교하여 의미 범주를 할당하였다. 부모 의미 범주를 이용한 실험의 best 5 결과에 정답인 단어를 해당 의미범주에 추가하는 과정을 거쳐 의미 범주를 확장하였다.

기존의 워드 임베딩을 이용한 실험과 재구성된 벡터를 이용한 실험을 진행하여 성능을 비교하였다.

실험 결과에 의하면, 어휘 유사도와 유의어 확장 및 워드 임베딩 자질이 세종 전자사전의 의미 범주를 할당하기 위한 단어를 찾아내는 데 도움이 되었다. 그러나 어휘 유사도 자질은 단어 유사도를 찾는 데에는 도움이 되지만 어휘 의미를 구별하기에는 어려움이 있다. 유의어 자질은 의미 범주를 할당하는 데 도움이 되었다.

특히 재구성된 벡터가 성능 향상에 도움이 되었다. 어휘 유사도 실험의 결과만 봤을 때 재구성된 벡터가 성능이 2~3% 정도 높게 측정되었다. 의미적으로 유사하게 벡터를 가지도록 만들기 위한 목적의 취지에 맞게 벡터가 조정되었음을 확인 할 수 있다. 그리고 기존의 워드 임베딩보다 재구성된 벡터의 실험이 벡터를 활용한 자질에 가중치가 더 높으며, 이 또한 벡터로 인한 성능 향상임을 알 수 있다.

향후 연구로는 같은 의미 범주를 같은 단어들을 하나의 벡터로 표현하여 워드 임베딩을 계산하는 방안과 클러스터링을 적용하여 벡터로 의미 범주를 할당하는 연구를 진행해 볼 것이며, 다양한 워드 임베딩 모델을 적용할 것이며 기계 학습을 통한 모델을 구현하여 추가 실험을 진행할 계획이다.

References

[1] Daniel Jurafsky, James H. Martin, "Speech and Language Processing," Vol. 2, pp. 546, 2007.

[2] National Institute of Korean Language, Final Achievement of the 21st Century Sejong Plan, Ministry of Culture, Sports and Tourism, 2010.

[3] Hearst, M. A., "Automatic Acquisition of Hyponyms from LargeText Corpora," Association for Computational Linguistics, pp. 539-545, 1992.

[4] Cederberg, S. and Widdows, D., "Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction," *Proc. of the Conference on Natural Language Learning-2003*, pp. 111-118, 2003.

[5] Verginica Barbu Mititelu, "Automatic Extraction of Patterns Displaying Hyponym-Hypernym Co-Occurrence from Corpora," *Proc. of First Central European Student Conference in Linguistics*, 2003.

[6] SANG, Erik Tjong Kim; HOFMANN, Katja; DE RIJKE, Maarten, "Extraction of Hypernymy Information from Text," *Interactive Multi-modal Question-Answering*, pp. 223-245, 2011.

[7] Pang. Chan-Seong and Lee Hae-Yun, A Study of the Automatic Extraction of Hypernyms arid Hyponyms from the Corpus, The Korean Society for Cognitive Science, Vol. 19, 2008. (in Korean)

[8] Choi Yu-Mi and Sakong Chul, Development of Algorithm for the Automatic Extraction of Broad Term, *The 5th Proceedings of the Korean Society for information Management Conference*, pp. 227-230, 1998. (in korean)

[9] Baroni, Marco, et al., "Entailment above the word level in distributional semantics," *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 23-32, 2012.

[10] Rei, Marek, and Ted Briscoe., "Looking for Hyponyms in Vector Space," *the Conference on Natural Language Learning*, pp. 68-77, 2014.

[11] Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A., "Retrofitting word vectors to semantic lexicons," arXiv preprint arXiv:1411.4166, 2014.

[12] Mikolov, Tomas, et al., "Efficient estimation of word representations in vector space," *CoRR*, abs/1301.3781, 2013.

[13] Available: https://air.changwon.ac.kr/~airdemo/kg_tagger/, 2016-09-10.
[14] Gerard Salton, A. Wong, and C. S. Yang, A vector space model for information retrieval, Communications of the ACM, pp. 613-620, 1975.
[15] Singhal, Amit, "Modern information retrieval: A brief overview," *IEEE*, pp. 35-43. 2001.
[16] Manwar, A. B., et al., "A Vector space model for information retrieval: A MATLAB approach," *Indian Journal of Computer Science and Engineering (IJCSE)*, pp. 222-229, 2012.
[17] Danielsson, Per-Erik, "Euclidean distance mapping," *Computer Graphics and image processing*, pp. 227-248, 1980.
[18] Pearson, K., "Notes on the history of correlation," *Biometrika*, Vol. 13, pp. 25-45, 1920.

박 다 솔

2014년 창원대학교 컴퓨터공학과 학사 2017년 창원대학교 친환경해양플랜트 FEED 공학과(정보통신·컴퓨터전공)석사. 현재 창원대학교 친환경해양플랜트 FEED공학과(정보통신·컴퓨터전공) 박사. 관심분야는 자연어처리, 딥러닝, 정보추출

차 정 원

숭실대학교(학사). 포항공과대학교(석사, 박사). USC/ISI(박사후연수). 2004년~현재 창원대학교 컴퓨터공학과 교수. 관심분야는 자연어처리, 기계학습, 정보검색