우리말샘 사전을 이용한 단어 의미 유사도 측정 모델 개발

A Word Semantic Similarity Measure Model using Korean Open Dictionary

김 호 용*,**, 이 민 호*,**, 서 동 민*,**
한국과학기술정보연구원*,
과학기술연합대학원대학교**

Hoyong Kim*,**, Min-Ho Lee*,**, Dongmin Seo*,**
Korea Institute of Science and Technology Information*,
University of Science & Technology**

요약

단어 의미 유사도 측정은 정보 검색이나 문서 분류와 같이 자연어 처리 분야 문제를 해결하는 데 큰 도움을 준다. 이러한 의미 유사도 측정 문제를 해결하기 위하여 단어의 계층 구조를 사용한 기존 연구들이 있지만 이는 단어의 의미를 고려하고 있지 않아 만족스럽지 못한 결과를 보여주고 있다. 본 논문에서는 국립국어원에서 간행한 표준국어대사전에 50만 어휘가 추가된 우리말샘 사전을 기반으로 하여 한국어 단어에 대한 계층 구조를 파악했다. 그리고 단어의 용례를 word2vec 모델에 학습하여 단어의 문맥적 의미를 파악하고, 단어의 정의문을 sent2vec 모델에 학습하여 단어의 사전적 의미를 파악했다. 또한, 구축된 계층 구조와 학습된 word2vec, sent2vec 모델을 이용하여 한국어 단어 의미 유사도를 측정하는 모델을 제안했다. 마지막으로 성능 평가를 통해 제안하는 모델이 기존 모델보다 향상된 성능을 보임을 입증했다.

I. 서론

단어 의미 유사도 측정은 자연어 처리 분야 문제를 해결하는 데 큰 도움을 준다. 정보 검색 시, 검색 단어와 연관된 문서를 찾는 데 사용되고, 여러 분야의 문서를 분류해내는 데도 유용하게 사용된다. 프린스턴 대학에서는 영어에 대한 대규모 어휘 데이터베이스인 워드넷[1]을 구축하여 영단어 의미 유사도 측정에 기여했다. 워드넷은 명사, 동사, 형용사 그리고 부사를 synset이라고 하는 인지 동의어 집합으로 그룹화한 데이터베이스로, 이 synset은 개념적 의미론과 어휘 연결 관계에 의하여 상호 연결되어 있다. 이러한 워드넷의 계층 구조로부터 단어 사이의 최단 거리나 깊이를 사용하여 단어 사이의 의미적 유사도를 파악하는 방법이 많이 사용되고 있다. 국내에서는 현재 ETRI에서 제공하는 국립국어원 표준국어대사전을 기반으로 구축한 WiseWordNet[2]이 개발되어 오픈 API 형태로 서비스가 제공되고 있지만, 한국어 단어의 의미 유사도를 측정함에 있어 반대 의미를 포함하지 않았고 신조어에 대한 데이터도 부족한 상황이다. 그래서 본 논문에서는 국립국어원에서 시범 운영 중인 사용자 참여형 온라인 국어사전인 우리말샘 사전을 이용한 단어 의미 유사도 측정 모델을 개발했다.

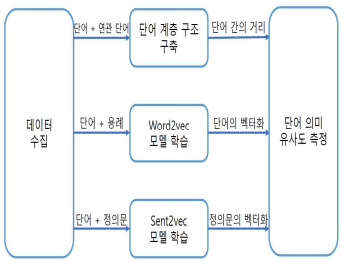
Ⅱ. 제안 방안

1. 시스템 구성도

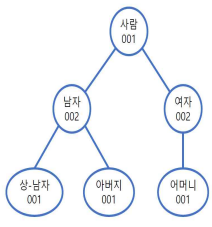
그림 1은 본 논문에서 제안하는 단어 의미 유사도 측정 시스템 구성도를 보여준다. 제안하는 시스템은 우리말샘 사전에서 제공하는 오픈 API를 사용하여 단어에 대한 데이터를 수집한 다음, 단어의 계층 구조 구축, word2vec 모델을 이용한 단어 벡터화, 그리고 sent2vec 모델을 이용한 정의문 벡터화, 마지막으로 단어 의미 유사도를 측정하는 과정으로 구성된다.

2. 단어의 계층 구조 구축

우리말샘 사전의 사전 검색 오픈 API와 사전 내용 오픈 API를 사용하여 단어의 정의문, 용례, 연관 단어, 연관 단어 유형, 링크 대상 코드를 수집했다. 연관 단어 유형에는 비슷한말, 반대말, 상위어, 하위어가 있다. 링크 대상 코드는 연관 단어의 고유 번호인 타겟 코드를 의미한다. 그리고 그림 2와 같이, 데이터 수집 이후 연관 단어의 유형에 따라 검색 단어와 연관 단어의 계층 구조를 트리 구조로 표현했다. 비슷한말은 검색 단어와 형제 노드로, 반대말은 새로운 트리로 구성했다. 또한 상위어는 검색 단어의 부모 노드로, 하위어는 검색 단어의 자식 노드로 구성하여 검색 단어와 연관 단어에 대한 트리 모양의 계층 구조를 구축했다.



▶▶ 그림 1. 단어 의미 유사도 측정 시스템 구성도



▶▶ 그림 2. 트리 모양의 단어 계층 구조 예시

3. word2vec 모델을 이용한 단어 벡터화

word2vec[3]은 벡터 공간 상에 단어를 벡터로 표현하여 단어의 문맥적 의미를 수치적으로 표현하는 비지도 기계 학습 모델이다. 이를 이용하여 우리말샘 사전으로부터 수집한 검색 단어의 용례를 학습시킴으로써 검색 단어의 문맥적 의미를 파악했다. 이후, 학습된 word2vec 모델을 이용하여 단어를 벡터화했다.

4. sent2vec 모델을 이용한 정의문 벡터화

sent2vec[4]은 word2vec의 확장 모델로 문장을 연속적인 단어들의 수열로 변환하고, 문장 속의 각 단어들을 word2vec 모델과 같은 방식으로 벡터화한 다음, 문장을 벡터로 표현하는 모델이다. 제안하는 모델에서는 수집한 모든 정의문을 sent2vec 모델에 학습시킨 후, sent2vec 모델을 이용하여 단어의 정의문을 벡터로 표현하고 두 단어의 정의문 벡터 간의 코사인 유사도를 계산하여 정의문 간의 의미 유사도를 수치적으로 표현했다.

5. 단어 의미 유사도 측정

단어의 계층 구조 구축, word2vec 모델을 이용한 단어 벡터화 그리고 sent2vec 모델을 이용한 정의문 벡터화 결과를 다음 수식을 통하여 두 단어 간의 의미 유사도를 측정했다.

$$sim_{w2v}(w_1, w_2) = cosSim\{w2v(w_1), w2v(w_2)\} \quad (1)$$

$$sim_{s2v}(w_1, w_2) = cosSim\{s2v(s_1), s2v(s_2)\} \quad (2)$$

$$sim(w_1, w_2) = c\{\alpha \bullet sim_{w2v}(w_1, w_2) + (1-\alpha) \bullet sim_{s2v}(w_1, w_2)\} \quad (3)$$

$$\alpha = \frac{1}{1 + \leq ngth(w_{1,}w_2)} \quad (4)$$

sim_{w2v}는 학습된 word2vec 모델로 두 단어를 벡터화한 다음 코사인 유사도를 측정한 값이다. $w2v(w)$는 학습된 word2vec 모델로부터 구한 단어 w에 대한 벡터이다. sim_{s2v}는 학습된 sent2vec 모델로 두 단어의 정의문을 벡터화한 다음 코사인 유사도를 측정한 값이다. $s2v(s)$는 학습된 sent2vec 모델로부터 구한 단어 w의 정의문 s에 대한 벡터이다. $sim(w_{1,}w_2)$은 두 단어의 의미 유사도 값으로 -1 이상 1 이하의 값을 가진다. c는 상수로 두 단어가 반대말 관계이면 -1, 그렇지 않은 경우 1의 값을 가진다. 또한 계층 구조에서 두 단어의 거리가 멀어질수록 문맥적 유사도가 감소한다고 가정하여 단어의 계층 구조로부터 구한 가중치 α를 수식에 추가했다. α를 구하는 방식은 수식(4)와 같다. $\leq ngth$는 계층 구조에서 두 단어 w_1과 w_2 사이의 거리를 의미한다. 예를 들어, 그림 2의 '아버지001'과 '어머니001'의 $\leq ngth$는 4가 되고 이에 따라 가중치 α는 0.2가 된다. 이렇게 해서 구한 가중치와 단어 벡터 간의 코사인 유사도, 정의문 벡터 간의 코사인 유사도를 이용하여 두 단어 간의 의미 유사도를 측정했다.

Ⅲ. 실험 결과

본 논문에서 제안한 방법과 기존 모델의 성능을 비교하기 위하여 비교 내용에 따른 단어들을 선정하고, ETRI WiseWordNet(WWN)에서 제공하는 어휘관계 분석 API를 통해 기존 모델과의 성능평가를 수행했다. 먼저, 실험을 위하여 우리말샘 사전으로부터 단어 4654개에 대한 데이터를 수집하였고, 데이터 전처리 결과 약 1만 문장의 용례 데이터를 얻을 수 있었다.

표 1은 신조어 및 연관 없는 단어 사이의 의미 유사도 측정 결과로, 기존 모델은 신조어에 대해서 결과를 도출할 수 없었고 연관이 없는 단어에 대한 의미적 유사도 또한 제안 모델(WSL)이 더 좋은 결과를 보여주었다. 일례로, '남자002'와 '요리006'을 비교할 때, 제안 모델은 '남자002'를 "남성으로 태어난 사람"라는 뜻을 가진 단어로 잘 검색한 반면, 기존 모델은 '남자002'를 '남자002'의 동음이의어인 "'국자01'의 방언"라는 뜻을 가진 '남자001'로 잘못 검색하였다.

표 1. 신조어 및 연관 없는 단어 의미 유사도 측정 결과

	남자002		
	상-남자001	요리006	취미007
w2v	0.6950	0.1375	0.1589
s2v	0.6671	0.3985	-0.0832
WSL	0.6811	0.2680	0.0378
WWN	-	0.4000	0.3250
	여자002		
	상-남자001	요리006	취미007
w2v	-0.5915	0.1886	0.2339
s2v	-0.6638	0.4072	-0.0751
WSL	-0.6457	0.2979	0.0794
WWN	-	0.3250	0.3250

* WSL: Word2vec + Sent2vec + Layer-level

감사의 글

본 연구는 한국과학기술정보연구원의 「연구데이터 오픈플랫폼 구축·활용 지원(K-18-L11-C04-S01)」 사업으로부터 지원을 받아 수행된 연구임.

■ 참 고 문 헌 ■

[1] WordNet, https://wordnet.princeton.edu/, 18.04.15
[2] WiseWordNet, http://aiopen.etri.re.kr/, 18.04.15
[3] Tomas, M., Kai, C., Greg, C., Jeffrey D. "Efficient Estimation of Word Representations in Vector Space", 2013
[4] Matteo, P. Prakhar, G., Martin, J. "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features", 2017