키워드 확장을 통한 효율적인 유의어 검출 방법

지기용*, 박지수**, 손진곤*1)
*한국방송통신대학교 대학원 정보과학과
**경기대학교 교양학부
e-mail:wlrldyd@knou.ac.kr

Efficient Synonym Detection Method through Keyword Extension

Ki Yong Ji*, JiSu Park**, Jin Gon Shon*
*Dept of Computer Science, Graduate School
Korea National Open University
**Division of General Studies, Kyonggi University

요 약

인공지능의 발달로 사람이 사용하는 자연어 형태의 문장을 통해 정보를 주고받는 질의응답 시스템이 주목받고 있다. 이러한 질의응답 시스템은 자연어로 구성된 사용자의 질의문에서 의도를 정확하게 파악해야 한다. 단순히 질의어의 키워드에 의존한 검색은 단어의 중의성을 고려하지 않아 질의문의 의도를 정확히 파악하는 데 문제가 있다. 이런 문제점을 해결하기 위해 질의문의 의미와 맥락에 따른 연관성을 이용하여 유의어를 확장하는 방법이 연구되고 있다. 본 논문에서는 워드 임베딩을 통해 생성된 단어 유사도를 이용하여 질의문에서 추출된 키워드를 확장하는 방법을 제안한다.

1. 서론

인공지능의 발달로 사람이 사용하는 자연어 형태의 문장을 통해 정보를 주고받는 질의응답 시스템이 주목받고 있다. 질의응답 시스템은 사용자로부터 자연어로 구성된 질의문을 입력받아 사용자가 원하는 답변을 제공해주고 있다. 현재의 질의응답 시스템은 사용자가 질문한 문장과 예제 데이터베이스에서 유사한 발화를 검색하여 응답을 구성한다. 단순히 질의문의 유사도를 이용하여 검색할 경우 질의문의 의도를 정확히 파악하는 데 한계가 있어 부적절한 답변을 하게 된다.

자연어 질의문에는 사용자가 원하는 응답 정보의 내용에 대해서 가장 많은 정보를 보유하고 있다. 사용자 질의어는 평균 2.21개 정도로서 함축적인 의미를 지니고 있다[1]. 단순히 질의문의 키워드에 의존한 검색은 의미 및 중의성을 고려하지 않아 질의문의 의도를 정확히 파악하는 데 한계가 있다. 이런 문제점을 해결하기 위해 질의어의 의미와 맥락에 따른 연관성을 이용하여 유의어를 확장하는 방법이 연구되고 있다.

언어의 특징을 이용하여 자연어를 통계적으로 접근하여 구축된 언어모델 이용하여 유의어를 확장하려는 많은 시도가 있었다. 이런 확장 방법에는 시소러스(thesaurus), 클러스터링(clustering), 적합도 피드백(relevance feedback) 등이 있다[1].

효율적인 질의문 분석을 위해서는 질의문의 의미와 맥락이 유사한 단어를 추출해야 한다. Word2Vec은 단어의 의미와 맥락을 고려할 수 있는 특성 벡터를 반영할 수 있는 특징이 있다. Word2Vec를 이용한 통계적 언어모델(statistical language model)을 구축하여 키워드 확장에 사용한다.

본 논문에서는 질의문에서 형태소 분석을 통하여 키워드를 추출하고 Word2Vec을 이용한 워드 임베딩을 통해서 단어 간의 유사도를 이용한 키워드 확장 방법을 제안한다.

2. 관련연구

2.1. 자연어 처리

자연어 처리는 형태소 분석, 구문 분석, 의미 분석 등 크게 3가지로 나눌 수 있다.

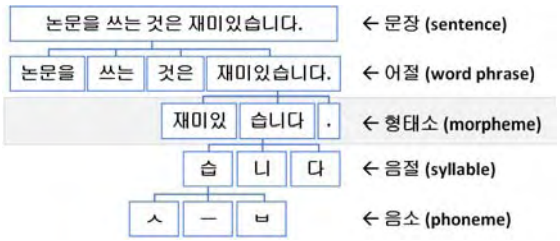
형태소 분석에서 형태소는 의미의 최소단위이다. 문장에서 형태소(최소의 유의미 단위)를 분리하고 단어를 문법적인 성질(특성)에 따라 분류하여 품사(어휘적 역할)를 결정하는 작업과정을 형태소 분석이라고 한다.

구문 분석은 형태소 분석 결과를 기반으로 문장을 이루고 있는 명사구, 동사구, 부사구 등의 구문을 묶어주는 것뿐만 아니라, 주어, 술어, 목적어 등과 같은 주요한 문장 구성성분을 밝혀내고 그들 사이의 구문 관계를 분석하여 문장의 문법적 구조를 결정하는 기술이다.

의미 분석은 문장을 구성하는 단어들의 의미를 구분하고, 통합적으로는 문장 구성성분들 사이의 의미적 관계를 논리적으로 밝혀내어 문장의 전체적 의미를 파악하는 기술이다. 특히 단어의 의미 구분은 한 단어가 둘 이상의 의

1) 교신저자

미가 있는 경우에 적용되며, 문맥상에 사용된 단어의 의미를 결정하는 것을 단어의미 중의성 해소(Word sense disambiguation)라고 한다[2].



(그림1) 언어의 단위

2.2. 워드 임베딩

문자들을 통계적으로 분석하기 위해서는 문자로 구성된 데이터를 벡터의 형태로 변환해서 표현해야 한다. 워드 임베딩(word embedding)을 이용하여 유의미한 단어들을 벡터 공간상에 표현할 수 있다. 유의미한 단어들을 가깝게 배치하여 어휘 의미를 표현하기 위한 연구가 진행되고 있다.

문자를 유의미한 벡터로 바꾸는 가장 손쉬운 방법으로 원-핫 인코딩이 있다. N개의 단어를 각각 N차원의 벡터로 표현하는 방식이다. 단어 하나에 인덱스 정수를 할당한다는 점에서 단어 주머니(bag of words)라 부르기도 한다. 단순히 하나의 요소만 1이고 나머지 모두 0인 희소 벡터 형태를 표현하기 때문에 단어 간 존재하는 유의어, 반의어와 같은 특정한 관계나 의미를 전혀 반영하지 못한다[3].

[0 0 0 1 0 0 0]

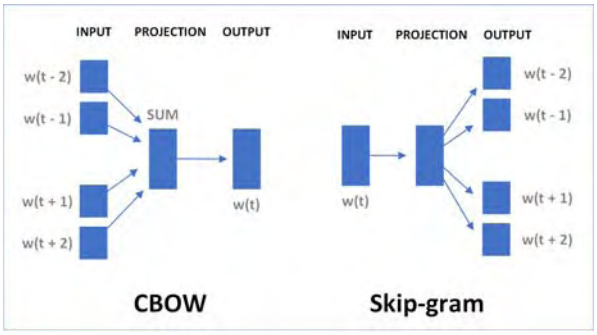
(그림2) 원-핫 인코딩

자연어 처리에 있어서 단어의 특징을 반영하지 못하는 문제점을 개선하여, 단어의 의미를 최대한 담아 단어를 벡터로 바꾸는 워드 임베딩 모델을 고안하게 된다. “단어의 주변을 보면 그 단어를 안다”라는 아이디어를 바탕으로, 문장의 맥락으로 단어를 예측하거나 단어로 맥락을 예측하는 방법(predictive method)이 있고 Word2Vec이 주목받고 있다.

2.3. Word2Vec

Word2Vec는 단어의 의미를 내포하는 벡터의 형태로 단어를 표현하는 기법으로 가장 대표적이다. Word2Vec는 특정 공간상에서 같은 문맥(context)을 갖는 단어들이 가까운 거리를 가진다는 아이디어인 분산 가설(Distributional Hypothesis)에서 출발한다. 문맥(context)에서 주변의 단어를 학습하여 벡터로 표현하기 때문에 같은 문맥을 갖는 단어들은 가까운 거리를 가지며 유의한 의미를 같는 특징을 갖게 된다. 맥락(context)에서 주변 단어를 학습하는 특징을 갖은 Word2Vec은 주어진 문장에 대한 문법적 해석이 가능하며, 단어의 거리를 통해 의미론적 추론도 가능하다[4].

Word2Vec의 알고리즘은 두 가지 방식이 있다. 하나는 맥락으로 단어를 예측하는 CBOW(continuous bag of words) 모델이 있고, 두 번째는 단어로 맥락을 예측하는 skip-gram 모델이다[5].



(그림3) CBOW 와 Skip-gram 아키텍처 비교

CBOW는 문장 내에서 주변 단어들을 모델에게 제공하고, 가운데 빈칸에 올 단어가 무엇인지 맞추는 방식으로 워드 임베딩을 학습한다. 즉 K개만큼의 주변 단어가 주어지면 중심에 올 단어의 조건부 확률을 계산하는 방식이다. Skip-gram은 한 단어를 모델에 제공하면, 모델은 이 주변에 어떤 단어들이 놓일지 맞히는 방식으로 학습한다. 즉 주어진 중심단어 주변의 K개 단어가 어떤 것이 나타날지 조건부 확률을 계산하는 방식이다[6].

2.4. 통계적 언어모델

언어 모델링은 자연어 안에서 규칙성을 찾아내고, 그 규칙성을 이용하기 위한 과정이다. 언어 모델링을 통해서 얻은 언어모델은 기계 번역이나 문자 인식, 철자 교정 등 다양한 방법으로 시스템의 정확도를 높이기 위한 중요한 요소가 된다.

언어모델이란 단어 순서에 대한 확률분포를 이야기한다. 즉 어떤 주어진 문맥(context)상에서 다음에 나올 적절한 글자 혹은 단어, 문장을 예측하는 확률 모델이다.

언어모델은 규칙에 기반을 둔 것과 통계적인 것의 두 가지로 나눌 수 있다[7]. 규칙에 기반을 둔 언어모델은 어떤 전문가가 일일이 문법 규칙을 만들기 때문에 비교적 정확하지만, 전문가의 노력이 든다는 단점이 있다. 반면 통계적 언어모델은 어떤 규칙을 사람이 알려주지 않고, 통계적 알고리즘에 의해서 스스로 학습하여 규칙을 만들어 가기 때문에 통계적 알고리즘을 만들기 위한 고수준의 노력이 필요하다.

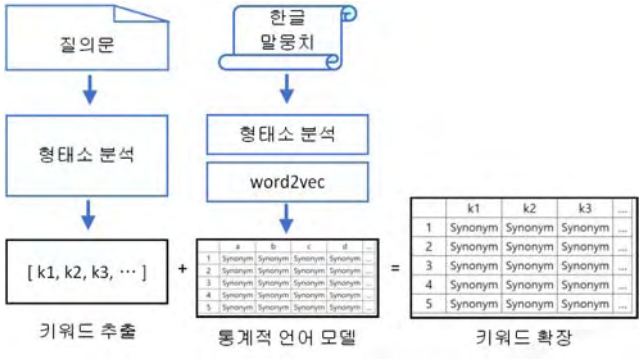
3. Word2Vec을 이용한 키워드 확장

질의문에는 사용자가 원하는 정보의 내용을 가장 많이 포함하고 있다. 그러므로 질의문에서보다 많은 정보를 얻기 위한 키워드 확장이 필요하다.

본 키워드 확장 방법은 질의문에서 키워드를 추출하고, 비정형 말뭉치에서 Word2Vec을 이용한 통계적 언어모델을 구축하여 유사도에 의한 유의어를 추출하여, 질의문의

키워드를 확장하는 것이다.

키워드 확장 모델은 아래의 (그림4)와 같다.



(그림4) 키워드 확장 모델

키워드 확장 모델의 키워드 추출은 질의문에서 형태소를 분석하여 명사에 해당하는 단어를 키워드로 사용한다. (그림5)는 질의문에서 자연어 처리 과정을 거쳐 명사의 단어를 추출하는 과정이다. 질의문을 형태소 분석하여 명사에 해당하는 단어를 질의문의 키워드로 사용한다.

(그림5) 키워드 추출 과정

통계적 언어모델을 구축하기 위하여 한글 말뭉치를 수집하였고 형태소 분석을 거쳐서 Word2Vec을 이용하여 구축한다. 파이썬을 활용하여 Word2Vec 워드 임베딩을 (그림6)과 같이 적용하여 모델을 구축한다.

```
### word2vec
token_morpheme=tokenize(txtData) #말뭉치를 형태소분석
model = gensim.models.Word2Vec(
            token_morpheme
                #형태소 분석된 단어
            , workers = 4
                #사용할 cpu의 크기
            , size=20
                #벡터의 차원의 크기
            , min_count =10
                #최소 단어의 수 보다 적게 발생할 경우 단어 무시
            , window= 2
                #컨텍스트 창의 크기
            , sample = 1e-3
                #빈번하게 등장하는 단어 다운 샘플링 0.0001
            , sg= 1
                #학습모델 skip-gram 적용
        )

model.save("feature20_context2")
```

(그림6) Word2Vec 적용 모델 방법

형태소 태깅된 토큰을 20차원의 벡터로 바꿔주고 주변 단어(window context)는 중심단어에서 앞뒤로 2개를 보고 학습하고, 단어의 출현 빈도가 10 미만인 단어는 분석에서 제외한다. Word2Vec의 학습 모델은 skip-gram을 적용하여 모델을 구축한다.

질의문에서 추출된 키워드 집합의 단어와 Word2Vec으로 구현된 통계적 언어모델에 적용하면 키워드 확장을 할 수 있다.

4. 구현 및 실험

키워드 추출을 위해서 '자연어 질의 분석 및 확장을 이용한다'라는 질의문을 형태소 분석하였다. <표1>은 자연어 질의문을 형태소 분석한 결과이다. <표2>는 질의문의 형태소 분석의 결과에서 명사의 키워드를 추출한 것이다.

<표1> 자연어 질의문 형태소 분석 결과

'자연어'	'질의'	'분석'		'및'		'확장'	'을'		'이용한'	'다'
Noun	Noun	Noun		Josa		Noun	Josa		Verb	Eomi

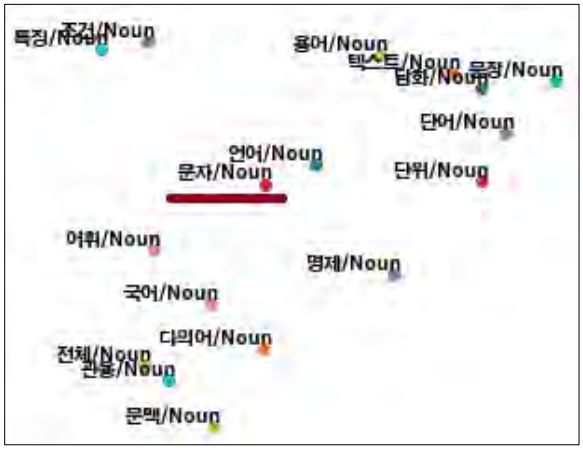
<표2> 키워드 추출

	k1	k2	k3	k4
key	'자연어'	'질의'	'분석'	'확장'

Word2Vec을 이용한 통계적 언어모델을 구축하였다. 통계적 언어 모델링의 결과로 (그림7)은 단어의 유사성에 의한 벡터의 속성값을 보여주고 있다. (그림8)은 '문자' 키워드의 유의어 분포를 2차원 형태로 표현한 결과이다. 통계적 언어모델을 이용하여 질의문의 키워드를 확장하기 위한 명사 키워드의 유의어를 이용하여 질의문의 키워드 확장을 하였다.

```
날씨/Noun
[-0.20733197  0.41331387 -0.15038924 -0.09204266 -0.22719064 -0.41671002
 -0.55535185  0.3285325  -0.00457635 -0.01023822 -0.30071506  0.01003095
  0.0005655   0.43008518  0.02826326 -0.44638926 -0.23796776  0.01087246
 -0.3355633  -0.34328398]
구름/Noun
[-0.30980858  0.29105046 -0.11398003  0.03896837 -0.20907126 -0.5656408
 -0.614029    0.41475448  0.19400273 -0.09935964 -0.37791017 -0.10870996
 -0.01513821  0.46599346  0.17003939 -0.46887404 -0.01712172  0.08367393
 -0.34965846 -0.2920114 ]
논문/Noun
[-0.22189894  0.20337394 -0.12587233  0.10989933 -0.5619966  -0.09222969
 -0.2068156   0.0719031   0.1655027  -0.17909282 -0.46969756  0.34756875
  0.53671545  0.83330154 -0.00891556 -0.55574775 -0.38025567 -0.20508939
 -0.15801011 -0.236908241
푸른/Noun
[-0.32831854  0.2759734  -0.13098401  0.01791424 -0.3391986  -0.5587698
 -0.650437    0.33071584  0.20646994  0.06400781 -0.3000711  -0.21463023
  0.0013686   0.3740759   0.31382442 -0.43192095 -0.18290891  0.06601322
 -0.39255673 -0.31536162]
밀음/Noun
[-0.15868881  0.16922344  0.11463726  0.07896692 -0.46617728 -0.40918908
 -0.4138079   0.20953059  0.02260016 -0.11086731 -0.29237548  0.04524129
  0.50870585  0.7255103   0.03151826 -0.49876022 -0.06544685  0.0477428
 -0.21643208 -0.05671332]
```

(그림7) 단어의 차원의 속성값



(그림8) '문자' 키워드의 유의어 분포

'자연어 질의 분석 및 확장을 이용한다'라는 질의문을 정확히 이해하기 위해서 키워드 확장을 통해서 질의문에서보다 많은 정보를 얻을 수 있다. 질의문을 형태소 분석하여 명사 키워드를 추출하고 한글 말뭉치에서 Word2Vec을 이용한 통계적 언어모델을 구축하였다. 이를 이용하여 키워드 확장을 통한 효율적인 유의어 검출 결과는 아래의 <표3>과 같다.

<표3> 키워드 확장을 통한 효율적인 유의어 검출 결과

자연어	[('일본어/Noun', 0.9933676719665527), ('인공어/Noun', 0.9864007830619812), ('영어/Noun', 0.938560962677002), ('』 『/Foreign', 0.894574761390686), ('2001/Number', 0.8914785981178284)]
질의	[('유도/Noun', 0.8963927626609802), ('향상/Noun', 0.895574688911438), ('학습/Noun', 0.8931348323822021), ('질/Noun', 0.8922499418258667), ('지원한/Verb', 0.892141044139862 1)]
분석	[('명확/Noun', 0.9648461937904358), ('명료/Noun', 0.9555449485778809), ('화용/Noun', 0.9499156475067139), ('프로그램/Noun', 0.9435765743255615), ('복잡/Noun', 0.941755473613739)]
확장	[('범위/Noun', 0.9400607943534851), ('세력/Noun', 0.9397995471954346), ('규모/Noun', 0.9388046860694885), ('강화/Noun', 0.9019788503646851), ('으로써/Eomi', 0.8982183337211609)]

<표3>은 키워드 확장을 통한 효율적인 유의어검출 결과를 확인하였다. '자연어'에서는 ['일본어', '인공어', '영어']의 유의어가 검출되었고 '확장'에서는 ['범위', '세력', '규모']의 유의어가 검출되었으며 '분석'에서는 ['명확', '명료', '화용']의 유의어가 검출되었고 '질의'에서는 ['유도', '향상', '학습']의 유의어가 검출되었다. 키워드의 확장을 통해서 질의문에서보다 많은 정보를 얻기 위한 유의어가 검출되었다.

5. 결론

본 논문에서 질의문에서보다 많은 정보를 얻기 위한 연구를 진행하였다. 이를 위해서 자연어 처리 분야의 형태소 분석과 워드 임베딩을 이용한 통계적 언어모델을 구축하였다. 통계적 언어모델을 바탕으로 키워드 확장을 통한 효율적인 유의어검출 모델을 제안하였다. 통계적 언어모델을 구축에 있어서 맥락에 따른 연관성을 이용하여 유의어를 확장하기 위해서 Word2Vec 이용하였다. 질의문의 키워드 확장 모델을 이용하여 효율적인 유의어검출을 하였다.

향후 연구에서는 효율적인 유의어의 키워드 확장을 통해서 중의성 해소를 위한 연구를 진행할 계획이다. 이를 통해서 보다 정확한 질의문에 의도를 파악하기 위한 연구에 활용하고자 한다.

참고문헌

[1] 신동하, 김창복, "한글 워드임베딩과 아프리오리를 이용한 검색 시스템의 질의어 확장", 한국항행학회논문지 2016.12

[2] 송민, "텍스트 마이닝", 청림, 2017

[3] 이수경, 이주진, 임성빈, "Brain's Pick : 단어 간 유사도 파악 방법", 카카오 AI 리포트, Vol.10, 2018.01

[4] 김정미, 이주홍, "word2vec을 활용한 RNN기반의 문서 분류에 관한 연구", 한국지능시스템학회, 2017.12

[5] Tomas mikolov, kai chen, Greg Corrado, Jeffercy Dean, "Efficient Estimation of Word Representations in Vector Space", 2013.09

[6] 원중호, 이한별, 문혜정, 손정, "텍스트 마이닝 기법을 이용한 경제 심리 관련 문서 분류", 한국은행 2017.04

[7] 김우성, 구명완, "통계적 언어 모델의 clustering 알고리즘과 음성인식에의 적용", 한국정보과학회, 1996.10