질의응답 시스템에서 형태소임베딩 모델과 GRU 인코더를 이용한 문장유사도 측정

이동건*O, 오교중*, 최호진*, 허정**
한국과학기술원(KAIST) 전산학부*
한국전자통신연구원(ETRI) 지식마이닝팀**
{hagg30, aomaru, hojinc}@kaist.ac.kr, jeonghur@etri.re.kr

Measuring Sentence Similarity using Morpheme Embedding Model and GRU Encoder for Question and Answering System

DongKeon Lee*O, KyoJoong Oh*, Ho-Jin Choi*, and Jeong Heo**
Korea Advanced Institute of Science and Technology (KAIST), School of computing*
Electronics and Telecommunications Research Institute (ETRI), Knowledge Mining Team**

요 약

문장유사도 분석은 문서 평가 자동화에 활용될 수 있는 중요한 기술이다. 최근 순환신경망을 이용한 인코더-디코더 언어 모델이 기계학습 분야에서 괄목할만한 성과를 거두고 있다. 본 논문에서는 한국어 형태소임베딩 모델과 GRU(Gated Recurrent Unit)기반의 인코더를 제시하고, 이를 이용하여 언어모델을 한국어 위키피디아 말뭉치로부터 학습하고, 한국어 질의응답 시스템에서 질문에 대한 정답을 유추 할 수 있는 증거문장을 찾을 수 있도록 문장유사도를 측정하는 방법을 제시한다. 본 논문에 제시된 형태소임베딩 모델과 GRU 기반의 인코딩 모델을 이용하여 문장유사도 측정에 있어서, 기존 글자임베딩 방법에 비해 개선된 결과를 얻을 수 있었으며, 질의응답 시스템에서도 유용하게 활용될 수 있음을 알 수 있었다.

주제어: 언어모델, 형태소임베딩, 순환신경망, gate recurrent unit(GRU), 문장유사도

1. 서론

최근 웹상의 많은 양의 문서와 글자 데이터를 기반으로 질의응답 시스템, 지능형 개인비서, 챗봇과 같은 자연어처리 기반의 응용프로그램이 개발되고 있다. 이런 응용프로그램을 개발하기 위해서는 웹 문서 속의 글자 데이터를 컴퓨터가 이해하고 분석을 할 수 있어야 하는데, 이를 위해 데이터마이닝 분야에서 사용되던 분류(classification) 또는 군집화(clustering) 기술이 널리 활용된다. 기존의 기술을 활용하기 위해서는 단어, 구, 문장을 실수와 같은 수치로 표현되어야 한다. 이를 위해서는 주로 언어모델을 학습하는 접근 방법을 취하게 되는데, 이는 각 단어 다음에 올수 있는 단어의 확률을 학습하는 것으로 uni-gram, n-gram, Bag of Word(BoW) model과 같은 여러 접근 방법이 제시되었으나, 최근에는 신경망네트워크를 활용하는 연구[1]가 활발히 이루어지고 있다.

Word2Vec[2]는 깊은 신경망네트워크의 한 종류인 Deep Belief Network(DBN)라는 모델을 사용하여 많은 단어로 구성된 문장 말뭉치 속에서 단어들 간의 언어모델을 학습하고 이를 기반으로 단어를 연속된 실수 벡터로 표현할 수 있게 만들어준다. 이 같은 신경망네트워크를 사용하여 언어모델을 학습함으로써 얻어지는 큰 이점은, 복잡한 자질 공학 과정 없이 단순하게 말뭉치 문장을 모델에 입력시키는 것만으로 언어모델을 학습하여 의미적으로 유사하거나, 구조적으로 비슷한 표현 순서를 보이는 단어나 구를 찾을 수 있다는 점에 있다.

이후, 구나 절, 문장, 문단 등을 실수 벡터로 표현하려는 연구로 확장되고 있다. 그 중에서 순환 신경망(recurrent neural network, RNN)에 기반을 둔 인코더-디코더 모델을 기계번역기술에 활용한 연구[3]에서 괄목할만한 성과를 거두고 있다. 문장은 단어의 순차적인 나열이기 때문에, 시계열 데이터를 학습하는데 적합한 모델인 RNN모델을 활용한다. 단어, 구, 문장을 실수 벡터로 표현하고, 학습된 언어모델로부터 최적의 번역 문장의 점수를 RNN 모델을 통해 계산해 낸다.

이 방법을 활용하면 한국어 말뭉치에서 언어모델을 학습할 수 있을 뿐만 아니라, 입력 문장을 벡터로 표현함으로써 두 문장이 입력되었을 때, 문장 간 유사도를 측정 할 수 있게 된다. 선행 연구에서는 글자임베딩 기술을 활용한 구문 유사도 측정[4] 연구에서는 문장을 음소 단위로 쪼개어 GRU를 통해 글자의 언어모델을 학습하고 주어진 문장 간의 유사도를 분석해 내는 연구를 진행하였다. [4]의 결과, 의미적으로 유사한 문장이지만 표현이 다른 경우, 유사도가 낮게 나오는 문제를 발견할 수 있었다.

본 논문에서는 한국어 문장을 실수 벡터로 표현할 수 있는 GRU 기반의 인코더를 제시하고, 이 모델과 형태소임베딩을 사용하여 입력된 두 문장의 문장유사도를 계산하고, 질의응답 시스템에서 주어진 질문에 유효한 증거문장을 찾는데 활용할 수 있는지 실험한다. 이를 통해, 한국어 문장 분석에 적합한 RNN을 이용한 언어모델 학습을 수행하고, 문장의 도메인이나 의미적인 정보를 포함하여 유사도를 분석하는 새로운 방법에 대해 제시한다.

2. 관련 연구

2.1 한국어 형태소임베딩

본 논문에서는 도메인 정보와 의미적 유사성을 반영하여 문장을 실수 벡터로 표현하기 위해서 한국어 형태소 임베딩 기술을 사용한다. 선행 연구[5]에서는 형태학상 교착어 계열인 한국어의 특성에 맞추어 형태소 별로 임베딩 모델을 학습하기 위해서 형태소임베딩 방법을 제시하였다. 영어에서의 단어는 기본적으로 어간(stem)을 독립적으로 쓰거나, 접사(affix)와 결합해서 동사만을 변형하여 사용한다. 이에 반해 한국어에서는 어간에 해당하는 어휘 형태소만 의미를 갖고, 어미나 조사에 해당하는 문법 형태소의 조합으로 단어를 구성한다. 따라서 그림 1과 같이 형태소임베딩[5]은 형태소 수준에서 어휘 형태소만을 추출하여 학습한다.

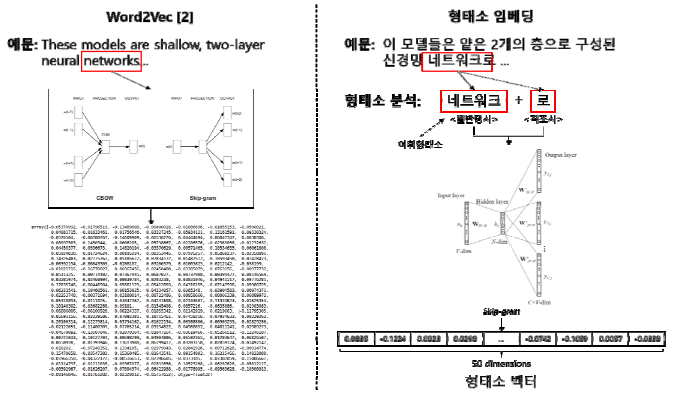

그림 1 Word2Vec[2]와 형태소임베딩[5]의 차이

추가로 형태소의 종류(type), 어깨 번호(의미 코드), 그리고 어휘 형태소에 결합된 어미나 조사에 해당하는 문법 형태소를 학습 자질로서 함께 학습한다. 이 같은 자질들을 어휘 형태소의 학습 자질로 사용하지 않은 경우, 문장 내에서 비슷한 위치에 등장하는 형태소가 비슷한 벡터로 학습될 수 있다. 이 경우, 쓰임이 다르거나, 반의어도 비슷한 벡터를 가진 형태소로 학습되어, 함의 자질 추출이나 함의 문장 생성 시 성능에 영향을 준다.

한국어 형태소임베딩의 언어모델 학습방법은 다음과 같다. 한국어 위키피디아 문서에 나오는 문장들을 형태소 분석기[6]를 통해 형태소 정보를 분석하며, 각 형태소에 대한 종류(type), 어깨 번호(의미 코드), 그리고 결합된 어미나 조사들을 학습 자질로 태그하여 말뭉치를 생성한다. 그리고 유사한 문맥 정보를 가진 단어와, 반의어를 언어 지식을 활용하여 정제한다. 그리고 DBN을 사용하여 말뭉치의 문장을 학습한다. 모델은 활성화 함수는 ReLU를 사용하였고, Dropout을 적용하였다. 10회 이상 등장하는 형태소들이 학습되었으며 학습은 3번 반복 수행한다. 학습된 한국어 형태소임베딩 모델의 벡터 크기는 100차원으로 고정시켰으며, 이는 뒤에 RNN으로 학습하는 문장 벡터를 100차원으로 고정시켰기 때문에 종속되었다.

2.2 GRU 기반 인코더-디코더

Word2Vec[2]가 깊은 신경망을 이용하여 언어모델을 학습하고 단어나 구를 실수 벡터로 표현할 수 있게 됨에 따라, 문장이나 문단도 실수 벡터로 표현하고자 하는 연구가 많이 수행되고 있다. 최근 새로운 RNN 모델의 한 종류인 GRU를 이용하여, 문장의 언어모델을 학습하고 새로운 문장을 생성하는 모델이 제시되었으며, 이를 기계번역 분야에 적용한 연구[3]가 발표되었다.

문장은 단어의 순차적인 나열이기 때문에, 문장의 앞에서부터 입력된 문맥 정보를 고려하면서 학습을 해야 한다. 따라서 신경망의 은닉 층에서 현재 입력 데이터와 이전까지 학습된 context unit까지 모두 입력으로 받아 처리를 할 수 있는 모델을 사용해야 한다. [3]에서는 이 같은 시계열 데이터를 학습하는데 적합한 모델인 RNN 모델을 활용한다. RNN 계열의 깊은 신경망으로는 gradient vanishing 문제를 해결한 Long Short Term Memory(LSTM)와, 이를 간소화한 gate recurrent unit(GRU) 등이 있다.

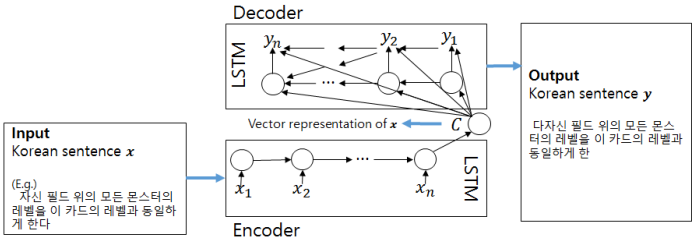

그림 2 LSTM 기반 인코더-디코더[3]

선행 연구[4]에서는 문장을 음소 단위로 쪼개어 문장을 인코딩하여 문장의 구문 유사도를 분석하고, 학습된 언어모델과 디코더를 이용하여 새로운 문장을 생성하는 연구를 수행하였다. 위키피디아 문장의 한글 한 글자를 초, 중, 종성을 쪼개 어절로 분리하여 각 단어를 알파벳으로 매핑 시킨 후에 하나의 알파벳을 구분하여 RNN의 입력으로 부여하는 character-RNN을 적용하였다. 이를 50차원의 임베딩 레이어를 이용하여 벡터로 매핑시켜 입력으로 설정하고, 어절들을 실수 벡터에 매핑시켰다. 그리고 128차원의 RNN-레이어에 대해서 인코더-디코더를 이용해, 입력 문장에 대해서 각각의 같은 길이의 문장을 생성해 내었다.

이 과정에서 음소를 분리하여 글자의 언어모델을 학습한 character-RNN 인코더는 같은 표현(사용된 단어가 같음)으로 작성된 문장의 경우 유사도를 잘 측정하였지만, 두 문장이 의미적으로 같지만 표현이 다른 경우 낮은 유사도를 보이는 것을 확인 할 수 있었다.

이 같은 문제는 질의응답 시스템, 기계번역, 함의문장 검출 및 생성, 문서 요약과 같이 문장 간의 의미적 유사도를 반영해야 하는 응용 분야에서 활용 할 수 없는 문제점이 발생한다. 본 연구는 이 같은 문제를 해결하기 위해서 어휘 형태소의 형태소 임베딩 결과를 활용하여 문장의 의미적 정보를 포함하여 언어모델을 학습하고 한국어 문장을 인코딩하는 방법을 제시하고자 한다.

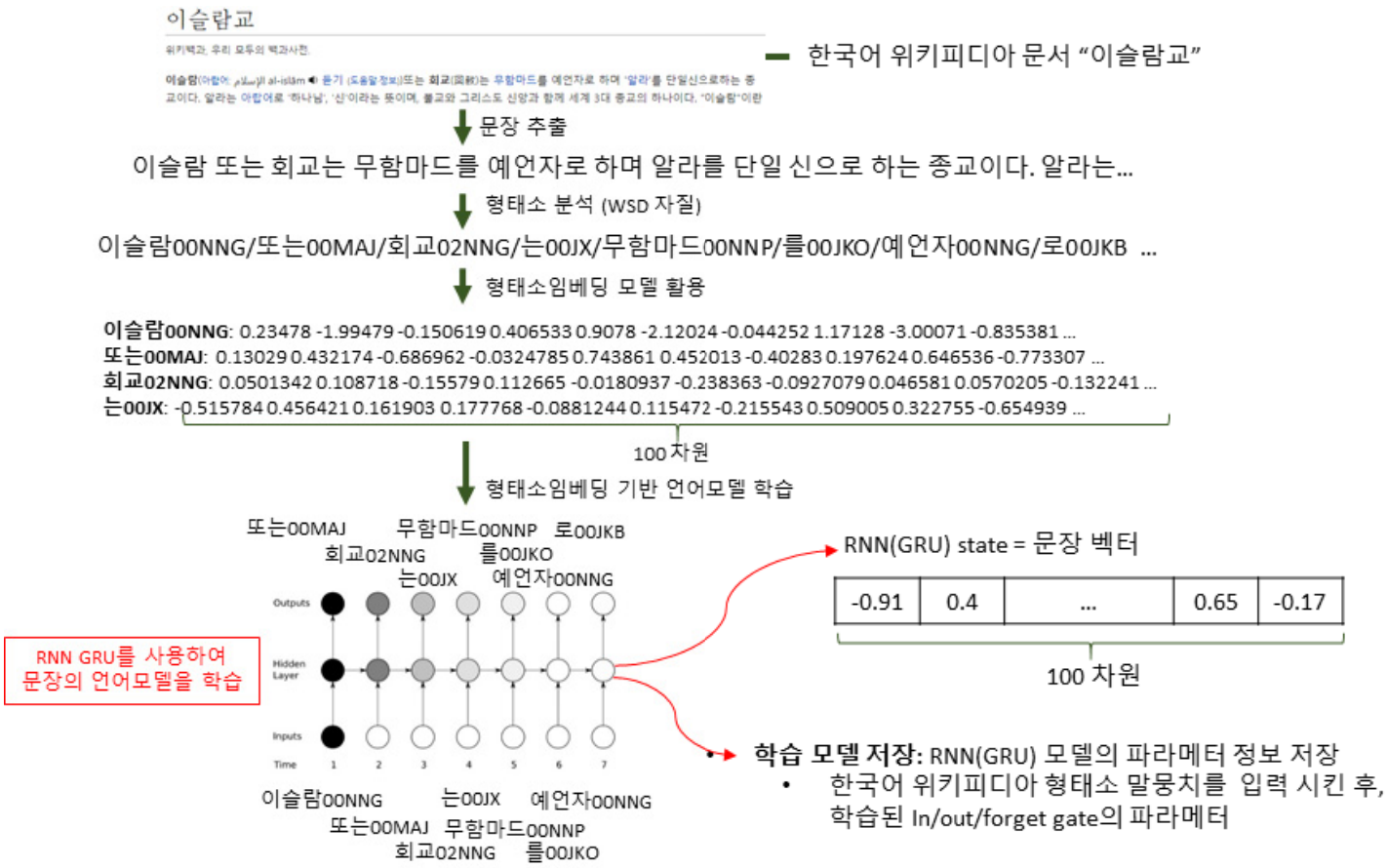

그림 3 형태소임베딩 모델 기반 한국어 문장 언어모델 학습 과정

3. 형태소임베딩 모델 기반 한국어 문장 인코딩

3.1 형태소임베딩 모델 기반 GRU 모델 학습

본 논문에서 제시하는 형태소임베딩 모델 기반 한국어 문장 언어모델 학습 과정은 그림 3과 같다. 우선 모델을 학습 시킬 한국어 말뭉치를 선택한다. 본 논문에서는 한국어 웹 백과사전인 위키피디아의 덤프 데이터(https://ko.wikipedia.org/wiki/위키백과:데이터베이스_다운로드)를 사용한다. 한국어 위키피디아 덤프 데이터(2016년 6월 기준)에서 제목과, 본문의 문장만 파싱하여 말뭉치를 구성한다. 숫자와 기호, 영문 및 한국어 문장만 정제하였을 때 2,619,773 문장이 파싱되었다.

이 말뭉치의 문장을 형태소분석기[6]를 통해 분석한 형태소 분석 자질과, 단어 의미 중의성 해소 결과를 학습 자질을 태그로 형태소임베딩에서 형태소 표현 방법따라 주석을 단다. 본 논문에서는 어께 번호와 형태소 종류를 자질로써 형태소의 주석으로 달았다. 본 연구에서 형태소분석결과 분석된 형태소는 77,172,818개이다.

선행학습으로 수행한 Word2Vec[2] 모델을 이용한 형태소 임베딩을 수행한 결과에서 해당 형태소의 형태소 벡터를 불러온다. 불러온 형태소 벡터를 입력과 출력으로 RNN 모델에 학습시킨다. 여기서 RNN 모델의 입력 층은 현재 형태소의 형태소임베딩 벡터이고, 출력 층은 다음 형태소의 형태소임베딩 벡터이다.

따라서 그림 3과 같이 첫 형태소를 학습할 때에는 "이슬람00NNG"에 해당하는 형태소임베딩 벡터를 입력으로 "또는00MAJ"에 해당하는 형태소임베딩 벡터를 출력으로 RNN을 학습시킨다. 본 논문에서는 100차원의 형태소임베딩 벡터를 사용한다. 이 방법으로 형태소 분석 자질로 주석이 달린 한국어 위키피디아 말뭉치 문장을 학습시킨다. GRU 모델을 이용한 학습에는 128개의 문장 길이를 사용하였다. 이는 평가용 말뭉치로 준비한 장학퀴즈 말뭉치의 평균 길이이다. 마지막으로 GRU in/out/forget gate의 인자를 저장하고 학습 모델로 저장한다.

3.2 GRU 모델 기반 문장 인코딩

이렇게 학습된 모델을 이용하여 한국어 인코더-디코더 언어모델의 비감독 학습을 통해 평가 문장에 대해 벡터를 추출한다. 문장을 벡터로 표현해 주는 GRU기반 인코더는 문장 길이만큼의 반복을 통해 매 반복 시 state 정보를 출력할 수 있는데, 이를 문장 벡터라고 한다. 본 논문에서는 GRU로부터 고정 100차원 벡터를 출력한다. 이 문장 벡터를 단층의 100×100 차원의 Softmax Layer를 이용하여 디코딩하였다. 마지막으로, 디코딩 된 결과와 전위된 출력에 대한 L2 Error를 최소화하도록 학습하였다. One iteration의 본 논문에서 제시하는 GRU 기반 문장 인코딩 과정은 그림 4와 같다.

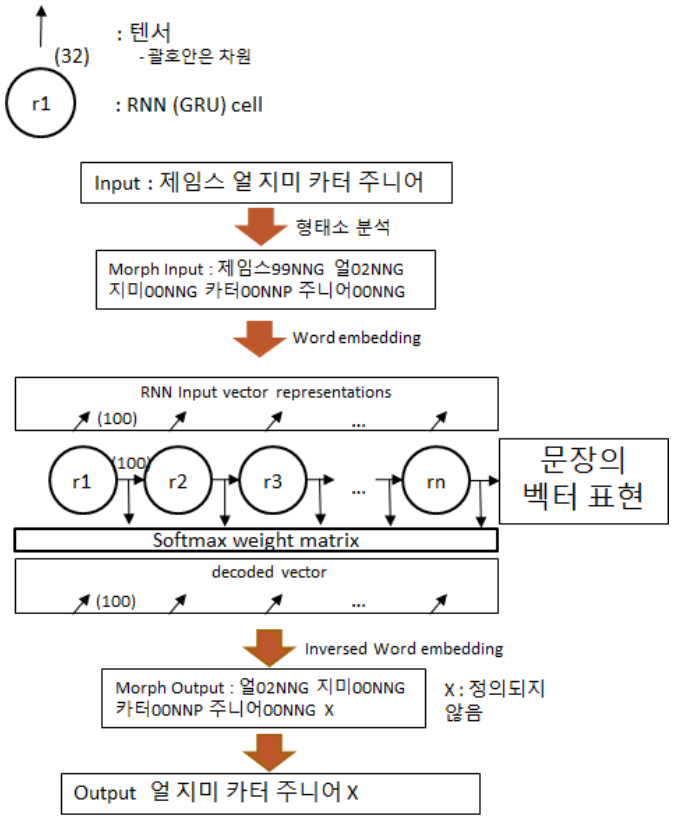

그림 4 GRU 기반 문장 인코딩 과정

4. 실험 및 결과

4.1 평가 데이터 수집 및 전처리

질의응답 시스템에서 형태소임베딩 모델과 GRU 인코더를 이용한 문장유사도 측정을 위한 실험을 위하여 평가셋을 구축하였다. 평가를 위한 평가셋은 TV 프로그램 "장학퀴즈"에서 기 출제된 문제를 정리한 문서에서 2,882 개의 질문에 대한 질문문장과 예상 답안을 찾을 수 있는 문장(증거문장), 증거문장의 문서명, 예상답안, 예상답안의 등장여부 등으로 구성된 tuple 총 195,998개로 구성되어 있다. (이하 장학퀴즈 평가셋)

이 평가셋의 질문문장과 증거문장의 문장유사도를 측정하기 위해서, 해당 문장을 형태소분석기를 통해, 한국어 문장 인코딩 모델에 사용 할 수 있도록, 형태소의 자질(어께번호, 형태소 종류)을 태그로 주석을 달았다. 그리고 선행 학습된 형태소임베딩 모델로부터 평가셋 말뭉치 문장의 형태소 각각에 대한 형태소임베딩 벡터를 불러온다.

4.2 실험 설계

본 논문에서 사용한 실험 데이터는 평가 말뭉치에서 문서명, 예상 답안을 데이터로써 사용하지 않기 때문에, 19만개의 tuple에 대해서 증거문장과 질문의 pair를 예상답안의 등장여부가 참인가의 flag로 나누었다.

다시 말하자면, (질문, 증거문장)의 pair를 key로 하여 참인 flag가 한번이라도 참으로 나타난 문장은 답을 찾는데 도움이 되는 문장 (관련문장), 해당 경우에 flag가 한 번도 참으로 나타나지 않은 문장 (비관련문장)으로 전체 문장을 분류하였다. 중복을 제거한 후 전체 19만개 tuple, 2882 질문문장에 대해서 관련문장은 8486개, 비관련문장은 128793개로 나타났다. 이를 주어진 GRU 모델을 이용하여 각 pair의 순위를 매겼고, 전체 문제 개수 대비 관련문의 순위가 1위인 문제의 개수 또는 5위 내인 문제의 개수를 평가 지표로 하였다.

4.3 문장유사도 분석 결과

평가셋의 2,882개의 질문문장 중에서 200 문장을 선별하여 제시한 인코딩 모델을 이용하여 문장유사도를 측정하고, 측정된 유사도 순으로 정렬한 후, 질문 별 top-10의 증거문장이 질문문장과 얼마나 유사한지 5-Likert scale 방법으로 사용자 평가를 진행하였다.

1점: 관련 없음
2점: 증거문장의 일부분이 질문문장과 관련 있으나, 두 문장이 같은 주제에 대해 설명하고 있지 않음
3점: 증거문장의 일부분이 질문문장과 관련 있으며, 두 문장이 같은 주제에 대해 설명
4점: 증거문장과 질문문장이 많은 부분이 관련 있으며, 같은 주제에 대한 설명
5점: 질문문장과 증거문장의 표현이 거의 일치 하며, 같은 주제에 대해 설명

표 1 문장유사도 수동 평가 결과 (200문장)

평가자	A	B
평가 평균	2.7025	2.6120

표 1에서와 같은 문장유사도 측정 결과에 대한 평가를 얻을 수 있었다. 표 2에서는 수동 평가 결과에 대한 Cohen's kappa 계수를 분석하였다. 2점 이하는 유사하지 않음으로 분석하였고, 3점 이상은 유사한 문장으로 분석하였다. 그 결과 Kappa 계수는 0.489로 수동 평가 결과 평가자 사이에서 상호 유의미하게 평가됨을 알 수 있었다.

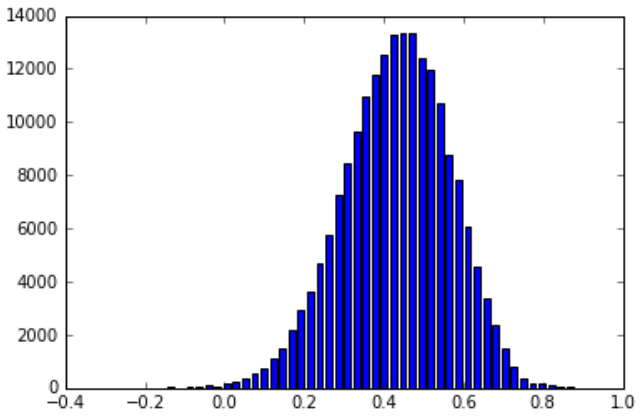

그림 5 전체 평가 문장의 유사도 분포

그림 5는 전체 평가셋에 대한 유사도 분포 결과를 나타낸 것으로 전체적으로 크게 편향되지 않은 정규분포를 따르는 것으로 보임을 알 수 있었다.

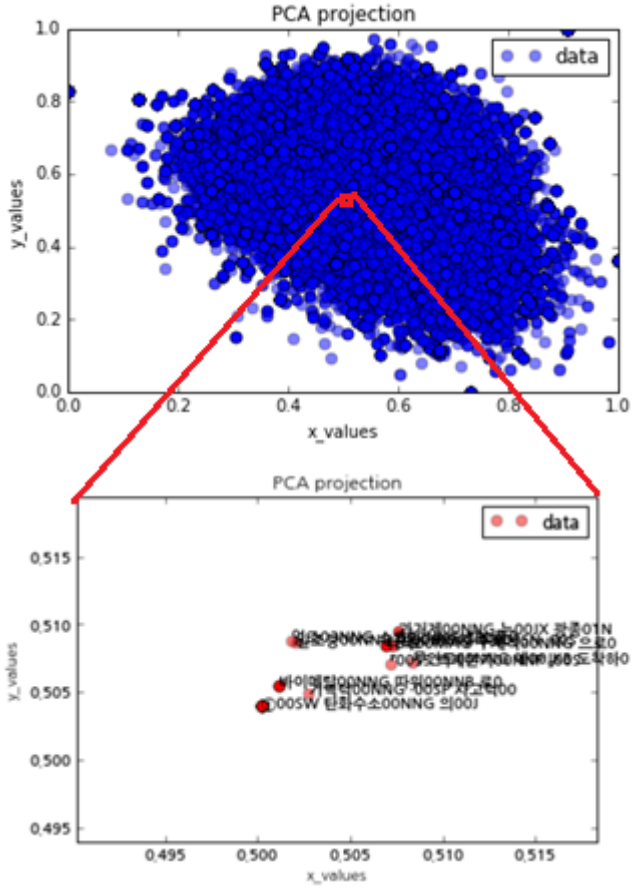

그림 6 평가셋 문장의 문장벡터 PCA 분석 결과

그림 6은 평가셋 문장의 문장벡터 PCA 분석 결과이다. 이를 [0.5, 0.51] 영역에 대해서 확대하였고, 해당 부분 영역의 상위 10개중 기호를 제외한 형태소에 대해서 전체 영역의 TF-IDF를 비교하였다. (표 2 참조) 해당 영역 내의 총 문장 수는 255개였다.

표 2 부분 영역과 전체 영역의 TF-IDF 비교

형태소명	영역 내 TF-IDF	영역 이외 TF-IDF
형태00NNG	7.0882	5.5725
특성01NNG	6.5774	6.8226
지방족00NNG	6.3951	NAN
화합물00NNG	6.2409	NAN
탄소01NNG	6.1074	NAN
CnH00SL	5.9896	NAN
와00JC	5.9896	3.2100
있01VA	5.7019	2.5657

그 결과 지방족, 화합물, 탄소, CnH (분자식) 등이 포함되는 문장은 해당하는 영역 내에서만 나타났다. 이로부터 모델이 전체 테스트 문장을 연관된 주요 키워드에 의해서 군집화하고 있음을 알 수 있었다.

다만, 유사한 문장일 경우 급격하게 한 점으로 수렴하는 경우를 확인 할 수 있었다. 이는 모델이 비교적 단순한 구조로 이뤄져있기 때문에 충분한 복잡도를 가지지 못하는 것이 이유로 생각된다.

4.4 장학퀴즈 평가셋에서의 문장유사도 적용 결과

표 3 Top-1 문장이 관련문장인 경우의 예

질문문장	순위	증거문장 / 유사도
[18] '왕'칭호를 처음 사용하고, 국호를 '신라'로 바꾼 지증왕은 우경을 실시하고, 순장을 금지하는 등 신라의 체제를 본격적으로 정비하였다. 지증왕이 정복한 곳으로 맞는 곳은 무엇일까? (정답 : 우산국)	1	지증왕(6세기 초) * 국호를 '신라', 왕호를 '왕'으로 바꿈, 행정 구역 정비* 우산국(울릉도, 독도) 복속, 우경과 수리 사업 장려법흥왕(6세기 초)* 중앙 집권 체제 정비 : 병부 및 상대등 제도 설치, 율령 반포, 17관등제 시행, 공복제 마련, 골품제 정비* 불교 공인, 독자 연호(건원) 사용 → 왕권의 초월적 지위 확보* 금관가야 정복(532) [true] [예상답안 : 우산국]
		0.324020117521
	2	정답 : 지증왕 2. ( )은 한강 유역을 장악하고 대가야를 정복하였다. [false] [예상답안 : 우산국]
		0.165295898914
	3	그 때문에 신라에선 줄기차게 사신을 보내 신임표를 내려줄 것을 요청하였고, 결국 당나라는 이세민의 신임표를 보내 선덕여왕을 '주국 낙랑군공 신라왕'으로 책봉하였다. [false] [예상답안 :낙랑군]
		0.158525586128
[24] 온도가 일정하게 유지될 때, 밀폐된 기체에 가해진 외부의 압력과 기체의 부피는 반비례한다는 법칙은 무엇일까? (정답 : 보일의 법칙)	1	온도가 일정하게 유지될 때, 밀폐된 기체에 가해진 외부의 압력과 기체의 부피는 반비례한다는 사실. 1660년 영국의 화학자 보일(Boyle, Robert)에 의해 실험적으로 발견되었다. [true] [예상답안 : 보일의 법칙]
		1.0
	2	기체 혼합물의 총압력은 혼합물 중 각 기체 부분압의 합계와 같다는 법칙. [false] [예상답안 :달톤의 법칙]
		0.312030851841
	3	프리스틀리는 식물이 이산화탄소를 흡수하고 산소를 내놓는 다고 하였다[광합성].달톤(John Dalton, 1766-1844)이 부분압의 법칙을 발견함에 [false] [예상답안 :달톤의 법칙]
		0.210616186261

표 3는 장학퀴즈 평가셋에서 예상 답안을 추측 할 수 있는 증거문장이 Top-1로 찾아진 예이다. 예시에서 알 수 있듯이 질문문장과 증거문장에서 공통적으로 사용된 핵심 어휘들이 많을수록 높은 유사도를 보이며, 이를 통해 정답을 유추 할 수 있음을 알 수 있었다.

표 4 테스트 데이터의 거짓음성 사례

질문문장	증거문장
우리 고유의 영토인 독도를 부르던 명칭이 아닌 것은 무엇일까?	독도는 과거에는 삼봉도(三峰島)라 불리었고 우산도, 가지도이라고도 부르다 고종18년(1881)부터 독도라고 부르게 되었다.
봄에 해당하는 절기가 아닌 것은 무엇일까?	봄에 해당하는 절기는 입춘, 우수(雨水),경칩(驚蟄), 춘분(春分)이다.
1950년대 우리나라는 원조물자를 이용하는'삼백산업'과 같은 소비재공업이 발달하였다.'삼백산업'에 해당하지 않는 산업은 무엇일까?	1950년대 한국산업의3대 성장부문인 제분·제당·면방직 공업을 지칭하는 말.

본 모델을 활용하여 평가셋에 존재하는 거짓 음성(False-negative) 결과도 찾아 낼 수 있었다. 표 4 에 표시된 증거문장들은 해당하는 질문문장의 답을 유추하는데 사용 될 수 있지만, 평가셋에서는 정답을 유추할 수 없는 문장으로 태깅되어 있다. 이런 현상은 주로 질문에서 부정 어휘가 사용되었고, 증거문장이 답에 반대되는 내용을 가지는 질문에서 나타났다. 동등 의미를 가지는 문장들에 대한 정보가 태깅되지 않은 데이터에 대한 무감독 학습에 의한 결과이기 때문에, 부정 의미를 이해하는 것이 힘들다. 추가적인 학습 데이터를 이용해서 문맥의 의미를 좀 더 명확하게 이해한다면 추가적인 성능 향상이 있을 것이다.

표 5 기존 Char-RNN 모델[4]과 비교

모델명	참 적용 비율 (coverage rate) (상위 n개 유사도 분석 결과에 질문의 정답을 유추 할 수 있는 문장이 포함된 질문 문장의 개수 / 총 질문 문장의 개수)	
	top 1	top 5
Char-RNN	16.51%	47.88%
형태소임베딩 + Morph-RNN	27.93%	51.63%

표 5에서는 기존에 연구한 Char-RNN 문장유사도 모델[4]과 본 논문의 Morph-RNN 모델의 평가용 질문셋(2,882 문장)에서 올바르게 증거 문장을 찾은 비율, 참 적용 비율을 측정해 보았다. 각 모델의 유사도 분석 결과로부터, 각 질문문장의 상위 n개 증거문장(n = 1, 5)에 대해서 정답을 유추 가능한 증거문장이 포함된 경우의 질문 문장 개수와 총 질문 문장 개수의 비를 측정하였다. 그 결과, 두 경우 모두 Char-RNN [4]에 비해 본 논문에서 제시한 Morph-RNN 모델이 나아진 결과를 보였다. 이는 본 논문에서 제시한 모델을 질의응답 시스템을 위한 문장유사도 분석에 보다 잘 활용될 수 있음을 보인다.

5. 결론

본 논문에서는 형태소임베딩 모델과 GRU 기반의 RNN 모델을 이용하여 한국어 문장을 실수 벡터로 표현할 수 있는 방법을 제시하고, 입력된 두 문장의 문장유사도를 계산하고, 질의응답 시스템에서 주어진 질문에 유효한 증거문장을 찾는데 활용할 수 있는지 실험하였다. 이를 통해 문장의 유사도를 분석하는 방법으로 기존의 char-RNN 모델을 이용한 방법[4]에 비해서 더 나은 결과를 얻을 수 있었다. 그뿐만 아니라, 평가셋에 존재하는 오류를 찾아 낼 수도 있었다.

본 논문에서 제시된 문장유사도 측정 방법을 활용하여 현재는 질의응답 시스템에 국한되어 있지만, 문장 생성, 유사도 분석, 문서 요약 등의 여러 영역 등 다양한 자연어처리 분야에서도 본 논문에서 제시하는 방법을 적용해 볼 수 있을 것이다.

추후 부정 의미 파악을 위한 모델 개선과, 문장 유사도를 측정 할 수 있는 평가셋 데이터를 정제하고, 보다 객관적인 평가 지표를 활용하여 추가 평가 실험을 진행할 예정이다.

감사의 글

본 연구는 미래창조과학부 산업융합원천기술개발사업의 “휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발” 과제의 지원으로 수행되었음 (과제번호 R0101-16-0062)

참고문헌

[1] Y. Bengio, et al., "Neural probabilistic language models", Innovations in Machine Learning, Springer Berlin Heidelberg, pp. 137-186, 2006.

[2] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781, 2013.

[3] Kyunghyun Cho, et al., "Learning phrase representations using rnn encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 2014.

[4] 이동건, 오교중, 최호진, "RNN을 이용한 한국어 문장 간 구문 유사도 측정", 2016 한국컴퓨터종합학술대회, 2016.

[5] 오교중, 이동건, 임채균, 최호진, 허정, "함의 자질 추출을 위한 형태소임베딩", 2016 한국컴퓨터종합학술대회, 2016.

[6] 이충희, 임준호, 임수종, 김현기, "기분석사전과 기계학습 방법을 결합한 음절 단위 한국어 품사 태깅", 정보과학회논문지 제43권 제3호, pp.362-369, 2016.