RNN을 이용한 한국어 문장 간 구문 유사도 측정

이동건○ 오교중 최호진
한국과학기술원 전산학부
hagg30@kaist.ac.kr ndex.oh@gmail.com hojinc@kaist.ac.kr

Measuring the Syntactic Similarity between Korean Sentences Using RNN

DongKeon Lee○ KyoJoong Oh Ho-Jin Choi
School of Computing, Korea Advanced Institute of Science and Technology

요약

문장 간 구문 유사도(syntactic similarity) 검사는 문장 생성 모델의 자동평가 도구로 이용되는 중요한 기술이다. 최근 딥러닝을 이용한 문장 임베딩(sentence embedding), 즉 문장 간 인코더-디코더에 관한 연구들이 진행 되고 있으며, 이를 이용한 기계 번역이 괄목할 만한 성과를 거두고 있다. 본 연구에서는 패러프레이징의 평가 도구로서 인코더-디코더 모델을 활용하고자 한다. 본 논문에서는 한국어 문장을 한국어 위키피디아 말뭉치를 이용해 RNN(recurrent neural network)으로 학습한 인코더-디코더 모델을 이용한 문장 간 유사도 분석 실험을 실시하였다.

1. 서론

최근 기계를 통한 자동 문장 생성(sentence generation) 연구가 기계 번역(statistical machine translation), 문서 요약(document summarization), 의역(paraphrasing) 등의 자연어처리 분야와, 언론, 교육과 같은 산업 분야에서 활발히 연구되고 있다. 특히 언론 분야에서는 로봇 저널리즘(robot journalism)이 대두되고 있으며, AP통신, LA Times 등과 같은 언론사에서는 자동으로 기사를 작성하는데 활용하고 있다. 국내에서도 파이낸셜뉴스와 서울대 언론정보학과 연구진이 함께 증권시황 기사를 작성해 주는 로봇을 공개하였다. 이 같은 기사는 보도자료, 날씨, 자연 재해, 스포츠 경기 결과 등의 단순하고 정형화된 정보를 고정된 형태의 문장틀에 집어넣어 문장을 생성하고 있다. 복잡하고 긴 문장을 생성시키기 위해서는 별도로 언어에 대한 깊은 이해가 필요하다. 이를 언어모델 학습 또는 인코딩이라고 부른다 [1].

본 논문에서 제시하는 문장 간 구문 유사도 측정 방법은, 입력 된 문장과 자동으로 번역, 의역된 문장 간의 유사도를 측정하는 방법이다. 기존의 문장 간 유사도 측정 방법은 많은 언어학적 지식, 언어에 따른 구조적 지식, 복잡한 자질 공학 과정이 필요하다 [2]. 하지만 최근 기계학습 분야의 깊은 신경망을 이용한 인코더-디코더 방법을 통해, 이 같은 복잡한 과정을 쉽게 해결할 수 있게 되었다 [3]. 이 방법은 복잡한 자질 공학 과정 없이, 한 글 문장을 모델에 입력하는 것만으로, 어절과 문장에 대한 의미적, 구조적 정보들을 실수로 이루어진 다차원의 벡터 형태로 표현 할 수 있다 [4]. 이를 이용하여 기계 번역, 의역 문장 생성과 같은 문장 생성 모델에서 자동 평가의 도구로 이용할 수 있을 뿐만 아니라, 표절 검사와 같은 기술로 응용할 수 있다.

본 논문에서는 깊은 신경망의 한 종류인 RNN(Recurrent Neural Network)를 사용하여, 언어모델을 80만 문장의 한국어 위키피디아 말뭉치로부터 자동으로 학습한다. 그리고 2개의 입력 문장이 주어졌을 때, 문장 간의 구문 유사도(syntactic similarity)를 학습된 언어모델로부터 얻어진 임베딩 벡터 간의 코사인 유사도로 계산한다.

본 논문은 먼저 현재 관련 연구 동향에 대해서 살펴보고, 사용된 모델 구현 방법과 실험 방법을 설명한 뒤에 실험 결과를 거쳐 결론 및 추후 연구 방향에 대한 제시하는 것으로 구성되어 있다.

2. 관련 연구

자연어처리 분야의 여러 기술에서 각종 연구와 응용을 위해는 언어에 대한 의미적, 구조적 정보들에 대한 깊은 이해가 필요하다. 특히 단어의 등장 순서에 대한 확률 분포를 계산하는 기술을 언어모델 학습 이라고 하는데, 기본적으로 통계적인 이론을 통하여 한 단어(unigram) 또는 여러 단어(n-gram)의 등장 순서에 대한 조건부 확률을 계산하는 것을 말한다 [1]. 기존의 연구에서는 문장 내 자질들을 추출하고, 이를 활용하는 방법이 주를 이루었으나, 이 방법은 자질 추출 과정이 복잡하고, 추출된 자질에 따라 성능 차이가 많이 나는 문제가 있다. 따라서 복잡한 자질 공학 과정 없이 언어모델을 학습하려는 시도로 이어졌으며, 그 결과 심층 신뢰 신경망(deep belief network)의 구조와 사전 학습 방법을 통해 단어의 의미적, 구조적 정보를 실수 형태의 다차원 벡터로 표현

하는 워드임베딩 방법이 제안되었다 [4]. 이를 기반으로 여러 구조의 깊은 신경망 네트워크를 통해 언어모델을 학습하는 연구가 계속 되었으며, 이를 인코딩이라고 부른다.

문장 생성 분야에서는 문장의 형태로 결과물을 출력하는 기술이 필요하다, 그 중에서 디코딩 기술은 학습된 언어모델로부터, 어절 또는 구의 등장 확률을 조합하여 가장 확률이 높은 문장을 생성하는 기술로, 이 기술을 통해 자동으로 자연스러운 문장을 생성 시킬 수 있기 때문에 주로 기계 번역과 의역 문장 생성(paraphrase generation) 분야에서 사용되었다 [3,5]. 특히, 의역 문장 생성 연구에서는 입력된 문장과 의미적으로 같으면서 문법적으로 자연스러운 문장을 생성하는데, 별도로 학습한 언어모델을 사용하여 디코더를 통해 문장을 생성시키고 있다.

최근에는 인코더와 디코더를 하나의 모델로 구현하여 자연어처리에 활용하는 방법이 제시되고 있다. 관련 연구에서는 깊은 신경망 기술의 한 종류인 RNN을 이용하여, 영어 말뭉치로부터 언어모델을 자동으로 학습하고, 문장을 생성하는데 도움을 주는 인코더-디코더 모델을 개발하였으며, 영어 문장으로의 번역 결과로 괄목할만한 성과를 거두었다 [3]. 하지만 아직까지 인코더-디코더 모델에 대한 한국어 문장에 대한 접근은 많지 않으며, 아직 임베딩 기술에 대해서도 연구 초기 단계에 머무르고 있다.

문장 유사도 분석은 입력된 여러 문장 간의 유사도를 계산하는 방법으로, 기존 연구에서는 주로 같은 어절이 사용되었는지 계산하는 방법과, 앞서 설명한 문장 자질 추출 결과에 기반하여 구조적 유사성을 계산하는 방법이 연구되었다 [2]. 이 연구에서도 마찬가지로 복잡한 자질 공학 과정이 요구되었으며, 의미적 유사성은 측정하기 어려운 문제가 있었다 [6]. 따라서 의미와 구조적 유사성을 하나의 모델로 측정하기 위한 방법이 요구되었으며, 이를 위해 어절의 등장 순서와 등장 빈도를 확률 분포로 계산한 언어모델을 활용하고자 한다. 본 논문에서는 한국어 문장에 대한 인코더-디코더 모델을 제안하고 이를 통해 구문 유사도를 측정하려고 한다.

3. 언어모델 학습

본 논문에서 제안하는 모델은 Cho et al. [3]을 기반으로 만들어진 다양한 오픈 소스 RNN 모델들 중 세 가지를 선택하여 변형하였다.

이들 세 가지 모델은 문장을 입력 받고 전처리를 수행하는 입력 계층, 입력과 출력에서 각각 구문 인코딩, 디코딩을 위해 양 극단에 위치한 RNN 계층, 입력 문장을 벡터 표현으로 나타내는 임베딩 계층, 문장 형태로 바꿔 출력해주는 출력 계층으로 이뤄진다. 각 모델들은 80만 개의 한국어 위키피디아 문장을 학습 데이터로 하여 학습되었다.

모델1의 입력은 위키피디아 문장을 어절로 분리하여 8,000개의 빈도수가 높은 단어들을 뽑아 각각에 대해 색인들을 부여하여 자질로 잡았다. 이를 48차원의 임베딩 레이어를 이용하여 벡터로 매핑 시켜 입력으로 설정하고, 어절들을 실수 벡터에 매핑 시켰다. 그리고 128차원의 RNN-레이어에 대해서 인코더-디코더를 이용해, 입력 문장에 대해서 같은 길이의 문장을 생성해 내었다.

두 번째 모델은 위의 모델에서 어절을 형태소 단위로 분리하여 8000개 빈도수가 높은 형태소를 추출하여 같은 방식으로 생성하였다.

세 번째 모델은 문장의 한글 한 글자를 초, 중, 종성을 쪼개 어절로 분리하여 각 단어를 알파벳으로 변환시킨 뒤에 하나의 알파벳을 구분하여 RNN의 입력으로 부여하는 char-rnn을 적용하였다. 이를 50차원의 임베딩 레이어를 이용하여 벡터로 매핑 시켜 입력으로 설정하고, 어절들을 실수 벡터에 매핑 시켰다. 그리고 128차원의 RNN-레이어에 대해서 인코더-디코더를 이용해, 입력 문장에 대해서 각각의 같은 길이의 문장을 생성해 내었다.

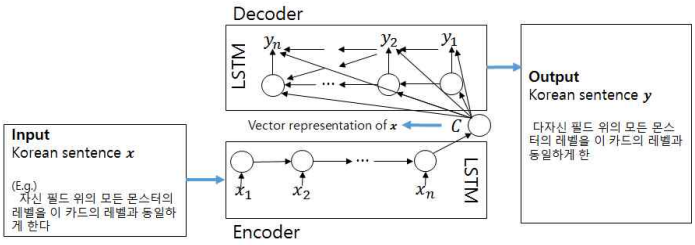

그림 1 인코더-디코더의 구조

RNN 출력을 50차원의 실수 벡터로 임베딩한다. 이 과정은 기본적으로 가변 길이의 입력 문장을 고정된 길이의 벡터 표현으로 압축(Squeeze)하여 나타낸 뒤에 입력 문장을 시프트(Shift)한 출력을 낼 수 있도록 하여 학습시킨다. 이는 Cho et al.의 인코더-디코더를 기반으로 만들었다 [3]. 그림 1은 이해를 돕기 위해 한국어 한 글자를 쉬프트 하였지만 char-rnn에서는 영문 알파벳을 시프트하여 목표 문장이 이뤄진다.

문장 유사도 분석 방법으로는 학습된 모델에 임의의 입력 문장을 부여한다. 입력 문장에 대한 인코딩 결과 출력되는 RNN의 최종 셀 상태(Cell State)가 문장의 실수 벡터로 놓을 수 있음을 확인했다. 이를 이용해 테스트 문장에 대한 코사인 유사성을 구하여 각 문장이 사람이 보기에도 비슷한 구문을 가지는지 보았다.

$$similarity(A,B) = \frac{A \cdot B}{||A||||B||}$$

4. 실험 및 결과

세 모델에 대해서 자체적으로 문장을 생성시켜내어 가장 좋은 성능을 보이는 모델을 하나를 뽑아 문장 인코딩을 시도 하였다. 앞선 두 모델들에 대해서는 형태소 분리 등의 입력 전처리를 시도하였음에도 문장 생성 결과가 대부분 문법적 의미가 없는 형태였다. 예상되는 문제점으로는 부족한 어휘량과 부적절한 임베딩 레이어 및

모델 설계의 적절성에 따른 결과라고 사료된다.

실험을 위해 장학퀴즈 평가 셋을 1,398개 문장을 사용하였다. 학습된 모델을 이용하여 해당 문장들에 대한 벡터 표현을 구하였다. 마지막으로 해당 출력을 서로 다른 문장에 대한 코사인 유사성을 계산하였다. 다음은 코사인 유사성 별 원 문장과 유사문장의 비교 결과이다.

유사성	원 문장	유사문장
0.9999 이상	'삼국지연의'에 나오는 제갈량과 관련이 없는 사자성어는 무엇일까?	동물과 관련이 없는 사자성어는 무엇일까?
	지금 보이는 숫자들은 일정한 규칙에 따른 것입니다. 규칙을 따를 때 ?에 들어갈 숫자는 무엇일까?	위 규칙을 따를 때, 물음표에 들어갈 숫자는 무엇일까?
	성별이 서로 대립하는 의미의 글자로 이루어진 단어가 아닌 것은 무엇일까?	서로 반대되는 뜻을 가진 한자로 이루어진 단어가 아닌 것은 무엇일까?
0.005 이하	심장 박동과 호흡 운동을 조절하는 생명 유지의 중추신경은 무엇일까?	서울 숭례문, 파리의 에펠탑 등 어떤 지역을 대표하는 표지
	'멧새, 멧돼지, 멧비둘기'에서 '멧'의 뜻	항공기의 속도 등을 잴 때 쓰는 것으로 음속을 기준으로 운동 물체의 속력을 나타내는 단위
	국제연합의 기준으로 국가 전체 인구 중 65세 이상의 노인 인구가 몇 % 이상을 차지할 때 고령화 사회라고 할까?	갑신정변으로 발행 중지. 박문국에서 발행한 우리나라 최초의 근대 신문.

표 1 문장 간의 유사성 비교

실험 결과는 위와 같다. 주로 동사 부분 등 문장의 마지막 부분이 일치하는 경우에 유사성이 높게 나타났으며, 문장의 앞부분이 일치하는 것은 비교적 유사성이 낮게 나타났다. 이것은 RNN의 Cell State 정보가 시퀀스가 길어짐에 따라 점차 잊히기 때문이라고 사료된다.

또한 문법이나 문맥적 의미에 대해서 상호 연관성이 보이는 경우에도 유사성이 비교적 낮게 나타나고 비슷한 단어가 발생하는 경우에만 유사성이 높게 나타났다. 이에 대한 원인으로는 부족한 학습 데이터와 문맥적인 정보에 대한 사전적인 정의를 데이터가 내포하지 않고 있기 때문이라고 사료되며, 이는 외부 지식베이스나 Labeled data에 의한 지도 학습(Supervised learning)이 필요하다.

5. 결론

말뭉치로부터 자동적으로 자질을 학습하고, 문장 간 구문 유사도 측정 방법으로 입력된 한국어 문장에 대해 한국어 구문 유사도에 대한 측정을 할 수 있게 했다. 단어가 일치하는 경우나, 일부 구문적 의미를 추출할 수 있는 것으로 보였다.

이후 연구에서는 모델을 보다 한국어 환경에 맞도록 변경하고 알맞은 전처리를 통하거나, 외부 지식 베이스를 활용하여 성능을 최적화시켜 나갈 계획이다. 또한 정량적인 평가를 위하여 알려진 언어분석기의 분석 결과를 토대로 유사성을 계산하거나, 사람이 점수를 매기는 등의 방법을 사용할 것이다.

마지막으로, 본 논문에서 제안한 한국어 구문 유사도 측정 방법을 활용하여 WiseQA 플랫폼 기술 개발 과제의 함의 문장 생성 문제에 적용하여, 그 결과를 기존의 방법과 비교해 볼 계획이다.

감사의 글

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

참고 문헌

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

[2] P. Achananuparp, et al., "The evaluation of sentence similarity measures." *Data Warehousing and Knowledge Discovery*, Springer Berlin Heidelberg, pp. 305-316, 2008.

[3] K. Cho, et al., "Learning phrase representations using RNN Encoder-Decoder for statistical machine translation," *arXiv preprint arXiv*:1406.1078, 2014.

[4] T. Mikolov, et al., “Distributed representations of words and phrases and their compositionality,” In *Proc. of Advances in Neural Information Processing Systems*, pp. 3111-3119, 2013.

[5] K. Oh, et al. "Paraphrase generation based on lexical knowledge and features for a natural language question answering system." In *Proc. of International Conference on Big Data and Smart Computing (BigComp)*, pp. 35-38, Feb. 2015.

[6] Y. Li, et al., "Sentence similarity based on semantic nets and corpus statistics," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 8, pp. 1138-1150, 2006.