딥러닝을 활용한 문장 예측 시스템

정진모, 지수진
고려대학교 컴퓨터정보통신 대학원, 전자부품연구원
e-mail : eo2587@gmail.com , sjin85@gmail.com

A Prediction System of Sentence using Deep Learning

Jin-mo Jung, Soo-jin Ji
Dept. of Computer & Information Technology, Korea University
Korea Electronics Technology Institute

요 약

본 논문은 기존에 주어진 문장 다음에 올 수 있는 문장에 대해 딥러닝을 활용하여 예측하는 시스템이며, 데이터 전처리, 문장 목적 파악, 문맥 파악의 세가지 파트로 구성되어 있다.
전처리 과정에서는 문장에 쓰인 단어에 대한 품사 정보를 Input Feature 로 추가한다. 이어서 문장 목적 파악을 위해서는 상황별로 문장을 표현하는 방법이나 단어들의 순서가 다르기 때문에 단어의 순서보다는 문장의 특징점을 학습한다.
마지막으로 문맥 파악을 위해서 이전 단계에서 학습된 문장별 목적 데이터를 기반으로 데이터의 시간적 흐름에 대한 학습을 진행함으로써 이후에 나올 수 있는 문장을 예측한다.

1. 서론

최근 하드웨어가 발전하고 및 데이터가 많아짐에 따라 인공지능을 활용하는 산업 분야는 점차 늘어가고 있다.
특히나 사용자의 의도를 파악하는 것은 인공지능의 시작점이라 볼 수 있다.
본 연구에서는 사용자가 말했던 내용을 기반으로 이후에 나올 수 있는 문장을 예측하는 시스템에 대한 연구를 하고자 한다.

2. 학습 구성

일반 이전 문장 다음에 나올 문장이 무엇인지를 파악하기 위해서는 이전 문장들의 전체적인 문맥을 파악하여야 한다.
문맥 파악을 위해서는 먼저 각 문장들이 의도하는 바를 파악하여 순서관계를 확인해야 하고, 문장들이 의도하는바를 파악하기 위해서는 문장을 구성하고 있는 단어들의 특징을 추출하여야 한다.
이러한 아이디어를 바탕으로 본 시스템은 세단계의 구성을 가지고 있다.

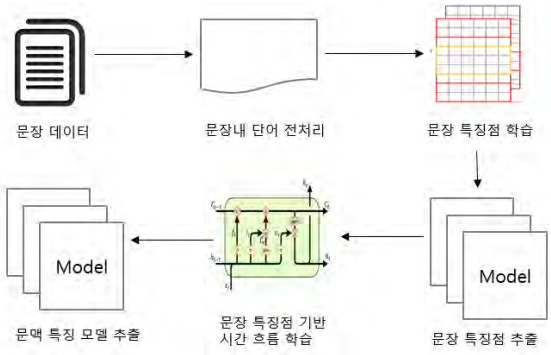

그림 1. 학습 과정

2.1 데이터 전처리

문장을 구성하는 단어들은 화자별로 표현하는 방식이나 단어들이 다르기 때문에 단어들을 아래 아이디어로 전처리 하였다. 기존 단어들은 Lemmatization 하여 단어의 표제로 전처리하고, 각 단어들이 가지고 있는 품사에 대한 정보를 학습시 데이터에 기존 단어 이외에 Input Feature 로 추가하였다.
단어들의 전처리는 Stanford CoreNLP 를 이용하였다.

* 전처리예
i. 대명사 -> PRP
: May I ask -> May **PRP** ask
ii. 고유명사와 외래어 -> NNP
: Is Mr. John in, please? -> Is **NNP** in, please?
iii. 숫자 -> CD
: Your number is 1234-5678. -> Your number is **CD**.
iv. 동사 및 축약형 -> 동사 원형 변환
: How are ~ -> How **be** ~

2.2 문장 목적 파악

문장들은 각 문장마다 고유의 목적을 가지고 있고 이러한 목적을 나타내는 특징들을 가지고 있다.

영어 문장을 예로 들면 'May I ~', 'Can I ~' 와 같은 문장은 형태는 다르지만 두 문장 모두 무언가에 대한 요청 또는 허락을 구하는 목적을 가지고 있음을 알 수 있다. 이러한 문장이 가지고 있는 목적을 나타내는 고유한 특징은 문장내에 있는 단어 순서 보다는 부분적으로 규칙적인 형태를 띄기 때문에 CNN 과 같은 특징 feature 기반의 학습 모델을 이용하여 학습하고자 한다.

학습을 위한 Input Data 는 문장의 단어 원본과 '전처리'과정에서 생성된 품사정보이다.

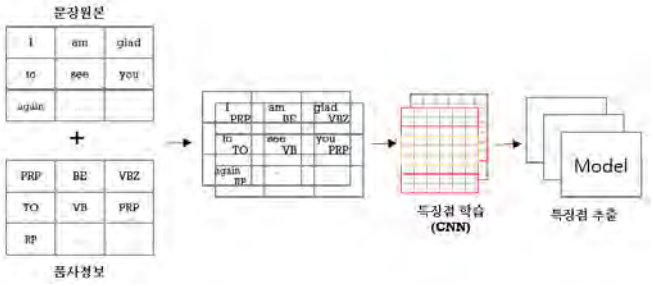

그림 2. 문장 특징점 추출예

- CNN 분류 클래스
  문장원본과 품사정보를 Inpute Feature 로 이용하여 CNN 을 통해 학습 시 추출되는 모델은 해당 문장내에 특정 품사와 단어 패턴이 표현되는지가 포함된다. 예를 들어 glad(VBZ) + to(TO) 과 같은 표현이 들어있는지를 학습을 통해 해당 문장이 감사 또는 기쁨을 나타내고 있음을 추출하고자 한다.

2.3 문맥 파악

앞선 과정에서 문장 목적을 파악 하였다 해도 문장이 구성된 순서에 의해 다음에 올 수 있는 문장의 종류가 달라지기 때문에 문장의 순서 기반으로 학습을 해야 한다.

본 시스템에서는 문맥을 학습 시키기 위해 문장 원본 이외에 앞선 과정에서 파악된 문장 목적 데이터를 Input Data 로 이용하여 시간의 흐름에 따라 변하는 데이터에 대해 할 수 있는 모델인 LSTM 을 이용하여 학습하고자 한다.

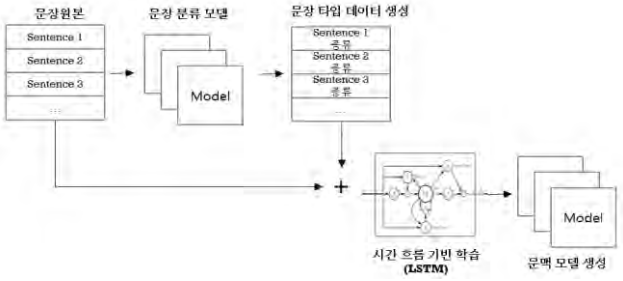

그림 3. 문맥 모델 생성 구성도

3. 실험 및 결과

본 시스템의 실험을 위한 트레이닝 데이터는 미리 10 가지 타입으로 분류된 영어 1000 문장을 이용하였다.

* 트레이닝 데이터 예
How are you? – greeting
Hello, How can I help you? - call
Thank you - appreciation
That's very nice of you. – compliment

데이터셋에서 테스트 데이터 비율은 15%로 설정 후 학습하였다.

문장의 특징점을 추출하여 문장에 대한 분류를 학습시키이 위한 방법으로는 CNN (convolutional neural network)을 이용하였다.

실험	단어 원본	품사 정보	단어 원본 + 품사 정보
100 데이터	65.2%	57.4%	68.6%
500 데이터	72.8%	65.2%	74.5%
1000 데이터	78.6%	73.3%	80.1%

표 1. 문장 특징점 추출 학습 결과

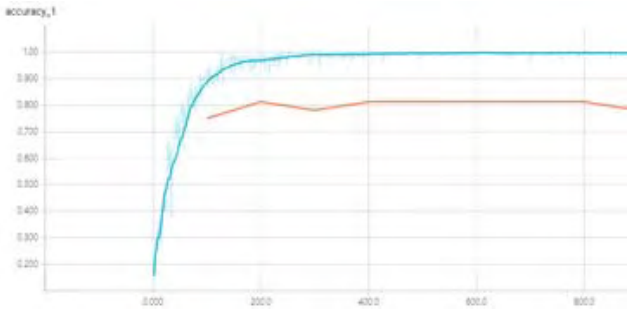

그림 4. 학습 정확도

위 실험을 통해 품사 정보나 단어 원본만을 이용하여 학습한 것보다 단어 원본 + 품사 정보를 포함시킨 학습이 정확도가 가장 높음을 알 수 있다. 1000 개의 학습 데이터 기준으로는 단어원본만 학습 시켰을 때 대비 1.5% 정도의 정확도 향상이 있었다.

4. 결론 및 향후 연구

본 논문에서는 '문장 목적 파악'까지의 실험 및 결과데이터가 포함되어 있다. 문맥 파악을 하고 의미있는 결과를 도출하기 위해서는 좀 더 많은 종류의 트레이닝 데이터 및 미리 분류된 문장 카테고리 데이터가 필요하다.

그러나 적은 데이터 기반에서도 문장의 목적을 파악하는데에는 문장 내 단어에 대한 품사 정보를 Input Feature 로 포함시키는 것이 효과적임을 알 수 있었다.

이러한 결과를 바탕으로 많은 학습 데이터를 확보한다면, ‘문맥 파악’을 위한 학습시에도 ‘문장 목적 파악’의 의해 생성된 문장 분류 데이터가 가공되지 않은 평문을(plain text) 이용하는 것보다 신뢰성이 높을 수 있음을 예상할 수 있다.

본 연구에서는 영어만을 이용하여 진행하였지만 향후 한글 각 단어들에 대한 품사 정의 방법을 연구하여 한글 문장에 대한 다음 문장 예측 시스템 관련해서도 연구가 진행되어야 할 것이다.

참고문헌

[1] "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification", Ye Zhang, Byron Wallace, 13 Oct 2015

[2] "A C-LSTM Neural Network for Text Classification",Chunting Zhou, 27 Nov 2015

[3] "Sequential short-text classification with recurrent and convolutional neural networks",Ji Young Lee,12 Mar 2016

[4] "Recurrent Convolutional Neural Networks for Text Classification.", Siwei Lai,

[5] 말뭉치로부터 자동 추출된 문맥 반영 구문 규칙을 이용한 영어 구문 분석, 조정미, 1995.9

[6] 딥러닝 기반 텍스트 질의응답을 위한 지식 추출 데이터 증강 기법, 조 휘 열, 2017.7

[7] 딥러닝을 이용한 형태소 분석 기반의 상품 카테고리 분류 기법, 김진삼, 2017,8